Advanced High-School Mathematics

David B. Surowski
Shanghai American School
Singapore American School

January 29, 2011

Preface/Acknowledgment

The present expanded set of notes initially grew out of an attempt to flesh out the International Baccalaureate (IB) mathematics "Further Mathematics" curriculum, all in preparation for my teaching this during during the AY 2007–2008 school year. Such a course is offered only under special circumstances and is typically reserved for those rare students who have finished their second year of IB mathematics HL in their junior year and need a "capstone" mathematics course in their senior year. During the above school year I had two such IB mathematics students. However, feeling that a few more students would make for a more robust learning environment, I recruited several of my 2006–2007 AP Calculus (BC) students to partake of this rare offering resulting. The result was one of the most singular experiences I've had in my nearly 40-year teaching career: the brain power represented in this class of 11 blue-chip students surely rivaled that of any assemblage of high-school students anywhere and at any time!

After having already finished the first draft of these notes I became aware that there was already a book in print which gave adequate coverage of the IB syllabus, namely the Haese and Harris text[1] which covered the four IB Mathematics HL "option topics," together with a chapter on the retired option topic on Euclidean geometry. This is a very worthy text and had I initially known of its existence, I probably wouldn't have undertaken the writing of the present notes. However, as time passed, and I became more aware of the many differences between mine and the HH text's views on high-school mathematics, I decided that there might be some value in trying to codify my own personal experiences into an advanced mathematics textbook accessible by and interesting to a relatively advanced high-school student, without being constrained by the idiosyncracies of the formal IB Further Mathematics curriculum. This allowed me to freely draw from my experiences first as a research mathematician and then as an AP/IB teacher to weave some of my all-time favorite mathematical threads into the general narrative, thereby giving me (and, I hope, the students) better emotional and

[1]Peter Blythe, Peter Joseph, Paul Urban, David Martin, Robert Haese, and Michael Haese, Mathematics for the international student; Mathematics HL (Options), Haese and Harris Publications, 2005, Adelaide, ISBN 1 876543 33 7

intellectual rapport with the contents. I can only hope that the readers (if any) can find some something of value by the reading of my stream-of-consciousness narrative.

The basic layout of my notes originally was constrained to the five option themes of IB: geometry, discrete mathematics, abstract algebra, series and ordinary differential equations, and inferential statistics. However, I have since added a short chapter on inequalities and constrained extrema as they amplify and extend themes typically visited in a standard course in Algebra II. As for the IB option themes, my organization differs substantially from that of the HH text. Theirs is one in which the chapters are independent of each other, having very little articulation among the chapters. This makes their text especially suitable for the teaching of any given option topic within the context of IB mathematics HL. Mine, on the other hand, tries to bring out the strong interdependencies among the chapters. For example, the HH text places the chapter on abstract algebra (Sets, Relations, and Groups) before discrete mathematics (Number Theory and Graph Theory), whereas I feel that the correct sequence is the other way around. Much of the motivation for abstract algebra can be found in a variety of topics from both number theory and graph theory. As a result, the reader will find that my Abstract Algebra chapter draws heavily from both of these topics for important examples and motivation.

As another important example, HH places Statistics well before Series and Differential Equations. This can be done, of course (they did it!), but there's something missing in inferential statistics (even at the elementary level) if there isn't a healthy reliance on analysis. In my organization, this chapter (the longest one!) is the very last chapter and immediately follows the chapter on Series and Differential Equations. This made more natural, for example, an insertion of a theoretical subsection wherein the density of two independent continuous random variables is derived as the convolution of the individual densities. A second, and perhaps more relevant example involves a short treatment on the "random harmonic series," which dovetails very well with the already-understood discussions on convergence of infinite series. The cute fact, of course, is that the random harmonic series converges with probability 1.

I would like to acknowledge the software used in the preparation of these notes. First of all, the typesetting itself made use of the industry standard, LaTeX, written by Donald Knuth. Next, I made use of three different graphics resources: *Geometer's Sketchpad*, *Autograph*, and the statistical workhorse *Minitab*. Not surprisingly, in the chapter on Advanced Euclidean Geometry, the vast majority of the graphics was generated through Geometer's Sketchpad. I like Autograph as a general-purpose graphics software and have made rather liberal use of this throughout these notes, especially in the chapters on series and differential equations and inferential statistics. Minitab was used primarily in the chapter on Inferential Statistics, and the graphical outputs greatly enhanced the exposition. Finally, all of the graphics were converted to PDF format via ADOBE® ACROBAT® 8 PROFESSIONAL (version 8.0.0). I owe a great debt to those involved in the production of the above-mentioned products.

Assuming that I have already posted these notes to the internet, I would appreciate comments, corrections, and suggestions for improvements from interested colleagues and students alike. The present version still contains many rough edges, and I'm soliciting help from the wider community to help identify improvements.

Naturally, my greatest debt of gratitude is to the eleven students (shown to the right) I conscripted for the class. They are (back row): Eric Zhang (Harvey Mudd), Jong-Bin Lim (University of Illinois), Tiimothy Sun (Columbia University), David Xu (Brown University), Kevin Yeh (UC Berkeley), Jeremy Liu (University of Virginia); (front row): Jong-Min Choi (Stanford University), T.J. Young (Duke University), Nicole Wong (UC Berkeley), Emily Yeh (University of Chicago), and Jong Fang (Washington University). Besides providing one of the most stimulating teaching environments I've enjoyed over

my 40-year career, these students pointed out countless errors in this document's original draft. To them I owe an un-repayable debt.

My list of acknowledgements would be woefully incomplete without special mention of my life-long friend and colleague, Professor Robert Burckel, who over the decades has exerted tremendous influence on how I view mathematics.

David Surowski
Emeritus Professor of Mathematics
May 25, 2008
Shanghai, China
dbski@math.ksu.edu
http://search.saschina.org/surowski

First draft: April 6, 2007
Second draft: June 24, 2007
Third draft: August 2, 2007
Fourth draft: August 13, 2007
Fifth draft: December 25, 2007
Sixth draft: May 25, 2008
Seventh draft: December 27, 2009
Eighth draft: February 5, 2010
Ninth draft: April 4, 2010

Contents

Chapter 1

Advanced Euclidean Geometry

1.1 Role of Euclidean Geometry in High-School Mathematics

If only because in one's "further" studies of mathematics, the results (i.e., *theorems*) of Euclidean geometry appear only infrequently, this subject has come under frequent scrutiny, especially over the past 50 years, and at various stages its very inclusion in a high-school mathematics curriculum has even been challenged. However, as long as we continue to regard as important the development of logical, deductive reasoning in high-school students, then Euclidean geometry provides as effective a vehicle as any in bringing forth this worthy objective.

The lofty position ascribed to deductive reasoning goes back to at least the Greeks, with Aristotle having laid down the basic foundations of such reasoning back in the 4th century B.C. At about this time Greek geometry started to flourish, and reached its zenith with the 13 books of Euclid. From this point forward, geometry (and arithmetic) was an obligatory component of one's education and served as a paradigm for deductive reasoning.

A well-known (but not well *enough* known!) anecdote describes former U.S. president Abraham Lincoln who, as a member of Congress, had nearly mastered the first six books of Euclid. By his own admission this was not a statement of any particular passion for geometry, but that such mastery gave him a decided edge over his counterparts is dialects and logical discourse.

Lincoln was not the only U.S. president to have given serious thought

to Euclidean geometry. President James Garfield published a novel proof in 1876 of the Pythagorean theorem (see Exercise 3 on page 4).

As for the subject itself, it is my personal feeling that the logical arguments which connect the various theorems of geometry are every bit as fascinating as the theorems themselves!

So let's get on with it ... !

1.2 Triangle Geometry

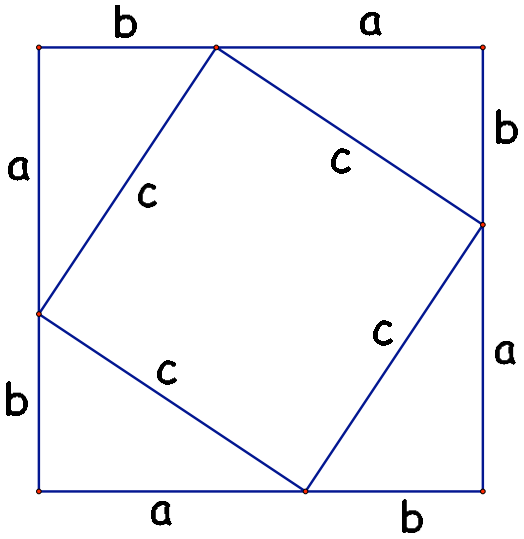
1.2.1 Basic notations

We shall gather together a few notational conventions and be reminded of a few simple results. Some of the notation is as follows:

A, B, C	labels of points
$[AB]$	The line segment joining A and B
AB	The length of the segment $[AB]$
(AB)	The line containing A and B
$\widehat{A}$	The angle at A
$C\widehat{A}B$	The angle between $[CA]$ and $[AB]$
$\triangle ABC$	The triangle with vertices A, B, and C
$\triangle ABC \cong \triangle A'B'C'$	The triangles $\triangle ABC$ and $\triangle A'B'C'$ are congruent
$\triangle ABC \sim \triangle A'B'C'$	The triangles $\triangle ABC$ and $\triangle A'B'C'$ are similar

1.2.2 The Pythagorean theorem

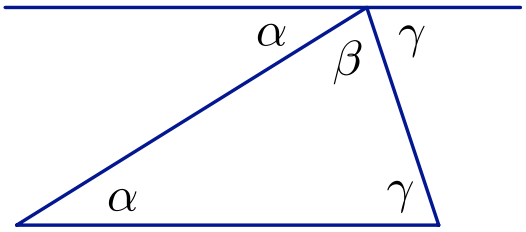
One of the most fundamental results is the well-known **Pythagorean Theorem**. This states that $\boldsymbol{a^2 + b^2 = c^2}$ in a right triangle with sides a and b and hypotenuse c. The figure to the right indicates one of the many known proofs of this fundamental result. Indeed, the area of the "big" square is $(a+b)^2$ and can be decomposed into the area of the smaller square plus the areas of the four congruent triangles. That is,



$$(a+b)^2 = c^2 + 2ab,$$

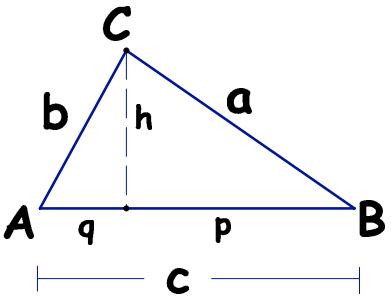
which immediately reduces to $a^2 + b^2 = c^2$.

Next, we recall the equally well-known result that the sum of the interior angles of a triangle is 180°. The proof is easily inferred from the diagram to the right.



EXERCISES

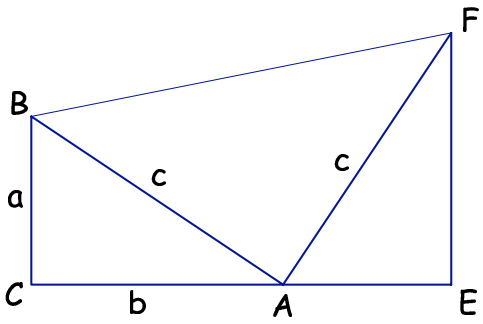
1. Prove **Euclid's Theorem for Proportional Segments**, i.e., given the right triangle $\triangle ABC$ as indicated, then



$$h^2 = pq, \quad a^2 = pc, \quad b^2 = qc.$$

2. Prove that the sum of the interior angles of a quadrilateral $ABCD$ is 360°.

3. In the diagram to the right, $\triangle ABC$ is a right triangle, segments $[AB]$ and $[AF]$ are perpendicular and equal in length, and $[EF]$ is perpendicular to $[CE]$. Set $a = BC$, $b = AB$, $c = AB$, and deduce President Garfield's proof[1] of the Pythagorean theorem by computing the area of the trapezoid $BCEF$.

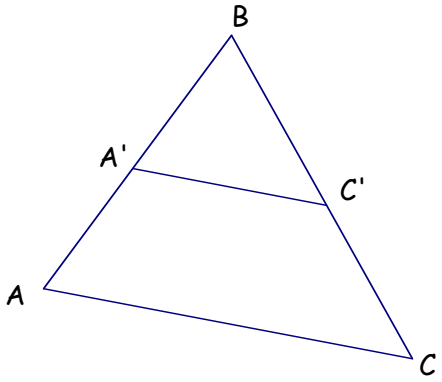


1.2.3 Similarity

In what follows, we'll see that many—if not most—of our results shall rely on the proportionality of sides in **similar triangles**. A convenient statement is as follows.

Similarity. Given the similar triangles $\triangle ABC \sim \triangle A'BC'$, we have that

$$\frac{A'B}{AB} = \frac{BC'}{BC} = \frac{A'C'}{AC}.$$



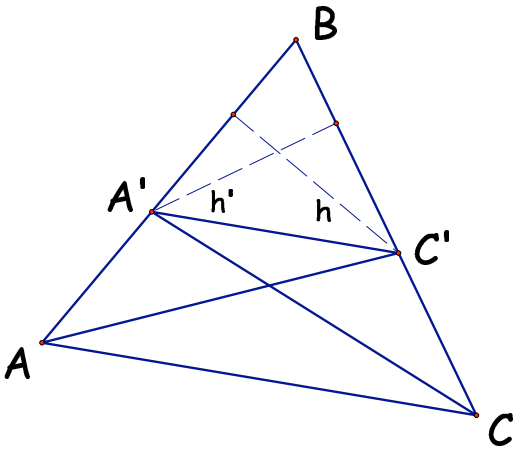
Conversely, if

$$\frac{A'B}{AB} = \frac{BC'}{BC} = \frac{A'C'}{AC},$$

then triangles $\triangle ABC \sim \triangle A'BC'$ are similar.

[1] James Abram Garfield (1831–1881) published this proof in 1876 in the JOURNAL OF EDUCATION (Volume 3 Issue 161) while a member of the House of Representatives. He was assasinated in 1881 by Charles Julius Guiteau. As an aside, notice that Garfield's diagram also provides a simple proof of the fact that perpendicular lines in the planes have slopes which are negative reciprocals.

PROOF. Note first that $\triangle AA'C'$ and $\triangle CA'C'$ clearly have the same areas, which implies that $\triangle ABC'$ and $\triangle CA'B$ have the same area (being the previous common area plus the area of the common triangle $\triangle A'BC'$). Therefore

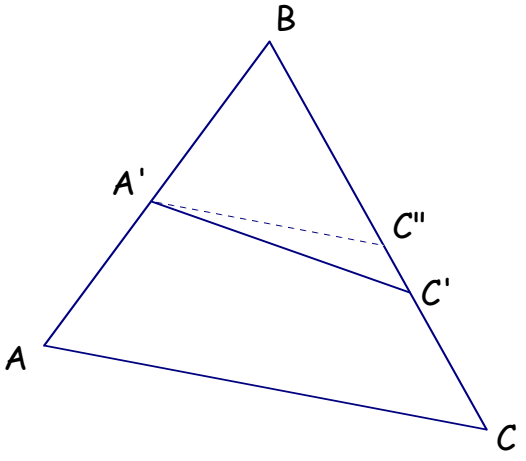


$$\begin{aligned}
\frac{A'B}{AB} &= \frac{\frac{1}{2}h \cdot A'B}{\frac{1}{2}h \cdot AB} \\
&= \frac{\text{area}\,\triangle A'BC'}{\text{area}\,\triangle ABC'} \\
&= \frac{\text{area}\,\triangle A'BC'}{\text{area}\,\triangle CA'B} \\
&= \frac{\frac{1}{2}h' \cdot BC'}{\frac{1}{2}h' \cdot BC} \\
&= \frac{BC'}{BC}
\end{aligned}$$

In an entirely similar fashion one can prove that $\dfrac{A'B}{AB} = \dfrac{A'C'}{AC}$. Conversely, assume that

$$\frac{A'B}{AB} = \frac{BC'}{BC}.$$

In the figure to the right, the point C'' has been located so that the segment $[A'C'']$ is parallel to $[AC]$. But then triangles $\triangle ABC$ and $\triangle A'BC''$ are similar, and so



$$\frac{BC''}{BC} = \frac{A'B}{AB} = \frac{BC'}{BC},$$

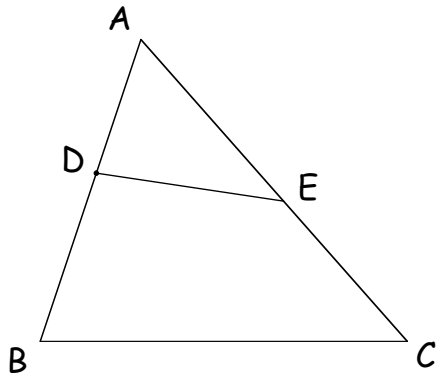
i.e., that $BC'' = BC'$. This clearly implies that $C' = C''$, and so $[A'C']$ is parallel to $[AC]$. From this it immediately follows that triangles

$\triangle ABC$ and $\triangle A'BC'$ are similar.

EXERCISES

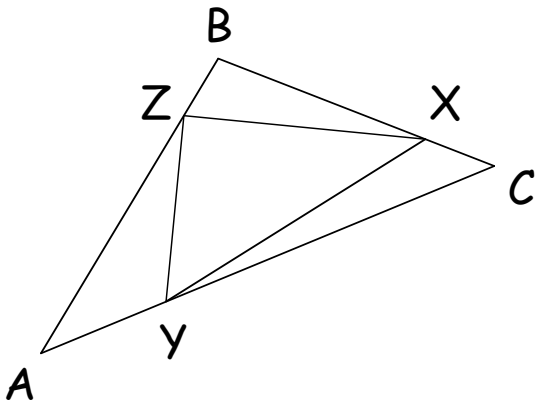
1. Let $\triangle ABC$ and $\triangle A'B'C'$ be given with $A\widehat{B}C = A'\widehat{B'}C'$ and $\dfrac{A'B'}{AB} = \dfrac{B'C'}{BC}$. Then $\triangle ABC \sim \triangle A'B'C'$.

2. In the figure to the right, $AD = rAB,\ AE = sAC$. Show that
$$\frac{\text{Area}\,\triangle ADE}{\text{Area}\,\triangle ABC} = rs.$$



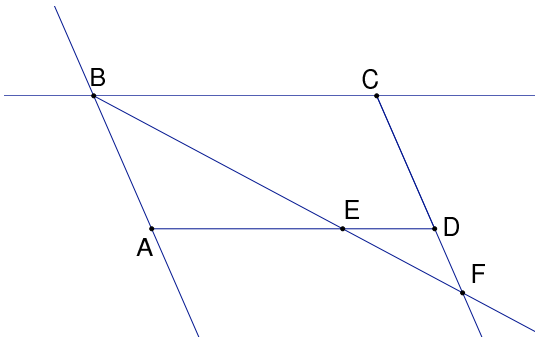
3. Let $\triangle ABC$ be a given triangle and let Y, Z be the midpoints of $[AC]$, $[AB]$, respectively. Show that (XY) is parallel with (AB). (This simple result is sometimes called the **Midpoint Theorem**)

4. In $\triangle ABC$, you are given that
$$\frac{AY}{YC} = \frac{CX}{XB} = \frac{BX}{ZA} = \frac{1}{x},$$
where x is a positive real number. Assuming that the area of $\triangle ABC$ is 1, compute the area of $\triangle XYZ$ as a function of x.

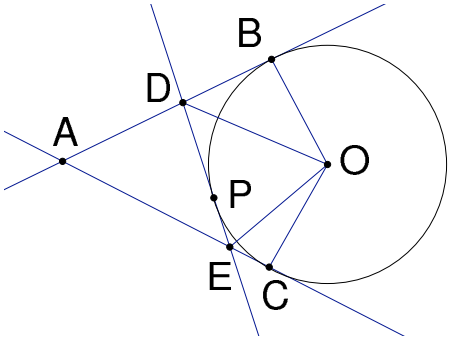


5. Let $ABCD$ be a quadrilateral and let $EFGH$ be the quadrilateral formed by connecting the midpoints of the sides of $ABCD$. Prove that $EFGH$ is a parallelogram.

6. In the figure to the right, $ABCD$ is a parallelogram, and E is a point on the segment $[AD]$. The point F is the intersection of lines (BE) and (CD). Prove that $AB \times FB = CF \times BE$.



7. In the figure to the right, tangents to the circle at B and C meet at the point A. A point P is located on the minor arc $\widehat{BC}$ and the tangent to the circle at P meets the lines (AB) and (AC) at the points D and E, respectively. Prove that $D\widehat{O}E = \frac{1}{2}B\widehat{O}C$, where O is the center of the given circle.



1.2.4 "Sensed" magnitudes; The Ceva and Menelaus theorems

In this subsection it will be convenient to consider the magnitude AB of the line segment $[AB]$ as "sensed,"[2] meaning that we shall regard AB as being either positive or negative and having absolute value equal to the usual magnitude of the line segment $[AB]$. The only requirement that we place on the signed magnitudes is that if the points A, B, and C are colinear, then

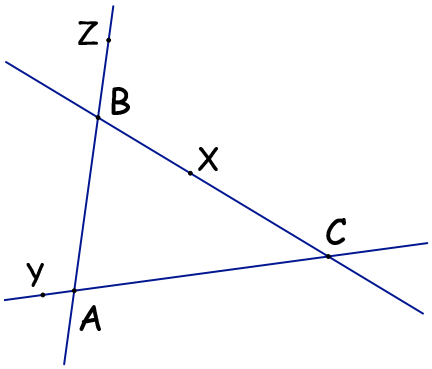
$$AB \times BC = \begin{cases} > 0 & \text{if } \overrightarrow{AB} \text{ and } \overrightarrow{BC} \text{ are in the same direction} \\ < 0 & \text{if } \overrightarrow{AB} \text{ and } \overrightarrow{BC} \text{ are in opposite directions.} \end{cases}$$

[2]IB uses the language "sensed" rather than the more customary "signed."

This implies in particular that for signed magnitudes,

$$\frac{AB}{BA} = -1.$$

Before proceeding further, the reader should pay special attention to the ubiquity of "dropping altitudes" as an auxiliary construction.



Both of the theorems of this subsection are concerned with the following configuration: we are given the triangle $\triangle ABC$ and points X, Y, and Z on the lines (BC), (AC), and (AB), respectively. Ceva's Theorem is concerned with the *concurrency* of the lines (AX), (BY), and (CZ). Menelaus' Theorem is concerned with the *colinearity* of the points X, Y, and Z. Therefore we may regard these theorems as being "dual" to each other.

In each case, the relevant quantity to consider shall be the product

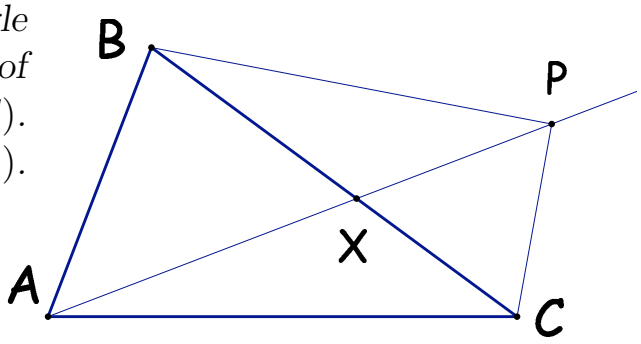
$$\frac{AZ}{ZB} \times \frac{BX}{XC} \times \frac{CY}{YA}$$

Note that each of the factors above is nonnegative precisely when the points X, Y, and Z lie on the segments $[BC]$, $[AC]$, and $[AB]$, respectively.

The proof of Ceva's theorem will be greatly facilitated by the following lemma:

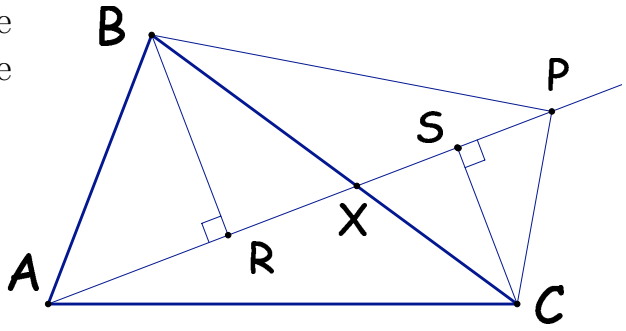
LEMMA. *Given the triangle* $\triangle ABC$*, let* X *be the intersection of a line through* A *and meeting* (BC)*. Let* P *be any other point on* (AX)*. Then*

$$\frac{\text{area } \triangle \text{APB}}{\text{area } \triangle \text{APC}} = \frac{BX}{CX}.$$



PROOF. In the diagram to the right, altitudes BR and CS have been constructed. From this, we see that

$$\begin{aligned}\frac{\text{area } \triangle \text{APB}}{\text{area } \triangle \text{APC}} &= \frac{\frac{1}{2}AP \cdot BR}{\frac{1}{2}AP \cdot CS} \\ &= \frac{BR}{CS} \\ &= \frac{BX}{CX},\end{aligned}$$

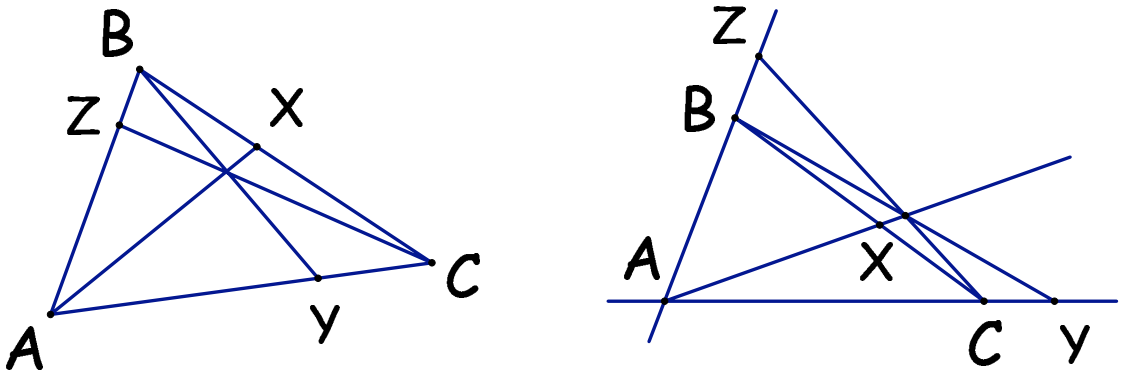


where the last equality follows from the obvious similarity $\triangle BRX \sim \triangle CSX$.

Note that the above proof doesn't depend on where the line (AP) intersects (BC), nor does it depend on the position of P relative to the line (BC), i.e., it can be on either side.

CEVA'S THEOREM. *Given the triangle* $\triangle ABC$*, lines (usually called* **Cevians** *are drawn from the vertices* A*,* B*, and* C*, with* X*,* Y*, and* Z*, being the points of intersections with the lines* (BC)*,* (AC)*, and* (AB)*, respectively. Then* (AX)*,* (BY)*, and* (CZ) *are concurrent if and only if*

$$\frac{AZ}{ZB} \times \frac{BX}{XC} \times \frac{CY}{YA} = +1.$$

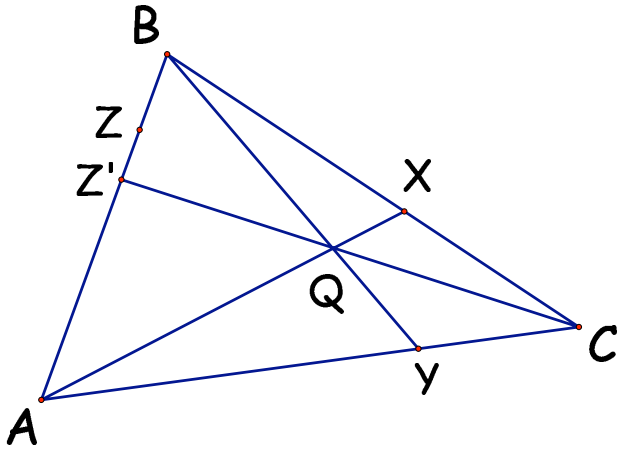


PROOF. Assume that the lines in question are concurrent, meeting in the point P. We then have, applying the above lemma three times, that

$$\begin{aligned} 1 &= \frac{\text{area}\,\triangle\text{APC}}{\text{area}\,\triangle\text{BPC}} \cdot \frac{\text{area}\,\triangle\text{APB}}{\text{area}\,\triangle\text{APC}} \cdot \frac{\text{area}\,\triangle\text{BPC}}{\text{area}\,\triangle\text{BPA}} \\ &= \frac{AZ}{ZB} \cdot \frac{BX}{XC} \cdot \frac{CY}{YA}. \end{aligned}$$

To prove the converse we need to prove that the lines (AX), (BY), and (CZ) are concurrent, given that

$$\frac{AZ}{ZB} \cdot \frac{BX}{XC} \cdot \frac{CY}{YZ} = 1.$$



Let $Q = (AX) \cap (BY)$, $Z' = (CQ) \cap (AB)$. Then (AX), (BY), and (CZ') are concurrent and so

$$\frac{AZ'}{Z'B} \cdot \frac{BX}{XC} \cdot \frac{CY}{YZ} = 1,$$

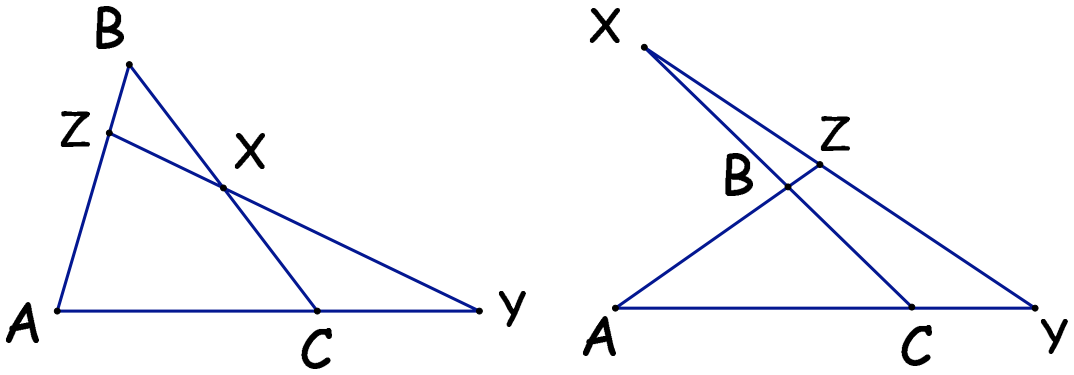
which forces

$$\frac{AZ'}{Z'B} = \frac{AZ}{ZB}.$$

This clearly implies that $Z = Z'$, proving that the original lines (AX), (BY), and (CZ) are concurrent.

Menelaus' theorem is a dual version of Ceva's theorem and concerns not **lines** (i.e., Cevians) but rather **points** on the (extended) edges of

the triangle. When these three points are collinear, the line formed is called a **transversal**. The reader can quickly convince herself that there are two configurations related to $\triangle ABC$:



As with Ceva's theorem, the relevant quantity is the product of the sensed ratios:

$$\frac{AZ}{ZB} \cdot \frac{BX}{XC} \cdot \frac{CY}{YA};$$

in this case, however, we see that either one or three of the ratios must be negative, corresponding to the two figures given above.

MENELAUS' THEOREM. *Given the triangle $\triangle ABC$ and given points X, Y, and Z on the lines (BC), (AC), and (AB), respectively, then X, Y, and Z are collinear if and only if*

$$\frac{AZ}{ZB} \times \frac{BX}{XC} \times \frac{CY}{YA} = -1.$$

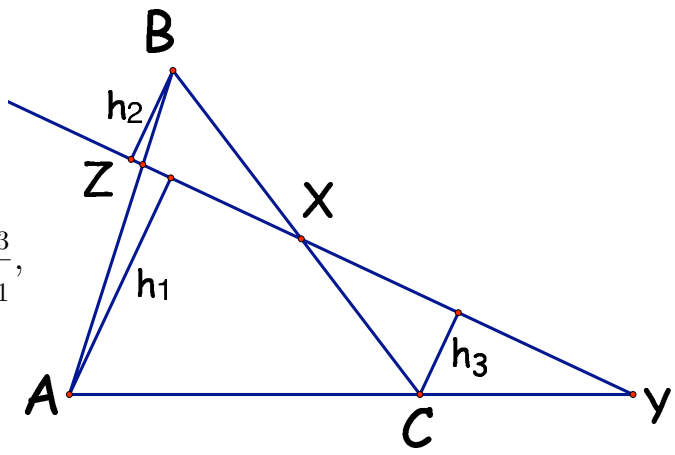
PROOF. As indicated above, there are two cases to consider. The first case is that in which two of the points X, Y, or Z are on the triangle's sides, and the second is that in which none of X, Y, or Z are on the triangle's sides. The proofs of these cases are formally identical, but for clarity's sake we consider them separately.

CASE 1. We assume first that X, Y, and Z are collinear and drop altitudes h_1, h_2, and h_3 as indicated in the figure to the right. Using obvious similar triangles, we get

$$\frac{AZ}{ZB} = +\frac{h_1}{h_2};\ \frac{BX}{XC} = +\frac{h_2}{h_3};\ \frac{CY}{YA} = -\frac{h_3}{h_1},$$

in which case we clearly obtain

$$\frac{AZ}{ZB} \times \frac{BX}{XC} \times \frac{CY}{YA} = -1.$$



To prove the converse, we may assume that X is on $[BC]$, Z is on $[AB]$, and that Y is on (AC) with $\frac{AZ}{ZB} \cdot \frac{BX}{XC} \cdot \frac{CY}{YA} = -1$. We let X' be the intersection of (ZY) with $[BC]$ and infer from the above that

$$\frac{AZ}{ZB} \cdot \frac{BX'}{X'C} \cdot \frac{CY}{YA} = -1.$$

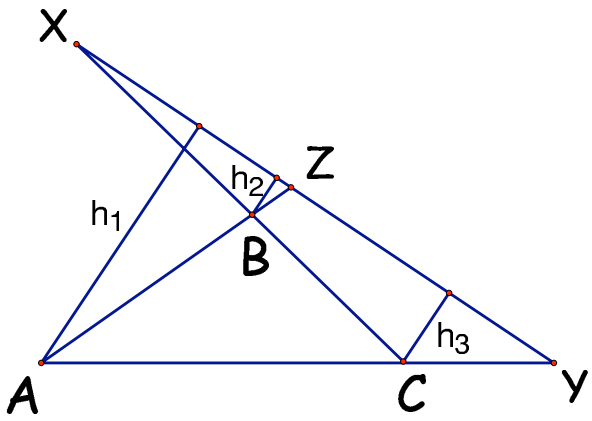
It follows that $\frac{BX}{XC} = \frac{BX'}{X'C}$, from which we infer easily that $X = X'$, and so X, Y, and Z are collinear.

CASE 2. Again, we drop altitudes from A, B, and C and use obvious similar triangles, to get

$$\frac{AZ}{ZB} = -\frac{h_1}{h_2};\ \frac{BX}{XC} = -\frac{h_2}{h_3};\ \frac{AY}{YC} = -\frac{h_1}{h_3};$$

it follows immediately that

$$\frac{AZ}{ZB} \cdot \frac{BX}{XC} \cdot \frac{CY}{YA} = -1.$$



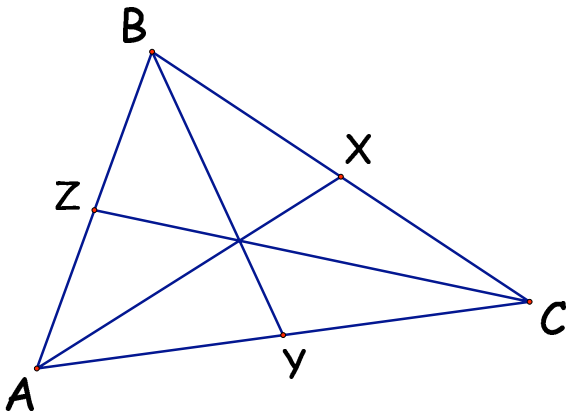
The converse is proved exactly as above.

1.2.5 Consequences of the Ceva and Menelaus theorems

As one typically learns in an elementary geometry class, there are several notions of "center" of a triangle. We shall review them here and show their relationships to Ceva's Theorem.

Centroid. In the triangle $\triangle ABC$ lines (AX), (BY), and (CZ) are drawn so that (AX) bisects $[BC]$, (BY) bisects $[CA]$, and (CZ) bisects $[AB]$ That the lines (AX), (BY), and (CZ) are concurrent immediately follows from Ceva's Theorem as one has that

$$\frac{AZ}{ZB}\cdot\frac{BX}{XC}\cdot\frac{CY}{YZ}=1\times 1\times 1=1.$$



The point of concurrency is called the **centroid** of $\triangle ABC$. The three Cevians in this case are called **medians**.

Next, note that if we apply the Menelaus' theorem to the triangle $\triangle ACX$ and the transversal defined by the points B, Y and the centroid P, then we have that

$$1=\frac{AY}{YC}\cdot\frac{CB}{BX}\cdot\frac{XP}{PA}\Rightarrow$$

$$1=1\cdot 2\cdot\frac{XP}{PA}\Rightarrow\frac{XP}{PA}=\frac{1}{2}.$$

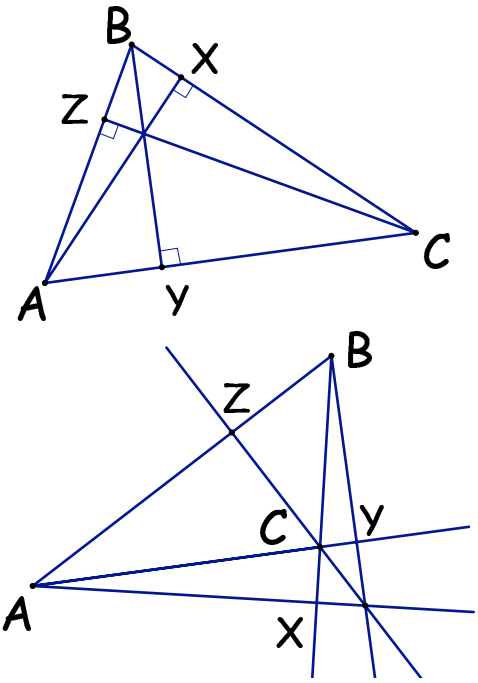
Therefore, we see that the distance of a triangle's vertex to the centroid is exactly 1/3 the length of the corresponding median.

Orthocenter. In the triangle $\triangle ABC$ lines (AX), (BY), and (CZ) are drawn so that $(AX) \perp (BC)$, $(BY) \perp (CA)$, and $(CZ) \perp (AB)$. Clearly we either have

$$\frac{AZ}{ZB}, \frac{BX}{XC}, \frac{CY}{YA} > 0$$

or that exactly one of these ratios is positive. We have

$$\triangle ABY \sim \triangle ACZ \Rightarrow \frac{AZ}{AY} = \frac{CZ}{BY}.$$



Likewise, we have

$$\triangle ABX \sim \triangle CBZ \Rightarrow \frac{BX}{BZ} = \frac{AX}{CZ} \text{ and } \triangle CBY \sim \triangle CAX$$

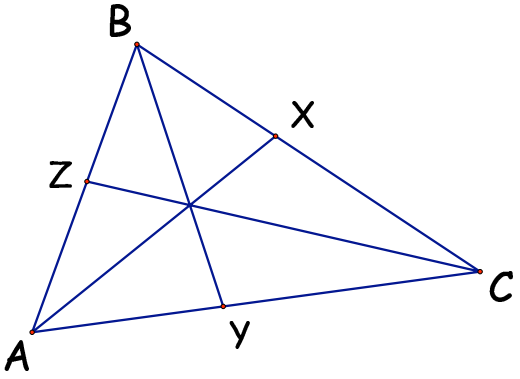
$$\Rightarrow \frac{CY}{CX} = \frac{BY}{AX}.$$

Therefore,

$$\frac{AZ}{ZB} \cdot \frac{BX}{XC} \cdot \frac{CY}{YA} = \frac{AZ}{AY} \cdot \frac{BX}{BZ} \cdot \frac{CY}{CX} = \frac{CZ}{BY} \cdot \frac{AX}{CZ} \cdot \frac{BY}{AX} = 1.$$

By Ceva's theorem the lines (AX), (BY), and (CZ) are concurrent, and the point of concurrency is called the **orthocenter** of $\triangle ABC$. (The line segments $[AX]$, $[BY]$, and $[CZ]$ are the **altitudes** of $\triangle ABC$.)

Incenter. In the triangle $\triangle ABC$ lines (AX), (BY), and (CZ) are drawn so that (AX) bisects $B\widehat{A}C$, (BY) bisects $A\widehat{B}C$, and (CZ) bisects $B\widehat{C}A$ As we show below, that the lines (AX), (BY), and (CZ) are concurrent; the point of concurrency is called the **incenter** of $\triangle ABC$. (A very interesting "extremal"

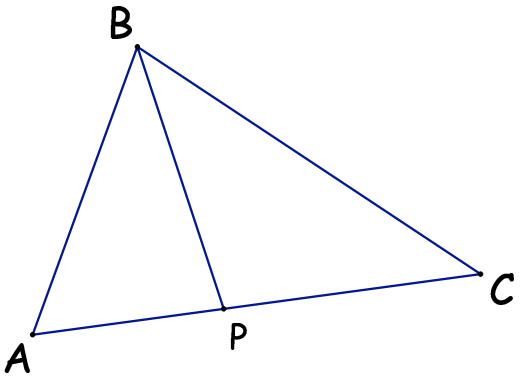


property of the incenter will be given in
Exercise 12 on page 153.) However, we shall proceed below to give another proof of this fact, based on Ceva's Theorem.

Proof that the angle bisectors of $\triangle ABC$ are concurrent. In order to accomplish this, we shall first prove the

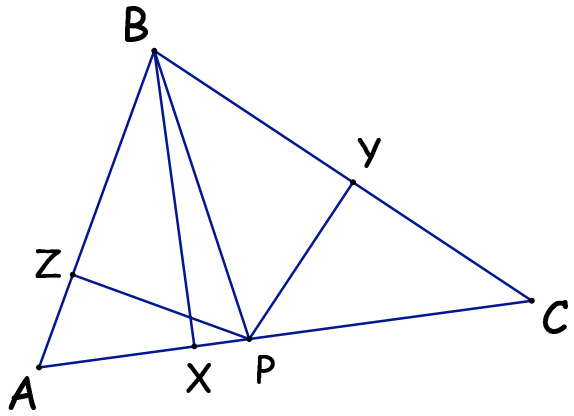
ANGLE BISECTOR THEOREM. *We are given the triangle $\triangle ABC$ with line segment $[BP]$ (as indicated to the right). Then*

$$\frac{AB}{BC} = \frac{AP}{PC} \Leftrightarrow A\widehat{B}P = P\widehat{B}C.$$



PROOF ($\Leftarrow$). We drop altitudes from P to (AB) and (BC); call the points so determined Z and Y, respectively. Drop an altitude from B to (AC) and call the resulting point X. Clearly $PZ = PY$ as $\triangle PZB \cong \triangle PYB$. Next, we have

$$\triangle ABX \sim \triangle APZ \Rightarrow \frac{AB}{AP} = \frac{BX}{PZ} = \frac{BX}{PY}.$$



Likewise,

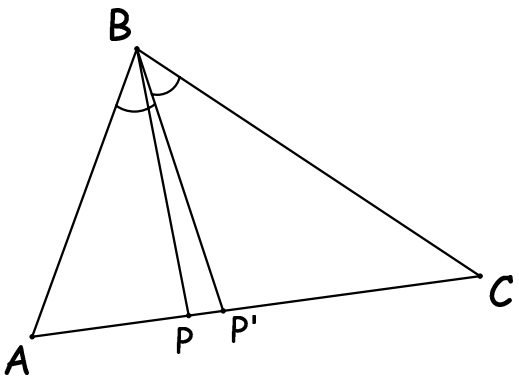
$$\triangle CBX \sim \triangle CPY \Rightarrow \frac{CB}{CP} = \frac{BX}{PY}.$$

Therefore,

$$\frac{AB}{BC} = \frac{AP \cdot BX}{PY} \cdot \frac{PY}{CP \cdot BX} = \frac{AP}{CP}.$$

($\Rightarrow$). Here we're given that $\frac{AB}{BC} = \frac{AP}{PC}$. Let P' be the point determined by the angle bisector (BP') of $A\widehat{B}C$. Then by what has already been proved above, we have $\frac{AP}{BC} = \frac{AP'}{P'C}$. But this implies that

$$\frac{AP}{PC} = \frac{AP'}{P'C} \Rightarrow P = P'.$$



Conclusion of the proof that angle bisectors are concurrent. First of all, it is clear that the relevant ratios are all positive. By the Angle Bisector Theorem,

$$\frac{AB}{BC} = \frac{AY}{YC},\ \frac{BC}{CA} = \frac{BZ}{ZA},\ \frac{AB}{AC} = \frac{BX}{XC};$$

therefore,

$$\frac{AZ}{BZ} \times \frac{BX}{XC} \times \frac{CY}{YA} = \frac{CA}{BC} \times \frac{AB}{AC} \times \frac{BC}{AB} = 1.$$

Ceva's theorem now finishes the job!

EXERCISES

1. The Angle Bisector Theorem involved the bisection of one of the given triangle's **interior** angles. Now let P be a point on the line (AC) **external** to the segment $[AC]$. Show that the line (BP) bisects the external angle at B if and only if

$$\frac{AB}{BC} = \frac{AP}{PC}.$$

2. You are given the triangle $\triangle ABC$. Let X be the point of intersection of the bisector of $B\widehat{A}C$ with $[BC]$ and let Y be the point of intersection of the bisector of $C\widehat{B}A$ with $[AC]$. Finally, let Z be the point of intersection of the *exterior* angle bisector at C with the line (AB). Show that X, Y, and Z are colinear.[3]

[3]What happens if the exterior angle bisector at C is parallel with (AB)?

3. Given $\triangle ABC$ and assume that X is on (BC), Y is on (AC) and Z is on (AB). Assume that the Cevians (AX) (BY), and (CZ) are concurrent, meeting at the point P. Show that

$$\frac{PX}{AX} + \frac{PY}{BY} + \frac{PZ}{CZ} = 1.$$

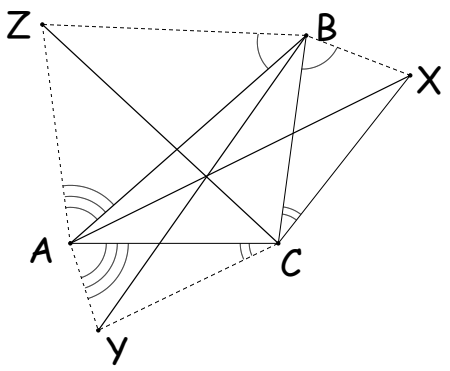
4. Given the triangle $\triangle ABC$ with incenter P, prove that there exists a circle $\mathcal{C}$ (called the **incircle** of $\triangle ABC$) with center P which is inscribed in the triangle $\triangle ABC$. The radius r of the incircle is often called the **inradius** of $\triangle ABC$.

5. Let $\triangle ABC$ have side lengths $a = BC$, $b = AC$, and $c = AB$, and let r be the inradius. Show that the area of $\triangle ABC$ is equal to $\frac{r(a+b+c)}{2}$. (Hint: the incenter partitions the triangle into three smaller triangles; compute the areas of each of these.)

6. Given the triangle $\triangle ABC$. Show that the bisector of the **internal** angle bisector at A and the bisectors of the **external** angles at B and C are concurrent.

7. Given $\triangle ABC$ and points X, Y, and Z in the plane such that

$$\begin{aligned} \angle ABZ &= \angle CBX, \\ \angle BCX &= \angle ACY, \\ \angle BAZ &= \angle CAY. \end{aligned}$$



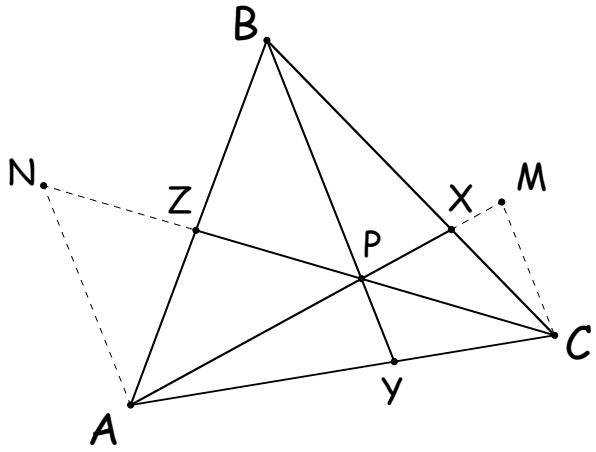
Show that (AX), (BY), and (CZ) are concurrent.

8. There is another notion of "center" of the triangle $\triangle ABC$. Namely, construct the lines l_1, l_2, and l_3 so as to be perpendicular bisectors of $[AB]$, $[BC]$, and $[CA]$, respectively. After noting that Ceva's theorem doesn't apply to this situation, prove directly that the lines l_1, l_2, and l_3 are concurrent. The point of concurrency is called the **circumcenter** of $\triangle ABC$. (Hint: argue that the point of concurrency of two of the perpendicular bisectors is equidistant to all three of the vertices.) If P is the circumcenter, then the common value $AP = BP = CP$ is called the **circumradius**

of the triangle $\triangle ABC$. (This is because the circumscribed circle containing A, B, and C will have radius AP.)

9. $\triangle ABC$ has side lengths $AB = 21$, $AC = 22$, and $BC = 20$. Points D and E are on sides $[AB]$ and $[AC]$, respectively such that $[DE] \parallel [BC]$ and $[DE]$ passes through the incenter of $\triangle ABC$. Compute DE.

10. Here's another proof of Ceva's theorem. You are given $\triangle ABC$ and concurrent Cevians $[AX]$, $[BY]$, and $[CZ]$, meeting at the point P. Construct the line segments $[AN]$ and $[CM]$, both parallel to the Cevian $[BY]$. Use similar triangles to conclude that



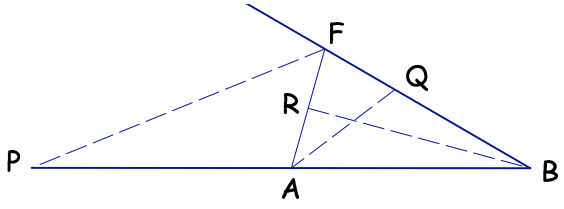
$$\frac{AY}{YC} = \frac{AN}{CM}, \frac{CX}{XB} = \frac{CM}{BP}, \frac{BZ}{ZA} = \frac{BP}{AN},$$

and hence that $\dfrac{AZ}{ZB} \cdot \dfrac{BX}{XC} \cdot \dfrac{CY}{YA} = 1$.

11. Through the vertices of the triangle $\triangle PQR$ lines are drawn lines which are parallel to the opposite sides of the triangle. Call the new triangle $\triangle ABC$. Prove that these two triangles have the same centroid.

12. Given the triangle $\triangle ABC$, let $\mathcal{C}$ be the inscribed circle, as in Exercise 4, above. Let X, Y, and Z be the points of tangency of $\mathcal{C}$ (on the sides $[BC]$, $[AC]$, $[AB]$, respectively) and show that the lines (AX), (BY), and (CZ) are concurrent. The point of concurrency is called the **Gergonne point** of the circle $\mathcal{C}$. (This is very easy once you note that $AZ = YZ$, etc.!)

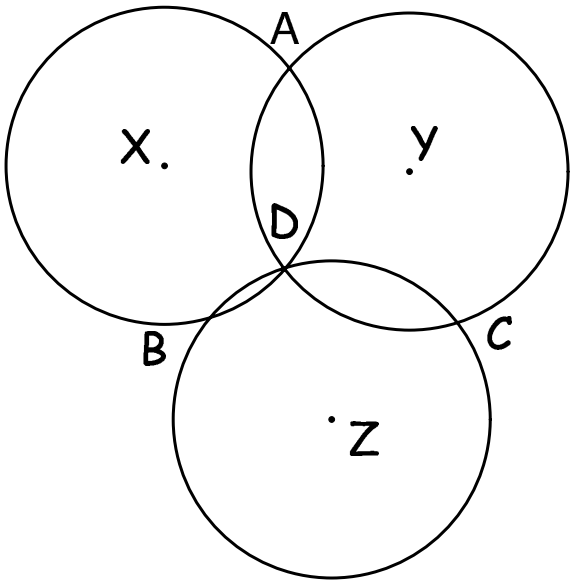
13. In the figure to the right, the dotted segments represent angle bisectors. Show that the points P, R, and Q are colinear.



14. In the figure to the right, three circies of the same radius and centers X, Y and Z are shown intersecting at points A, B, C, and D, with D the common point of intersection of all three circles.
Show that

    (a) D is the circumcenter of $\triangle XYZ$, and that

    (b) D is the orthocenter of $\triangle ABC$. (Hint: note that $YZCD$ is a rhombus.)



15. Show that the three medians of a triangle divide the triangle into six triangle of equal area.

16. Let the triangle $\triangle ABC$ be given, and let A' be the midpoint of $[BC]$, B' the midpoint of $[AC]$ and let C' be the midpoint of $[AB]$. Prove that

    (i) $\triangle A'B'C' \sim \triangle ABC$ and that the ratios of the corresponding sides are 1:2.

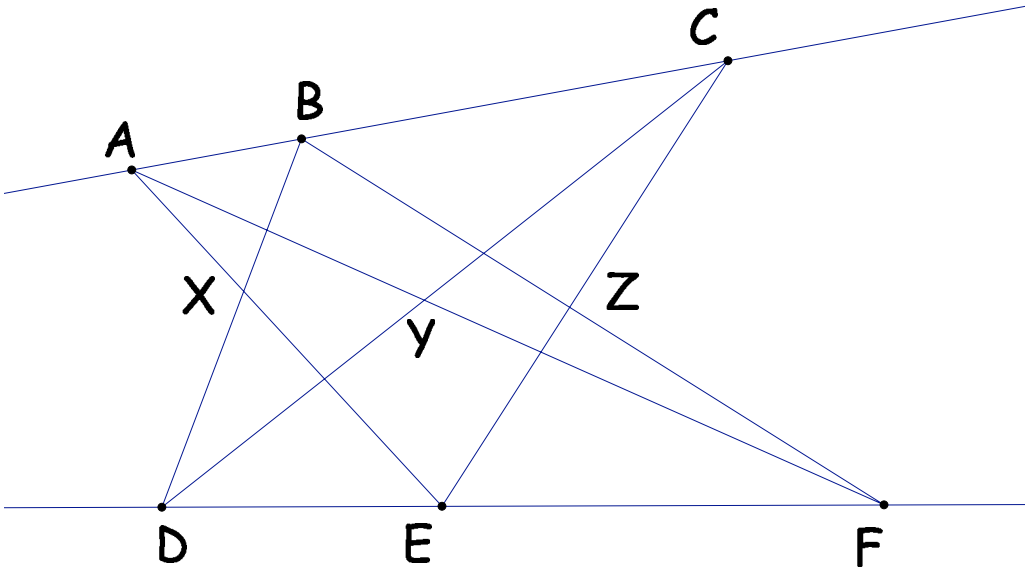
    (ii) $\triangle A'B'C'$ and $\triangle ABC$ have the same centroid.

    (iii) The four triangles determined within $\triangle ABC$ by $\triangle A'B'C'$ are all congruent.

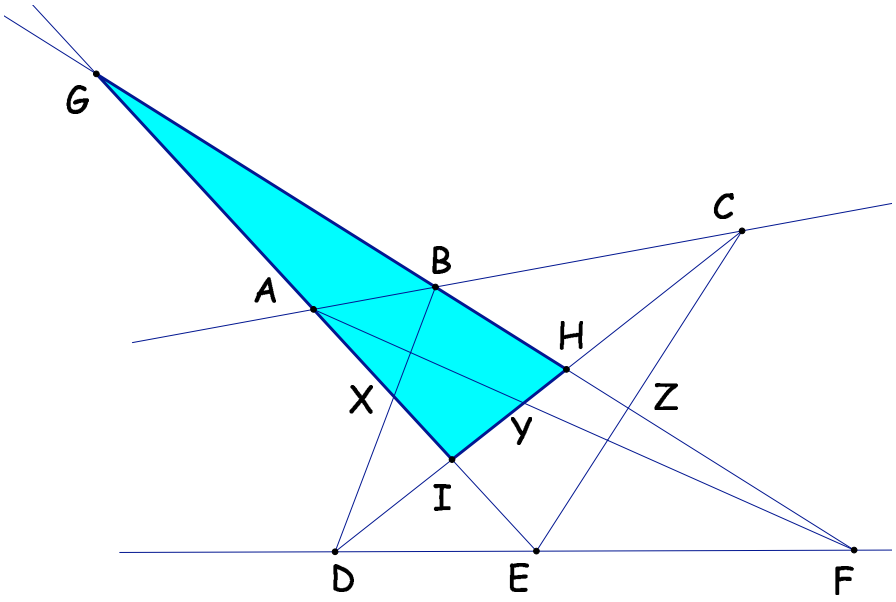
    (iv) The circumcenter of $\triangle ABC$ is the orthocenter of $\triangle A'B'C'$.

    The triangle $\triangle A'B'C'$ of $\triangle ABC$ formed above is called the **medial triangle** of $\triangle ABC$.

17. The figure below depicts a hexagram "inscribed" in two lines. Using the prompts given, show that the lines X, Y, and Z are colinear. This result is usually referred to **Pappus' theorem**.



Step 1. Locate the point G on the lines (AE) and (FB); we shall analyze the triangle $\triangle GHI$ as indicated below.[4]



Step 2. Look at the transversals, applying Menelaus' theorem to each:

[4]Of course, it may be that (AE) and (FB) are parallel. In fact, it may happen that all analogous choices for pairs of lines are parallel, which would render the present theme invalid. However, while the present approach uses Menelaus' theorem, which is based on "metrical" ideas, Pappus' theorem is a theorem only about incidence and colinearity, making it really a theorem belonging to "projective geometry." As such, if the lines (AE) and (BF) were parallel, then projectively they would meet "at infinity;" one could then apply a projective transformation to move this point at infinity to the finite plane, preserving the colinearity of X, Y, and Z

$[DXB]$, so $\dfrac{GX}{XI}\dfrac{ID}{DH}\dfrac{HB}{BG} = -1.$

$[AYF]$, so $\dfrac{GA}{AI}\dfrac{IY}{YH}\dfrac{HF}{FG} = -1.$

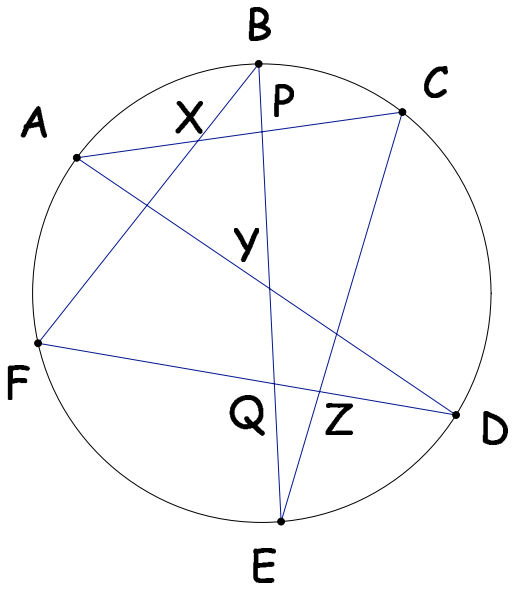
$[CZE]$ (etc.)

$[ABC]$ (etc.)

$[DEF]$ (etc)

Step 3. Multiply the above five factorizations of -1, cancelling out all like terms!

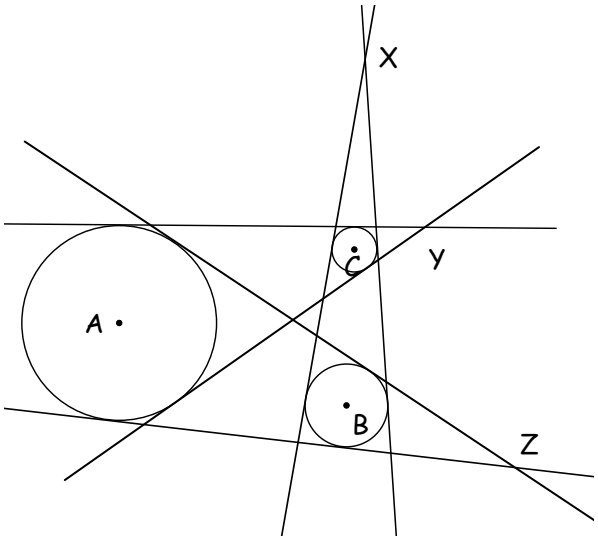
18. This time, let the hexagram be inscribed in a circle, as indicated to the right. By producing edges $[AC]$ and $[FD]$ to a common point R and considering the triangle $\triangle PQR$ prove **Pascal's theorem**, namely that points X, Y, and Z are colinear. (Proceed as in the proof of Pappus' theorem: consider the transversals $[BXF]$, $[AYD]$, and $[CZE]$, multiplying together the factorizations of -1 which each produces.)



19. A straight line meets the sides $[PQ]$, $[QR]$, $[RS]$, and $[SP]$ of the quadrilateral $PQRS$ at the points U, V, W, and X, respectively. Use Menelaus' theorem to show that

$$\frac{PU}{UQ} \times \frac{QV}{VR} \times \frac{RW}{WS} \times \frac{SX}{XP} = 1.$$

20. The diagram to the right shows three circles of different radii with centers A, B, and C. The points X, Y, and Z are defined by intersections of the tangents to the circles as indicated. Prove that X, Y, and Z are colinear.



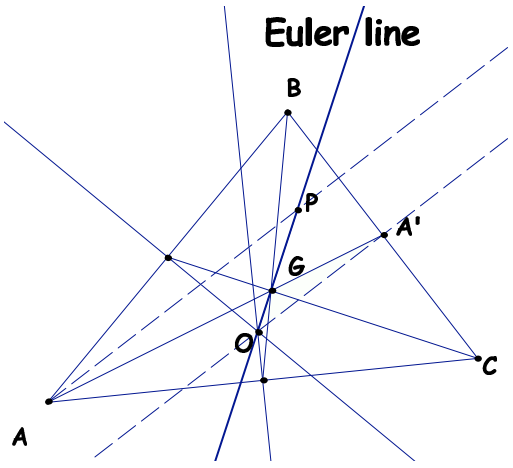
21. (**The Euler line.**) In this exercise you will be guided through the proof that in the triangle $\triangle ABC$, the centroid, circumcenter, and orthocenter are all colinear. The line so determined is called the **Euler line**.

    In the figure to the right, let G be the centroid of $\triangle ABC$, and let O be the circumcenter. Locate P on the ray $\overrightarrow{OG}$ so that $GP : OG = 2 : 1$.

    (a) Let A' be the intersection of (AG) with (BC); show that $\triangle OGA' \sim \triangle PGA$. (Hint: recall from page 13 that $GA : GA' = 2 : 1$.)

    (b) Conclude that (AP) and (OA') are parallel which puts P on the altitude through vertex A.

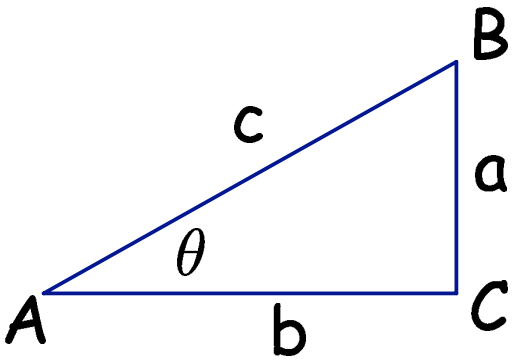
    (c) Similarly, show that P is also on the altitudes through vertices B and C, and so P is the orthocenter of $\triangle ABC$.



1.2.6 Brief interlude: laws of sines and cosines

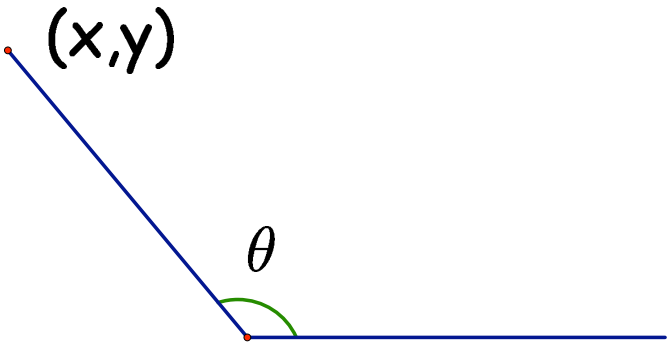
In a right triangle $\triangle ABC$, where $\widehat{C}$ is a right angle, we have the familiar **trigonometric ratios**: setting $\theta = \widehat{A}$, we have

$$\sin\theta = \frac{a}{c}, \quad \cos\theta = \frac{b}{c};$$



the remaining trigonometric ratios ($\tan\theta$, $\csc\theta$, $\sec\theta$, $\cot\theta$) are all expressable in terms of $\sin\theta$ and $\cos\theta$ in the familiar way. **Of crucial importance here is the fact that by similar triangles, these ratios depend only on θ an not on the particular choices of side lengths.**[5]

We can extend the definitions of the trigonometric functions to arbitrary angles using coordinates in the plane. Thus, if θ is any given angle relative to the positive x-axis (whose measure can be anywhere between $-\infty$ and ∞ degrees, and if (x, y) is any point on the terminal ray, then we set



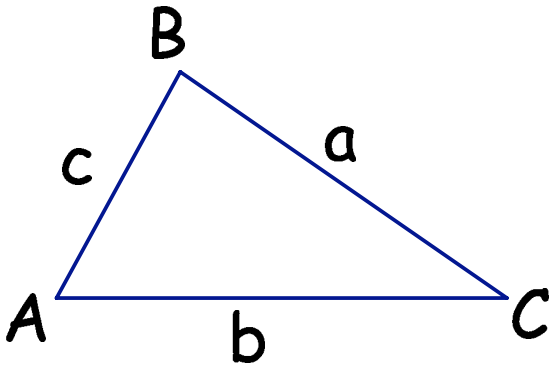
$$\sin\theta = \frac{y}{\sqrt{x^2+y^2}}, \quad \cos\theta = \frac{x}{\sqrt{x^2+y^2}}.$$

Notice that on the basis of the above definition, it is obvious that $\sin(180-\theta) = \sin\theta$ and that $\cos(180-\theta) = -\cos\theta$. Equally important (and obvious!) is the **Pythagorean identity**: $\sin^2\theta + \cos^2\theta = 1$.

[5] A fancier way of expressing this is to say that by similar triangles, the trigonometric functions are **well defined**.

LAW OF SINES. *Given triangle* $\triangle ABC$ *and sides* a, b, *and* c, *as indicated, we have*

$$\frac{\sin A}{a} = \frac{\sin B}{b} = \frac{\sin C}{c}.$$



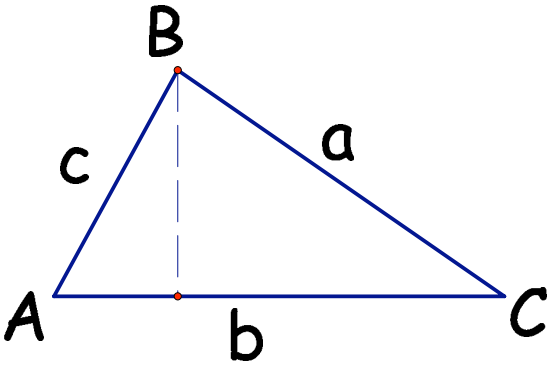
PROOF. We note that

$$\frac{1}{2}bc\sin A = \text{area}\,\triangle ABC = \frac{1}{2}ba\sin C,$$

and so

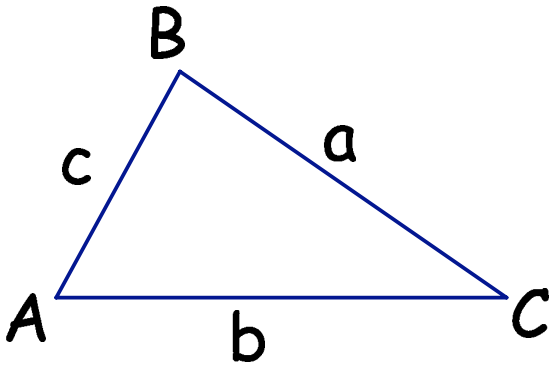
$$\frac{\sin A}{a} = \frac{\sin C}{c}.$$

A similar argument shows that $\dfrac{\sin B}{b}$ is also equal to the above.

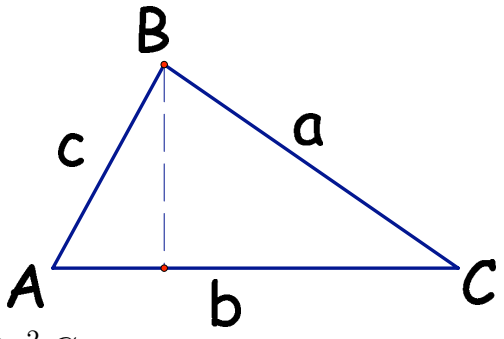


LAW OF COSINES. *Given triangle* $\triangle ABC$ *and sides* a, b, *and* c, *as indicated, we have*

$$c^2 = a^2 + b^2 - 2ab\cos C.$$



PROOF. Referring to the diagram to the right and using the Pythagorean Theorem, we infer quickly that



$$\begin{aligned} c^2 &= (b - a\cos C)^2 + a^2 \sin^2 C \\ &= b^2 - 2ab\cos C + a^2\cos^2 C + a^2 \sin^2 C \\ &= a^2 + b^2 - 2ab\cos C, \end{aligned}$$

as required.

EXERCISES

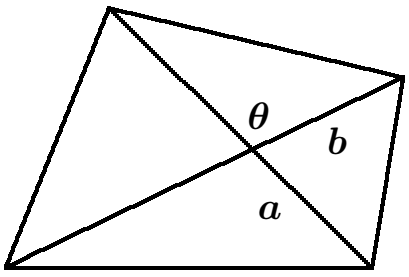
1. Using the Law of Sines, prove the Angle Bisector Theorem (see page 15).

2. Prove **Heron's formula**. Namely, for the triangle whose side lengths are a, b, and c, prove that the area is given by

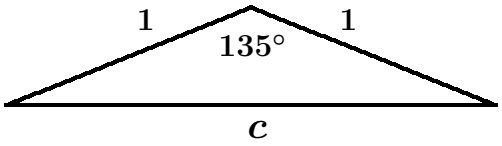
$$\text{area} = \sqrt{s(s-a)(s-b)(s-c)},$$

where $s = \dfrac{a+b+c}{2}$ = one-half the perimeter of the triangle. (Hint: if A is the area, then start with $16A^2 = 4b^2(c^2 - c^2\cos^2 A) = (2bc - 2bc\cos A)(2bc + 2bc\cos A)$. Now use the Law of Cosines to write $2bc\cos A$ in terms of a, b, and c and do a bit more algebra.)

3. In the quadrilateral depicted at the right, the lengths of the diagonals are a and b, and meet at an angle θ. Show that the area of this quadrilateral is $\frac{1}{2}ab\sin\theta$. (Hint: compute the area of each triangle, using the Law of Sines.)



4. In the triangle to the right, show that $c = \dfrac{\sqrt{1+i}+\sqrt{1-i}}{\sqrt[4]{2}}$ (where $i^2 = -1$)



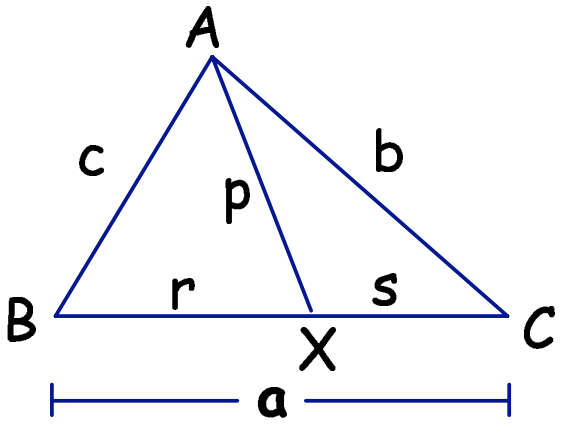
5. Given $\triangle ABC$ with C a right angle, let D be the midpoint of $[AB]$ and show that $\triangle ADC$ is isosceles with $AD = DC$.

6. Given $\triangle ABC$ with $BC = a$, $CA = b$, and $AB = c$. Let D be the midpoint of $[BC]$ and show that $AD = \frac{1}{2}\sqrt{2(b^2+c^2)-a^2}$.

1.2.7 Algebraic results; Stewart's theorem and Apollonius' theorem

STEWART'S THEOREM. *We are given the triangle $\triangle ABC$, together with the edge BX, as indicated in the figure to the right. Then*

$$a(p^2+rs) = b^2r + c^2s.$$



PROOF. We set $\theta = A\widehat{B}C$; applying the Law of Cosines to $\triangle AXB$ yields

$$\cos\theta = \frac{r^2+p^2-c^2}{2pr}.$$

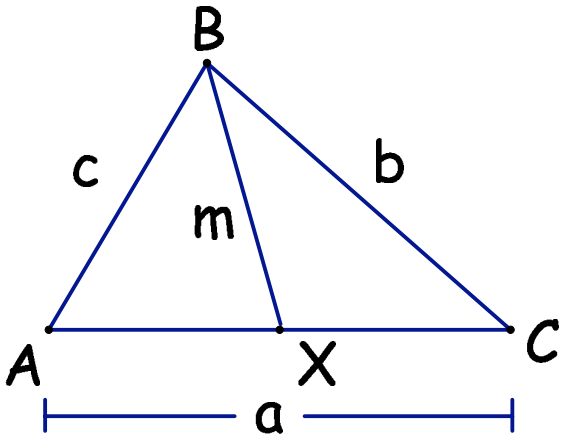
Applying the Law of Cosines to the triangle $\triangle BXC$ gives

$$\cos\theta \;=\; \frac{b^2 - s^2 - p^2}{2ps}.$$

Equating the two expressions and noting that $a = r+s$ eventually leads to the desired result.

COROLLARY [APOLLONIUS THEOREM]. *We are given the triangle $\triangle ABC$, with sides a, b, and c, together with the median BX, as indicated in the figure to the right. Then*

$$b^2 + c^2 = 2m^2 + a^2/2.$$



If $b = c$ (the triangle is isosceles), then the above reduces to

$$m^2 + (a/2)^2 = b^2.$$

This follows instantly from Stewart's Theorem.

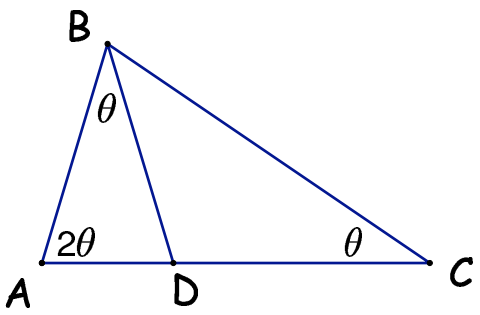
EXERCISES

1. Assume that the sides of a triangle are 4, 5, and 6.

   (a) Compute the area of this triangle.

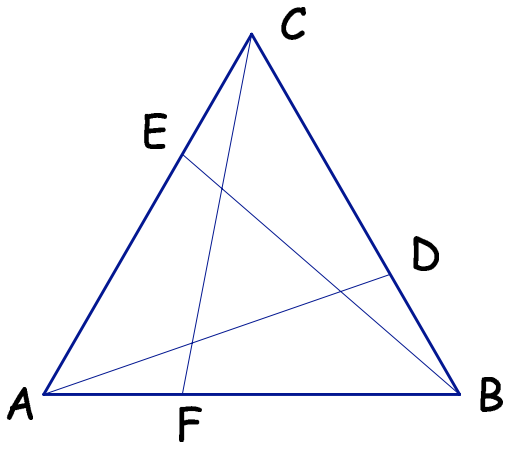
   (b) Show that one of the angles is twice one of the other angles.

2. **(The Golden Triangle)** You are given the triangle depicted to the right with $\triangle ABD \sim \triangle BCA$ Show that $\dfrac{DC}{AD} = \dfrac{\sqrt{5}+1}{2}$, the **golden ratio**.



3. Let $\triangle ABC$ be given with sides $a = 11$, $b = 8$, and $c = 8$. Assume that D and E are on side $[BC]$ such that $[AD]$, $[AE]$ trisect $B\widehat{A}C$. Show that $AD = AE = 6$.

4. You are given the equilateral triangle with sides of unit length, depicted to the right. Assume also that $AF = BD = CE = r$ for some positive $r < 1$. Compute the area of the inner equilateral triangle. (Hint: try using similar triangles and Stewart's theorem to compute $AD = BE = CF$.)

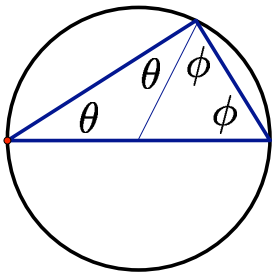


1.3 Circle Geometry

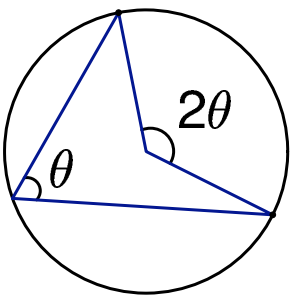
1.3.1 Inscribed angles

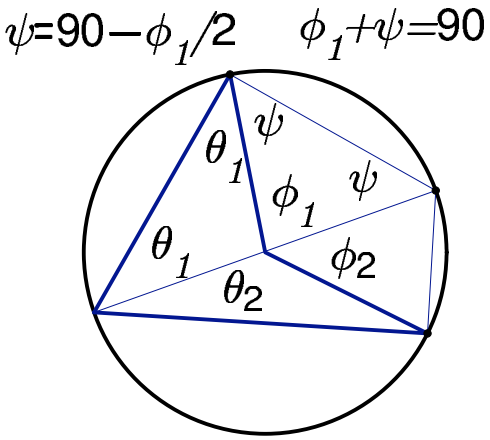
LEMMA. *If a triangle $\triangle ABC$ is inscribed in a circle with $[AB]$ being a diameter, then $A\widehat{C}B$ is a right angle.*

PROOF. The diagram to the right makes this obvious; from $2\theta + 2\phi = 180$, we get $\theta + \phi = 90°$.

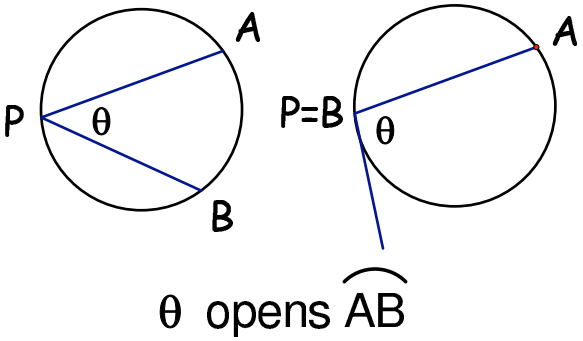


INSCRIBED ANGLE THEOREM. *The measure of an angle inscribed in a circle is one-half that of the inscribed arc.*



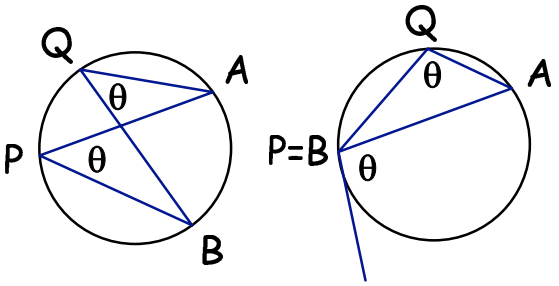


PROOF. We draw a diameter, as indicated; from the above lemma, we see that $\theta_1 + \psi = 90$. This quickly leads to $\phi_1 = 2\theta_1$. Similarly $\phi_2 = 2\theta_2$, and we're done.



Before proceeding, we shall find the following concept useful. We are given a circle and points A, B, and P on the circle, as indicated to the right. We shall say that the angle $A\widehat{P}B$ **opens** the arc $\widehat{AB}$. A degenerate instance of this is when B and P agree, in which case a tangent occurs. In this case we shall continue to say that the given angle **opens** the arc $\widehat{AB}$.

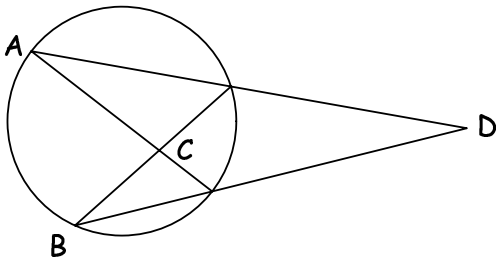
As an immediate corollary to the Inscribed Angle Theorem, we get the following:



COROLLARY. *Two angles which open the same are are equal.*

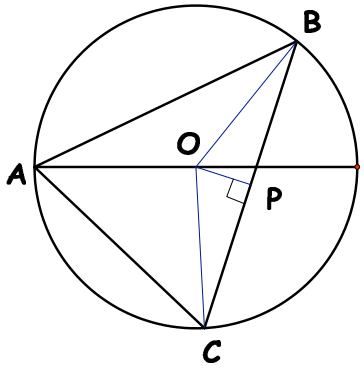
EXERCISES

1. In the diagram to the right, the arc $\widehat{AB}$ has a measure of 110° and the measure of the angle $A\widehat{C}B$ is 70°. Compute the measure of $A\widehat{D}B$.[6]

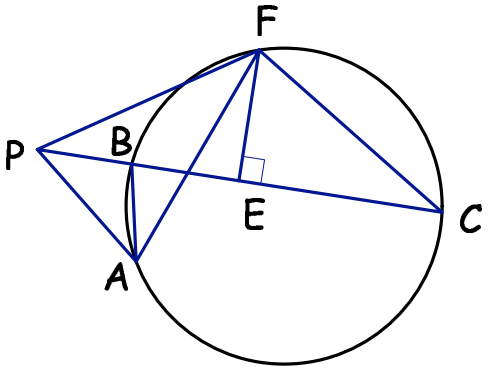


2. Let $[AB]$ be a diameter of the circle $\mathcal{C}$ and assume that C is a given point. If $A\widehat{C}B$ is a right angle, then C is on the circle $\mathcal{C}$.

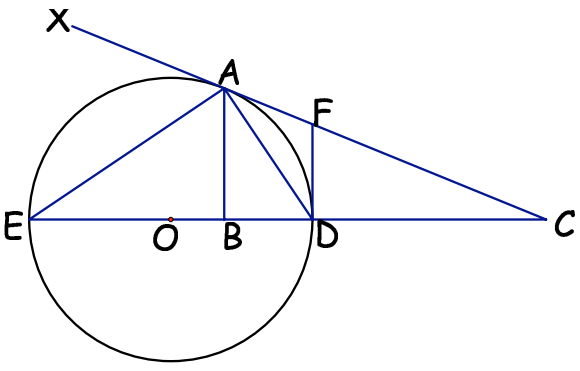
3. Let $\mathcal{C}$ be a circle having center O and diameter d, and let A, B, and C be points on the circle. If we set $\alpha = B\widehat{A}C$, then $\sin\alpha = BC/d$. (Hint: note that by the inscribed angle theorem, $B\widehat{A}C = P\widehat{O}C$. What is the sine of $P\widehat{O}C$?)



4. In the given figure $AF = FC$ and $PE = EC$.
   (a) Prove that triangle $\triangle FPA$ is isosceles.
   (b) Prove that $AB + BE = EC$.



5. A circle is given with center O. The points E, O, B, D, and E are collinear, as are X, A, F, and C. The lines (XC) and (FD) are tangent to the circle at the points A and D respectively. Show that
   (a) (AD) bisects $B\widehat{A}C$;
   (b) (AE) bisects $B\widehat{A}X$.



6. Let $\triangle ABC$ have circumradius R. Show that

$$\text{Area}\,\triangle ABC = \frac{R(a\cos A + b\cos B + c\cos B)}{2},$$

where $a = BC$, $b = AC$, and $c = AB$. (See exercise 5, page 17 for the corresponding result for the inscribed circle.)

Circle of Apollonius

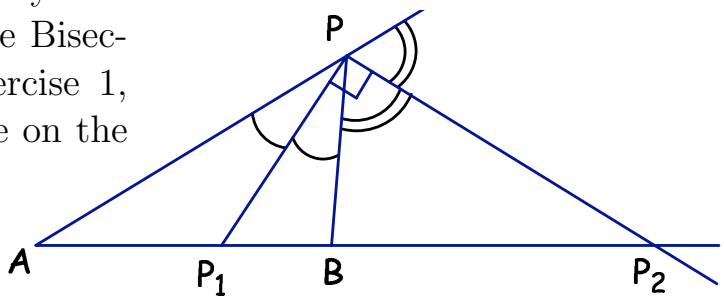
CIRCLE OF APOLLONIUS. *Assume that $c \neq 1$ is a constant and that A and B are two given points. Then the locus of points*

$$\left\{P \,\middle|\, \frac{PA}{PB} = c\right\}$$

is a circle.

PROOF. This is actually a very simple application of the Angle Bisector Theorem (see also Exercise 1, page 16). Let P_1 and P_2 lie on the line (AB) subject to

$$\frac{AP_1}{P_1B} = c = \frac{AP_2}{BP_2}.$$



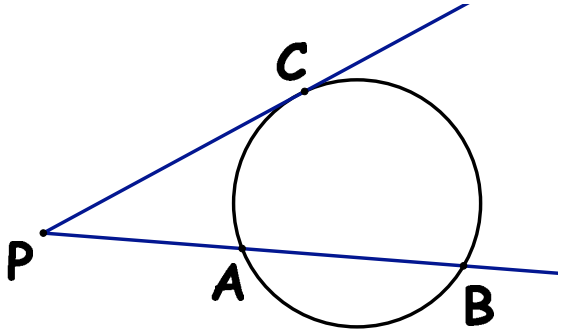
If we let P an arbitrary point also subject to the same condition, then from the Angle Bisector Theorem we infer that $A\widehat{P}P_1 = P_1\widehat{P}B$ and $B\widehat{P}P_2 = 180 - A\widehat{P}B$.

This instantly implies that $P_1\widehat{P}P_2$ is a right angle, from which we conclude (from Exercise 2, page 30 above) that P sits on the circle with diameter $[P_1P_2]$, proving the result.

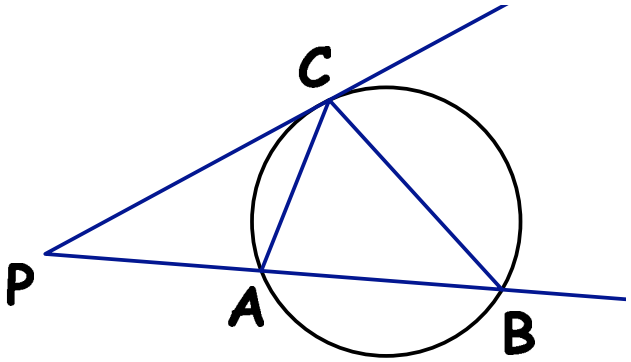
1.3.2 Steiner's theorem and the power of a point

SECANT-TANGENT THEOREM. *We are given the a circle, a tangent line* (PC) *and a secant line* (PA)*, where* C *is the point of tangency and where* $[AB]$ *is a chord of the circle on the secent (see the figure to the right. Then*

$$PC^2 \;=\; PA \times PB.$$



PROOF. This is almost trivial; simply note that $P\widehat{C}A$ and $A\widehat{B}C$ open the same angle. Therefore, $\triangle PCA \sim \triangle PBC$, from which the conclusion follows.



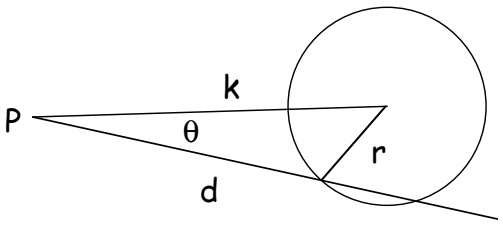
There is also an almost purely algebraic proof of this result.[7]

The following is immediate.

[7] If the radius of the circle is r and if the distance from P to the center of the circle is k, then denoting d the distance along the line segment to the two points of intersection with the circle and using the Law of Cosines, we have that $r^2 = k^2 + d^2 - 2kd\cos\theta$ and so d satisfies the quadratic equation

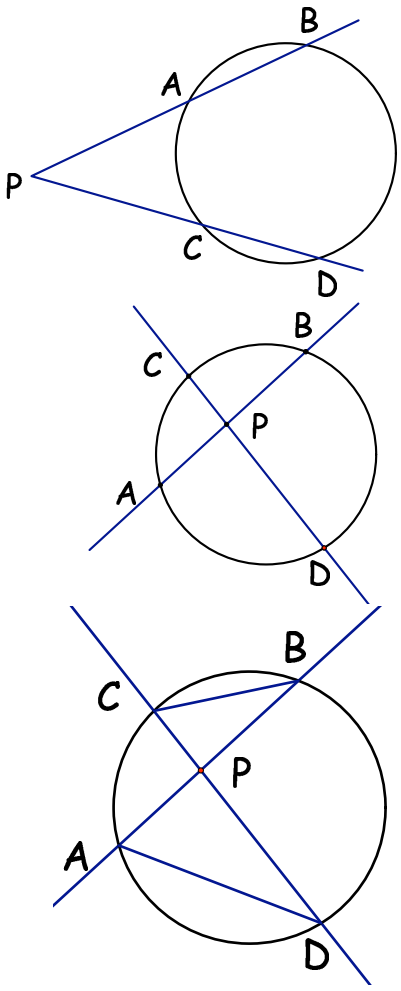
$$d^2 - 2kd\cos\theta + k^2 - r^2 = 0.$$

The product of the two roots of this equation is $k^2 - d^2$, which is independent of the indicated angle θ.



COROLLARY. **(Steiner's Theorem)** *We are given the a circle, and secant lines (PA) and (PC), where (PA) also intersects the circle at B and where (PC) also intersects the circle at D.*

$$PA \times PB \;=\; PC \times PD.$$



PROOF. Note that only the case in which P is interior to the circle needs proof. However, since angles $C\widehat{B}P$ and $P\widehat{D}A$ open the same are, they are equal. Therefore, it follows instantly that $\triangle PDA \sim \triangle PBC$, from which the result follows.

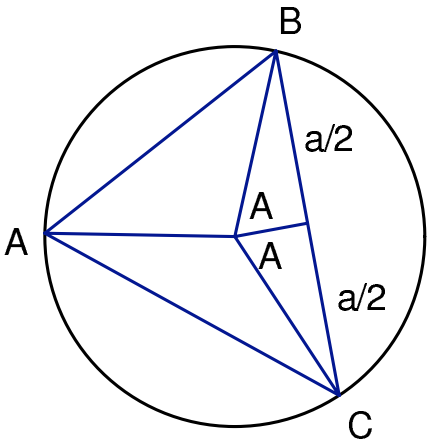
The product $PA \times PB$ of the distances from the point P to the points of intersection of the line through P with the given circle is independent of the line; it is called the **power of the point with respect to the circle**. It is customary to use *signed magnitudes* here, so that the power of the point with respect to the circle will be *negative* precisely when P is *inside* the circle. Note also that the power of the point P relative to a given circle $\mathcal{C}$ is a function only of the distance from P to the center of $\mathcal{C}$. (Can you see why?)

The second case of Steiner's theorem is sometimes called the "Intersecting Chords Theorem."

EXERCISES

1. In the complex plane, graph the equation $|z+16| = 4|z+1|$. How does this problem relate with any of the above?

2. Prove the "Explicit Law of Sines," namely that if we are given the triangle $\triangle ABC$ with sides a, b, and c, and if R is the circumradius, then



$$\frac{a}{\sin A} = \frac{b}{\sin B} = \frac{c}{\sin C} = 2R.$$

Conclude that the perimeter of the triangle is
$a+b+c = 2R(\sin A+\sin B+\sin C)$.

3. Let a circle be given with center O and radius r. Let P be a given point, and let d be the distance OP. Let l be a line through P intersecting the circle at the points A and A'. Show that

   (a) If P is inside the circle, then $PA \times PA' = r^2 - d^2$.

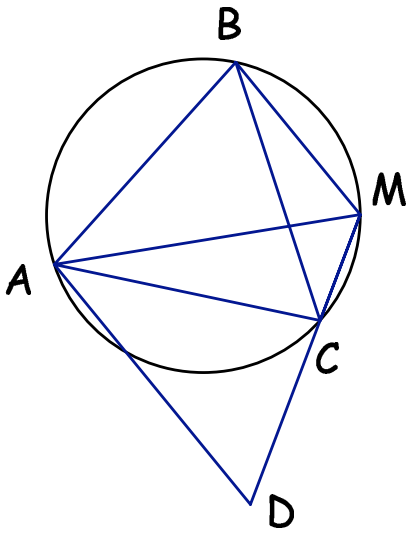
   (b) If P is outside the circle, then $PA \times PA' = d^2 - r^2$.

   Therefore, if we use sensed magnitudes in defining the power of P relative to the circle with radius r, then the power of P relative to this circle is always $d^2 - r^2$.

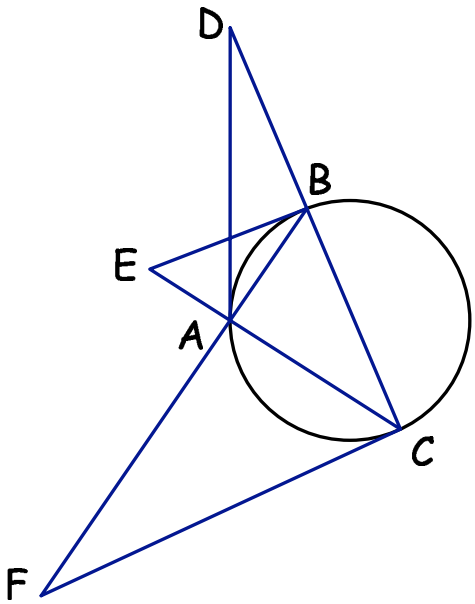
4. Given the circle $\mathcal{C}$ and a real number p, describe the locus of all points P having power p relative to $\mathcal{C}$.

5. Let P be a point and let $\mathcal{C}$ be a circle. Let A and A' be **antipodal** points on the circle (i.e., the line segment $[AA']$ is a diameter of $\mathcal{C}$). Show that the power of P relative to $\mathcal{C}$ is given by the vector dot product $\overrightarrow{PA} \cdot \overrightarrow{PA'}$. (Hint: Note that if O is the center of $\mathcal{C}$, then $\overrightarrow{PA}=\overrightarrow{PO} + \overrightarrow{OA}$ and $\overrightarrow{PA'}=\overrightarrow{PO} - \overrightarrow{OA}$. Apply exercise 3.)

6. Prove Van Schooten's theorem. Namely, let $\triangle ABC$ be an equilateral triangle, and let $\mathcal{C}$ be the circumscribed circle. Let $M \in \mathcal{C}$ be a point on the shorter arc $\widehat{BC}$. Show that $AM = BM + CM$. (Hint: Construct the point D subject to $AM = DM$ and show that $\triangle ABM \cong \triangle ACD$.)



7. The figure to the right shows the triangle $\triangle ABC$ inscribed in a circle. The tangent to the circle at the vertex A meets the line (BC) at D, the tangent to the circle at B meets the line (AC) at E, and the tangent to the circle at C meets the line (AB) at F. Show that D, E, and F are colinear. (Hint: note that $\triangle ACD \sim \triangle BAD$ (why?) and from this you can conclude that $\frac{DB}{DC} = \left(\frac{AB}{AC}\right)^2$. How does this help?)



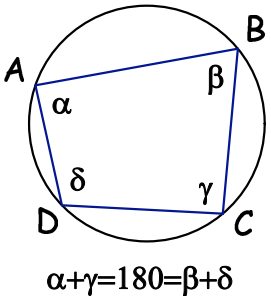
1.3.3 Cyclic quadrilaterals and Ptolemy's theorem

As we have already seen, any triangle can be incribed in a circle; this circle will have center at the circumcenter of the given triangle. It is then natural to ask whether the same can be said for arbitrary polygons. However, a moment's though reveals that this is, in general false even for quadrilaterals. A quadrilateral that can be incribed in a circle is called a **cyclic quadrilateral**.

THEOREM. *The quadrilateral $ABCD$ is cyclic if and only if*

$$A\widehat{B}C + C\widehat{D}A \;=\; C\widehat{A}B + B\widehat{C}D \;=\; 180^\circ. \quad (1.1)$$

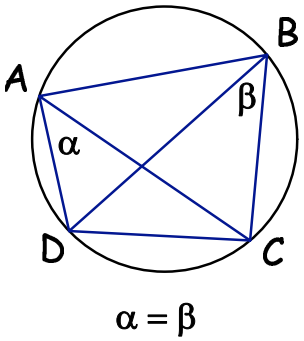
In other words, both pairs of opposite angles add to 180°.



PROOF. If the quadrilateral is cyclic, the result follows easily from the Inscribed Angle theorem. (Draw a picture and check it out!) Conversely, assume that the condition holds true. We let $\mathcal{C}$ be circumscribed circle for the triangle $\triangle ABC$. If D were inside this circle, then clearly we would have $A\widehat{B}C + C\widehat{D}A > 180^\circ$. If D were outside this circle, then $A\widehat{B}C + C\widehat{D}A < 180^\circ$, proving the lemma.

The following is even easier:

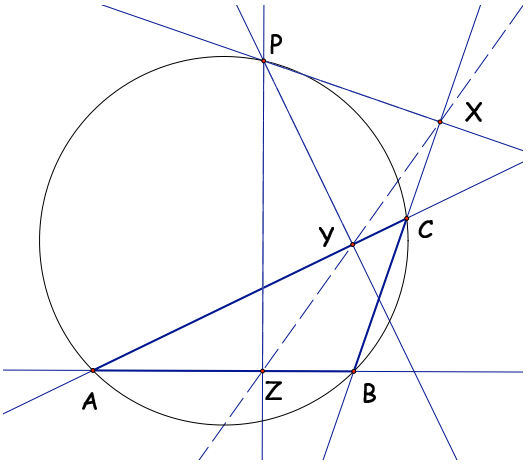
THEOREM. *The quadrilateral $ABCD$ is cyclic if and only if $D\widehat{A}C = D\widehat{B}C$.*



PROOF. The indicated angles open the same arc. The converse is also (relatively) easy.

Simson's line (Wallace's line). There is another line that can be naturally associated with a given triangle $\triangle ABC$, called *Simson's Line* (or sometimes *Wallace's Line*), constructed as follows.

Given the triangle $\triangle ABC$, construct the circumcenter $\mathcal{C}$ and arbitrarily choose a point P on the circle. From P drop perpendiculars to the lines (BC), (AC), and (AB), calling the points of intersection X, Y, and Z, as indicated in the figure below.



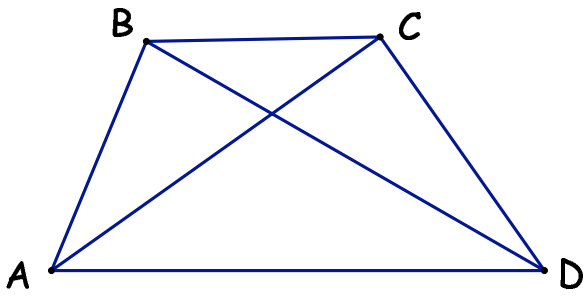
THEOREM. *The points X, Y, and Z, constructed as above are colinear. The resulting line is called* **Simson's line** (*or* **Wallace's line**) *of the triangle $\triangle ABC$.*

PROOF. Referring to the diagram we note that $P\widehat{Z}B$ and $P\widehat{X}B$ are both right angles. This implies that $X\widehat{P}Z + Z\widehat{B}X = 180°$ and so the quadrilateral $PXBZ$ is cyclic. As a result, we conclude that $P\widehat{X}Z = P\widehat{B}Z$. Likewise, the smaller quadrilateral $PXCY$ is cyclic and so $P\widehat{C}A = P\widehat{C}Y = P\widehat{X}Y$. Therefore,

$$\begin{aligned} P\widehat{X}Z &= P\widehat{B}Z \\ &= P\widehat{B}A \\ &= P\widehat{C}A \quad \text{(angles open the same arc)} \\ &= P\widehat{C}Y \\ &= P\widehat{X}Y, \end{aligned}$$

which clearly implies that X, Y, and Z are coliner.

PTOLEMY'S THEOREM. *If the quadrilateral $ABCD$ is cyclic, then the product of the two diagonals is equal to the sum of the products of the opposide side lengths:*



$$AC \cdot BD = AB \cdot CD + AD \cdot BC.$$

When the quadrilateral is not cyclic, then

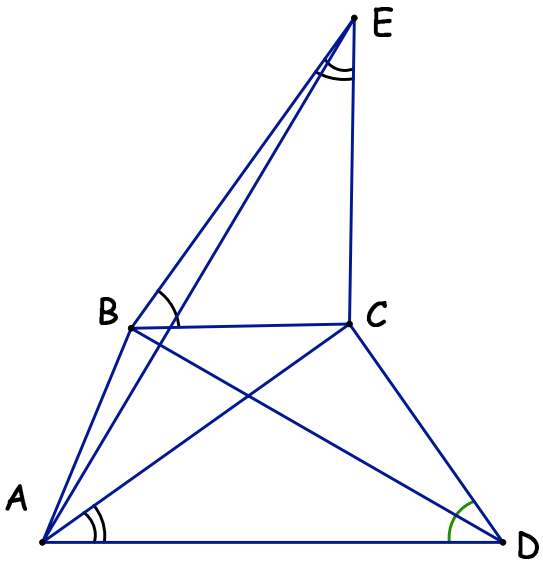
$$AC \cdot BD < AB \cdot CD + AD \cdot BC.$$

PROOF. Whether or not the quadrilateral is cyclic, we can construct the point E so that $\triangle CAD$ and $\triangle CEB$ are similar. This immediately implies that

$$\frac{CE}{CA} = \frac{CB}{CD} = \frac{BE}{DA},$$

from which we obtain

$$BE = \frac{CB \cdot DA}{CD}. \quad (1.2)$$



Also, it is clear that $E\widehat{C}A = B\widehat{C}D$; since also

$$\frac{CD}{CA} = \frac{CB}{CE},$$

we may infer that $\triangle ECA \sim \triangle BCD$. Therefore,

$$\frac{EA}{BD} = \frac{CA}{CD},$$

forcing

$$EA = \frac{CA \cdot DB}{CD}. \quad (1.3)$$

If it were the case that $ABCD$ were cyclic, then by (1.1) we would have

$$C\widehat{B}E + A\widehat{B}C = C\widehat{D}A + A\widehat{B}C = 180^\circ.$$

But this clearly implies that A, B, and E are colinear, forcing

$$EA = AB + BE$$

Using (1.2) and (1.3) we get

$$\frac{CA \cdot DB}{CD} = AB + \frac{CB \cdot DA}{CD},$$

proving the first part of Ptolemy's theorem.

Assume, conversely, that $ABCD$ is not cyclic, in which case it follows that

$$C\widehat{B}E + A\widehat{B}C \;=\; C\widehat{D}A + A\widehat{B}C \neq 180^\circ.$$

This implies that the points A, B, and E form a triangle from which it follows that $EA < AB + BE$. As above we apply (1.2) and (1.3) and get

$$\frac{CA \cdot DB}{CD} < AB + \frac{CB \cdot DA}{CD},$$

and so

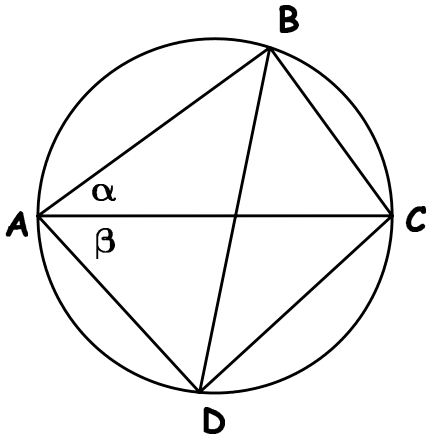
$$CA \cdot DB < AB \cdot CD + CB \cdot DA,$$

proving the converse.

Corollary. **(The Addition Formulas for Sine and Cosine)** *We have, for angles α and β, that*

$$\sin(\alpha+\beta) = \sin\alpha\cos\beta+\sin\beta\cos\alpha;\;\; \cos(\alpha+\beta) = \cos\alpha\cos\beta-\sin\alpha\sin\beta.$$

Proof. We shall draw a cyclic quadrilateral inside a circle having diameter $AC = 1$ (as indicated), and leave the details to the reader. (Note that by Exercise 3 on page 30, we have that $BD = \sin(\alpha + \beta)$ (see the figure). To obtain the addition formula for cos, note that $\cos\alpha = \sin(\alpha + \pi/2)$.)



EXERCISES

1. $[AB]$ and $[AC]$ are chords of a circle with center O. X and Y are the midpoints of $[AB]$ and $[AC]$, respectively. Prove that O, X, A, and Y are concyclic points.

2. Derive the Pythagorean Theorem from Ptolemy's theorem. (This is very easy!)

3. Derive Van Schooten's theorem (see page 35) as a consequence of Ptolemy's theorem. (Also very easy!)

4. Use the addition formula for the sine to prove that if $ABCD$ is a cyclic quadrilateral, then $AC \cdot BD = AB \cdot DC + AD \cdot BC$.

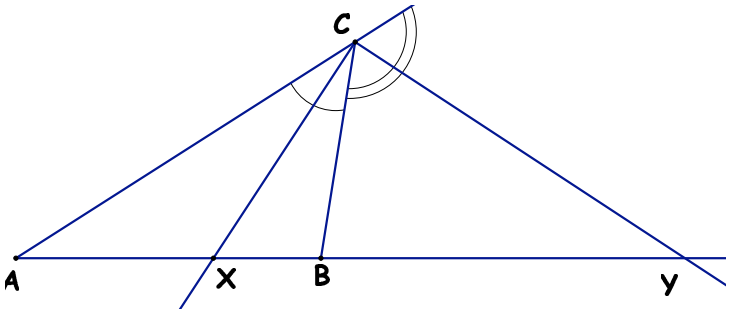
5. Show that if $ABCD$ is a cyclic quadrilateral with side length a, b, c, and d, then the area K is given by

$$K = \sqrt{(s-a)(s-b)(s-c)(s-d)},$$

where $s = (a+b+c+d)/2$ is the semiperimeter.[8]

1.4 Internal and External Divisions; the Harmonic Ratio

The notion of **internal** and **external** division of a line segment $[AB]$ is perhaps best motivated by the familiar picture involving internal and external bisection of a triangle's angle (see the figure to the right). Referring to this figure, we say that the point X divides the segment $[AB]$ **internally** and that the point Y divides the segment $[AB]$ **externally**. In general, if A, B, and X are colinear points, we



[8] This result is due to the ancient Indian mathematician Brahmagupta (598–668).

set $A;X;B = \dfrac{AX}{XB}$ (signed magnitudes); if $A;X;B > 0$ we call this quantity the **internal division** of $[AB]$, and if $A;X;B < 0$ we call this quantity the **external division** of $[AB]$. Finally, we say that the colinear points A, B, X, and Y are in a **harmonic ratio** if

$$A;X;B \;=\; -A;Y;B;$$

that is to say, when

$$\frac{AX}{XB} \;=\; -\frac{AY}{YB} \quad \text{(signed magnitudes)}.$$

It follows immediately from the Angle Bisector Theorem (see page 15) that when (BX) bisects the interior angle at C in the figure above and (BY) bisects the exterior angle at C, then A, B, X, and Y are in harmonic ratio.

Note that in order for the points A, B, X, and Y be in a harmonic ratio it is necessary that one of the points X, Y be interior to $[AB]$ and the other be exterior to $[AB]$. Thus, if X is interior to $[AB]$ and Y is exterior to $[AB]$ we see that A, B, X, and Y are in a harmonic ratio precisely when

$$\text{Internal division of } [AB] \text{ by } X \;=\; -\text{External division of } [AB] \text{ by } Y.$$

EXERCISES

1. Let A, B, and C be colinear points with $(A;B;C)(B;A;C) = -1$. Show that the **golden ratio** is the positive factor on the left-hand side of the above equation.

2. Let A, B, and C be colinear points and let $\lambda = A;B;C$. Show that under the 6=3! permutations of A, B, C, the possible values of $A;B;C$ are

$$\lambda,\ \frac{1}{\lambda},\ -(1+\lambda),\ -\frac{1}{1+\lambda},\ -\frac{1+\lambda}{\lambda},\ -\frac{\lambda}{1+\lambda}.$$

3. Let A, B, X, and Y be colinear points. Define the **cross ratio** by setting

$$[A, B; X, Y] \;=\; \frac{AX}{AY} \cdot \frac{YB}{XB} \;\text{ (signed magnitudes).}$$

Show that the colinear points A, B, X, and Y are in harmonic ratio if $[A, B; X, Y] = -1$.

4. Show that for colinear points A, B, X, and Y one has

$$[A, B; X, Y] = [X, Y; A, B] = [B, A; Y, X] = [Y, X; B, A].$$

Conclude from this that under the $4! = 24$ permutations of A, B, X, and Y, there are at most 6 different values of the cross ratio.

5. Let A, B, X, and Y be colinear points, and set $\lambda = [A, B; X, Y]$. Show that under the 4! permutations of A, B, X, and Y, the possible values of the cross ratio are

$$\lambda,\ \frac{1}{\lambda},\ 1-\lambda,\ \frac{1}{1-\lambda},\ \frac{\lambda}{\lambda-1},\ \frac{\lambda-1}{\lambda}.$$

6. If A, B, X, and Y are in a harmonic ratio, how many possible values are there of the cross ratio $[A, B; X, Y]$ under permutations?

7. Let A and B be given points.

   (a) Show that the locus of points $\{M \mid MP = 3MQ\}$ is a circle.

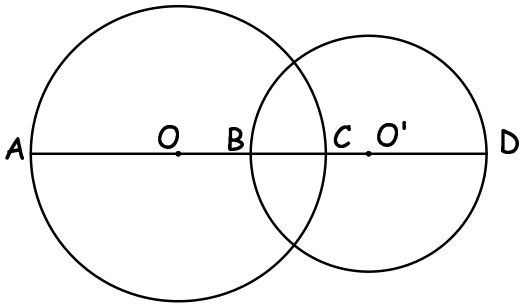
   (b) Let X and Y be the points of intersection of (AB) with the circle described in part (a) above. Show that the points A, B, X, and Y are in a harmonic ratio.

8. Show that if $[A, B; X, Y] = 1$, then either $A = B$ or $X = Y$.

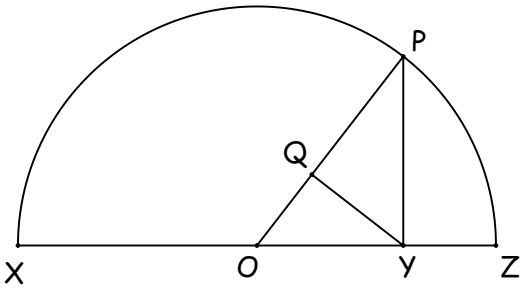
9. The **harmonic mean** of two real numbers is a and b is given by $\dfrac{2ab}{a+b}$. Assume that the points A, B, X, and Y are in a harmonic

ratio. Show that AB is the harmonic mean of AX and AY.[9]

10. The figure to the right depicts two circles having an **orthogonal intersection**. (What should this mean?) Relative to the diagram to the right (O and O' are the centers), show that A, C, B, and D are in a harmonic ratio.



11. The figure to the right shows a semicircle with center O and diameter XZ. The segment $[PY]$ is perpendicular to $[XZ]$ and the segment $[QY]$ is perpendicular to $[OP]$. Show that PQ is the harmonic mean of XY and YZ.



1.5 The Nine-Point Circle

One of the most subtle mysteries of Euclidean geometry is the existence of the so-called "nine-point circle," that is a circle which passes through nine very naturally pre-prescribed points.

To appreciate the miracle which this presents, consider first that arranging for a circle to pass through three noncollinear points is, of course easy: this is the circumscribed circle of the triangle defined by these points (and having center at the circumcenter). That a circle will not, in general pass through four points (even if no three are colinear)

[9]The harmonic mean occurs in elementary algebra and is how one computes the average rate at which a given task is accomplished. For example, if I walk to the store at 5 km/hr and walk home at a faster rate of 10 km/hr, then the average rate of speed which I walk is given by

$$\frac{2 \times 5 \times 10}{5 + 10} = \frac{20}{3} \text{ km/hr.}$$

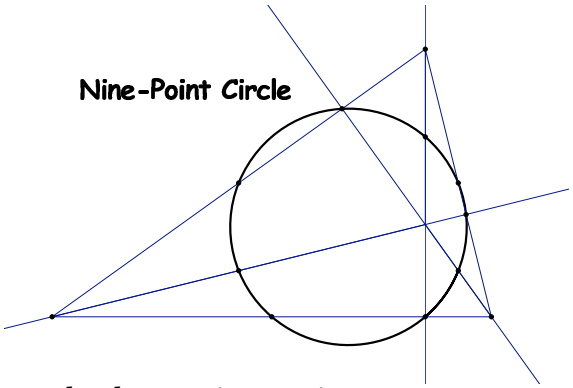
we need only recall that not all quadrilaterals are cyclic. Yet, as we see, if the nine points are carefully—but naturally—defined, then such a circle does exist!

THEOREM. *Given the triangle $\triangle ABC$, construct the following nine points:*

(i) The bases of the three altitudes;

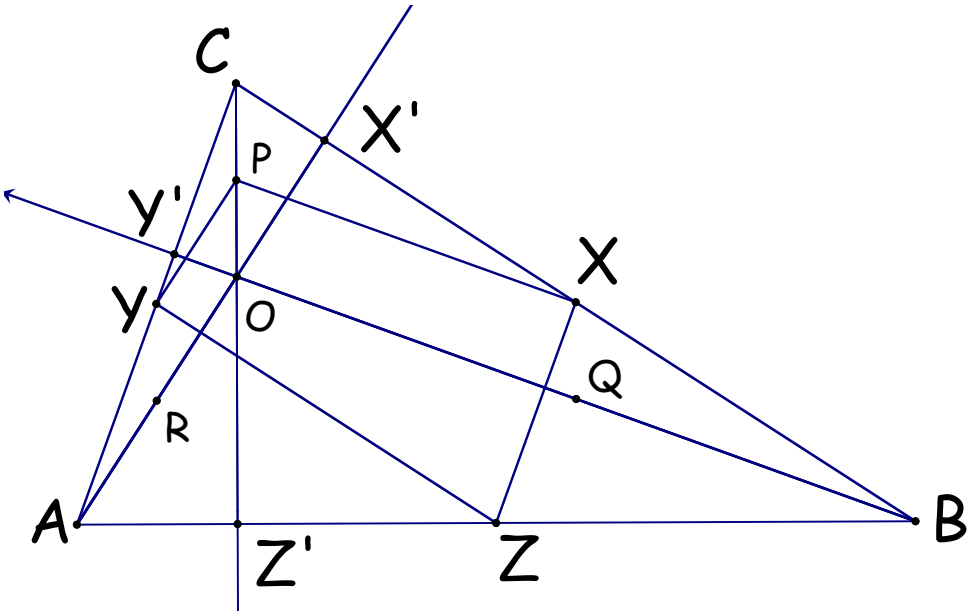
(ii) The midpoints of the three sides;

(iii) The midpoints of the segments joining the orthocenter to each of the vertices.



Then there is a unique circle passing through these nine points.

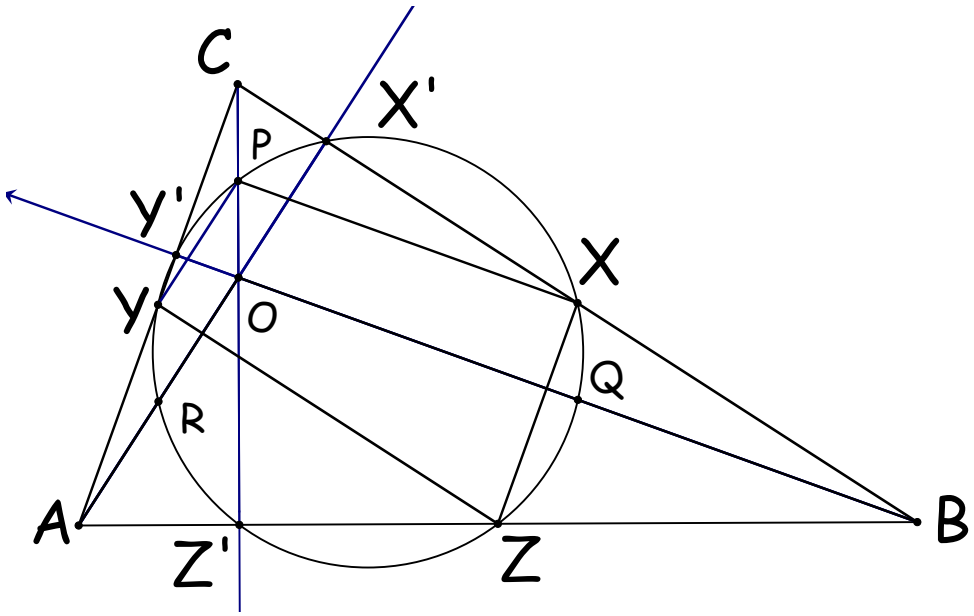
PROOF. Refer to the picture below, where A, B, and C are the vertices, and X, Y, and Z are the midpoints. The midpoints referred to in (iii) above are P, Q, and R. Finally, O is the orthocenter of $\triangle ABC$.



By the Midpoint Theorem (Exercise 3 on page 6 applied to $\triangle ACO$, the line (YP) is parallel to (AX'). Similarly, the line (YZ) is parallel to (BC). This implies immediately that $\angle PYZ$ is a right angle. SImilarly, the Midpoint Theorem applied to $\triangle ABC$ and to $\triangle CBO$ implies that

(XZ) and (AC) are parallel as are (PX) and (BY'). Therefore, $\angle PXZ$ is a right angle. By the theorem on page 35 we conclude that the quadrilateral $YPXZ$ is cyclic and hence the corresponding points all lie on a common circle. Likewise, the quadrilateral $PXZZ'$ is cyclic forcing its vertices to lie on a common circle. As three non-collinear points determine a unique circle (namely the circumscribed circle of the corresponding triangle—see Exercise 8 on page 17) we have already that P, X, Y, Z, and Z' all lie on a common circle.

In an entirely analogous fashion we can show that the quadrilaterals $YXQZ$ and $YXZR$ are cyclic and so we now have that P, Q, R, X, Y, Z, and Z' all lie on a common circle. Further analysis of cyclic quadrilaterals puts Y' and Z' on this circle, and we're done!



Note, finally, that the nine-point circle of $\triangle ABC$ lies on this triangle's Euler line (see page 22).

EXERCISES.

1. Prove that the center of the nine-point circle is the circumcenter of $\triangle XYZ$.

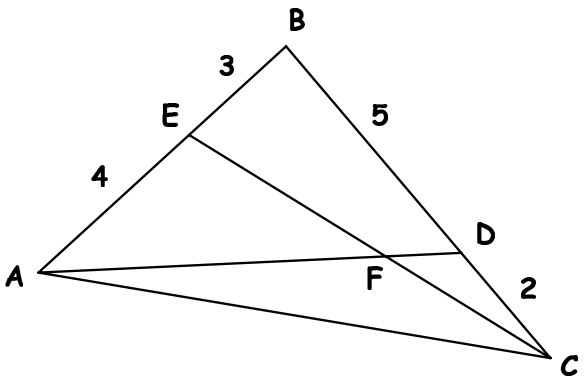
2. Referring to the above diagram, prove that the center of the nine-point circle lies at the midpoint of the segment $[NO]$, where N is the orthocenter of $\triangle ABC$.

3. Given $\triangle ABC$, let O be its orthocenter. Let $\mathcal{C}$ be the nine-point circle of $\triangle ABC$, and let $\mathcal{C}'$ be the circumcenter of $\triangle ABC$. Show that $\mathcal{C}$ bisects any line segment drawn from O to $\mathcal{C}'$.

1.6 Mass point geometry

Mass point geometry is a powerful and useful viewpoint particularly well suited to proving results about ratios—especially of line segments. This is often the province of the Ceva and Menelaus theorems, but, as we'll see, the present approach is both easier and more intuitive.

Before getting to the definitions, the following problem might help us fix our ideas. Namely, consider $\triangle ABC$ with Cevians $[AD]$ and $[CE]$ as indicated to the right. Assume that we have ratios $BE : EA = 3 : 4$ and $CD : DB = 2 : 5$. Compute the ratios $EF : FC$ and $DF : FA$.



Both of the above ratios can be computed fairly easily using the converse to Menalaus' theorem. First consider $\triangle CBE$. From the converse to Menelaus' theorem, we have, since A, F, and D are colinear, that (ignoring the minus sign)

$$1 = \frac{2}{5} \times \frac{7}{4} \times \frac{EF}{FC},$$

forcing $EF : FC = 10 : 7$.

Next consider $\triangle ABD$. Since the points E, F, and C are colinear, we have that (again ignoring the minus sign)

$$1 = \frac{4}{3} \times \frac{7}{2} \times \frac{DF}{FA},$$

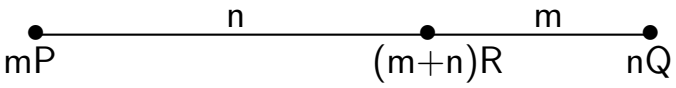
and so $DF : FA = 3 : 14$.

Intuitively, what's going on can be viewed in the following very tangible (i.e., physical) way. Namely, if we assign "masses" to the points of $\triangle ABC$, say

$$A \text{ has mass } \frac{3}{2};\ B \text{ has mass } 2;\ \text{ and } C \text{ has mass } 5,$$

then the point E is at the center of mass of the weighted line segment $[AB]$ and has mass $\frac{7}{2}$, and D is at the center of mass of the weighted line segment $[BC]$ and has mass 7. This suggests that F should be at the center of mass of both of the weighted line segments $[CE]$ and $[AD]$, and should have total mass $\frac{17}{2}$. This shows why $DF : FA = \frac{3}{2} : 7 = 3 : 14$ and why $EF : FC = 5 : \frac{7}{2} = 10 : 7$.

We now formalize the above intuition as follows. By a **mass point** we mean a pair (n, P)—usually written simply as nP—where n is a positive number and where P is a point in the plane.[10] We define an **addition** by the rule: $mP + nQ = (m + n)R$, where the point R is on the line segment $[PQ]$, and is at the center of mass inasmuch as $PR : RQ = n : m$. We view this as below.

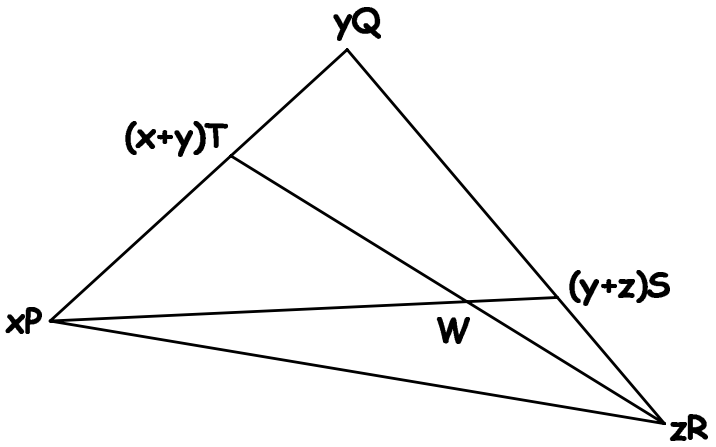


It is clear that the above addition is **commutative** in the sense that $xP + yQ = yQ + xP$. However, what isn't immediately obvious is that this addition is **associative**, i.e., that $xP+(yQ+zR) = (xP+yQ)+zR$ for positive numbers x, y, and z, and points P, Q, and R. The proof is easy, but it is precisely where the converse to Menelaus' theorem comes in! Thus, let

$$yQ + zR = (y + z)S, \quad xP + yQ = (x + y)T.$$

Let W be the point of intersection of the Cevians $[PS]$ and $[RT]$.

[10] Actually, we can take P to be in higher-dimensional space, if desired!



Applying the converse to Menelaus' theorem to the triangle $\triangle PQS$, we have, since T, W, and R are colinear, that (ignore the minus sign)

$$1 = \frac{PT}{TQ} \times \frac{QR}{RS} \times \frac{SW}{WP} = \frac{y}{x} \times \frac{y+z}{y} \times \frac{SW}{WP}.$$

This implies that $PW : WS = (y+z) : x$, which implies that

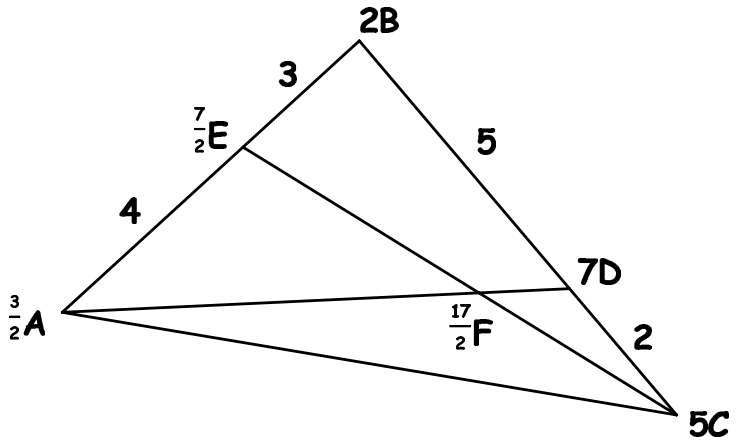
$$(x+y+z)W = xP + (y+z)S = xP + (yQ + zR).$$

Similarly, by applying the converse of Menelaus to $\triangle QRT$, we have that $(x+y+z)W = (x+y)T + zR = (xP+yQ)+zR$, and we're done, since we have proved that

$$xP + (yQ + ZR) = (x+y+z)W = (xP+yQ) + zR.$$

The point of all this is that given mass points xP, yQ, and zR, we may unambiguously denote the "center of mass" of these points by writing $xP + yQ + zR$.

Let's return one more time to the example introduced at the beginning of this section. The figure below depicts the relevant information. Notice that the assigments of masses to A, B, and C are uniquely determined up to a nonzero multiple.



The point F is located at the center of mass—in particular it is on the line segments $[AD]$ and $[CE]$; furthermore its total mass is $\frac{17}{2}$. As a result, we have that $AF : FD = 7 : \frac{3}{2} = 14 : 3$ and $CF : FE = \frac{7}{2} : 5 = 14 : 10$, in agreement with what was proved above.

We mention in passing that mass point geometry can be used to prove Ceva's theorem (and its converse) applied to $\triangle ABC$ when the Cevians $[AX]$, $[BY]$, and $[CZ]$ meet the triangle's sides $[BC]$, $[AC]$, and $[AB]$, respectively. If we are given that

$$\frac{AZ}{ZB} \times \frac{BX}{XC} \times \frac{CY}{YA} = 1,$$

we assign mass ZB to vertex A, mass AZ to vertex B, and mass $\frac{AZ \cdot BA}{XC}$ to vertex C. Since $ZB : \frac{AZ \cdot BX}{XC} = \frac{CY}{YA}$, we see that the center of mass will lie on the intersection of the three Cevians above. Conversely, if we're given the three concurrent Cevians $[AX]$, $[BY]$, and $[CZ]$, then assigning masses as above will place the center of mass at the intersection of the Cevians $[AX]$ and $[CZ]$. Since the center of mass is also on the Cevian $[BY]$, we infer that

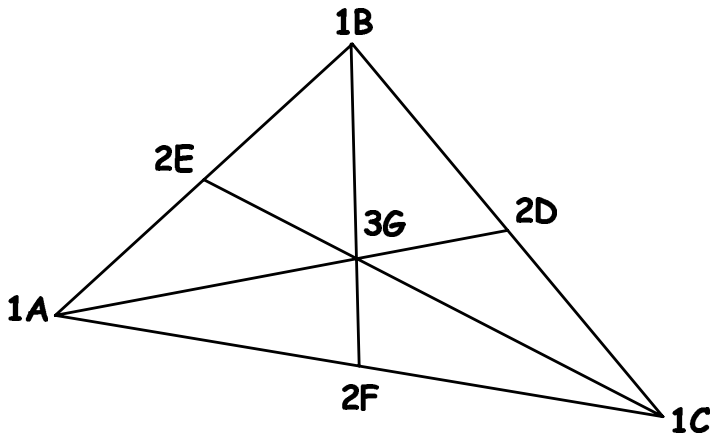
$$\frac{CY}{YA} = \frac{ZB \cdot XC}{AZ \cdot BX},$$

and we're done!

We turn to a few examples, with the hopes of conveying the utility of this new approach. We emphasize: the problems that follow can all be solved without mass point geometry; however, the mass point approach is often simpler and more intuitive!

EXAMPLE 1. *Show that the medians of $\triangle ABC$ are concurrent and the point of concurrency (the centroid) divides each median in a ratio of 2:1.*

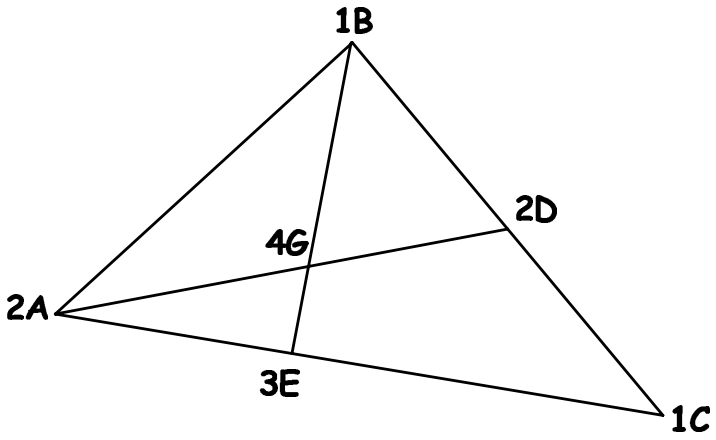
SOLUTION. We assign mass 1 to each of the points A, B, and C, giving rise to the following weighted triangle:



The point G, begin the center of mass, is on the intersection of all three medians—hence they are concurrent. The second statement is equally obvious as $AG : GD = 2 : 1$; similarly for the other ratios.

EXAMPLE 2. *In $\triangle ABC$, D is the midpoint of $[BC]$ and E is on $[AC]$ with $AE : EC = 1 : 2$. Letting G be the intersections of the Cevians $[AD]$ and $[BE]$, find $AG : GD$ and $BG : GE$.*

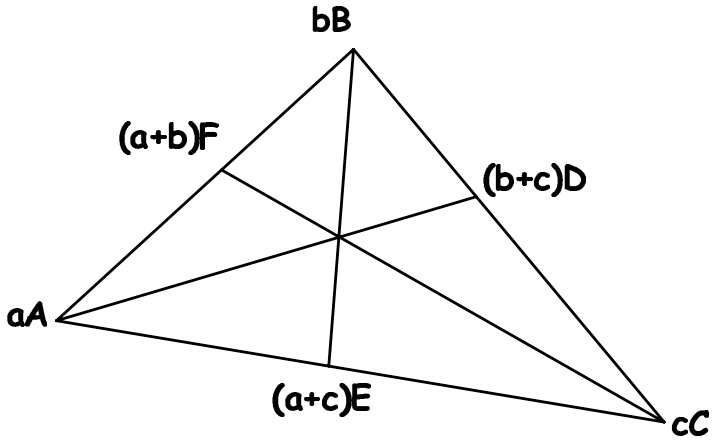
SOLUTION. The picture below tells the story:



From the above, one has $AG : GD = 1 : 1$, and $BG : GE = 3 : 1$.

EXAMPLE 3. *Prove that the angle bisectors of $\triangle ABC$ are concurrent.*

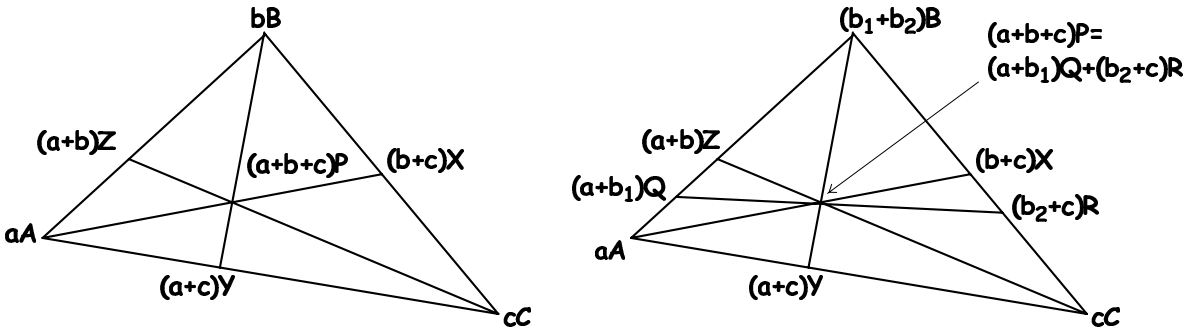
PROOF. Assume that $AB = c$, $AC = b$, $BC = a$ and assign masses a, b, and c to points A, B, and C, respectively. We have the following picture:



Note that as a result of the Angle Bisector Theorem (see page 15) each of the Cevians above are angle bisectors. Since the center of mass is on each of these Cevians, the result follows.

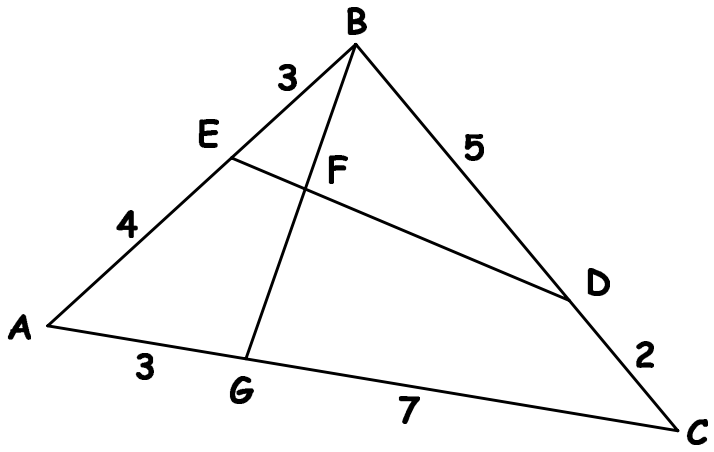
The above applications have to do with Cevians. The method of mass point geometry also can be made to apply to **transversals**, i.e., lines through a triangle not passing through any of the vertices. We shall discuss the necessary modification (i.e., **mass splting**) in the context of the following example.

SOLUTION. The above examples were primarily concerned with computing ratios along particular Cevians. In case a **transversal** is involved, then the method of "mass splitting" becomes useful. To best appreciate this, recall that if in the triangle $\triangle ABC$ we assign mass a to A, b to B, and c to C, then the center of mass P is located on the intersection of the three Cevians (as depicted below). However, suppose that we "split" the mass b at B into two components $b = b_1 + b_2$, then the center of mass P will not only lie at the intersection of the concurrent Cevians, it will also lie on the transversal $[XZ]$; see below:

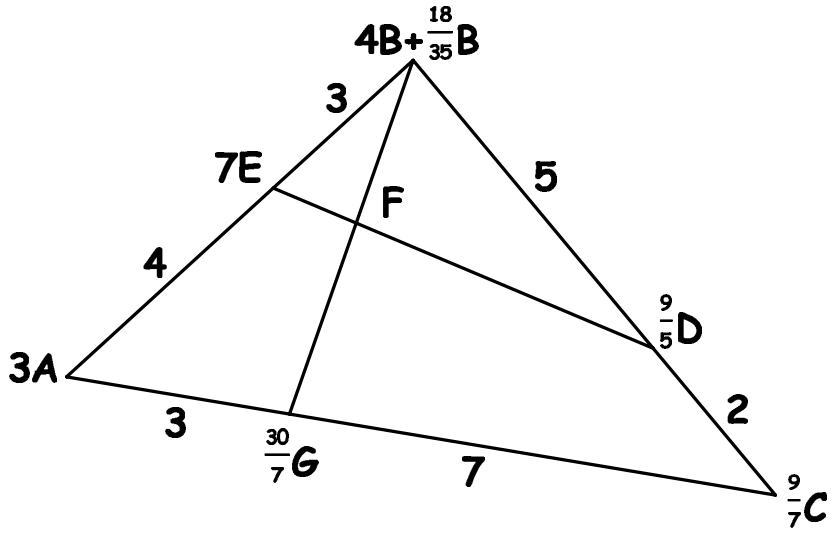


Note that in the above diagram, $QP : PR = (b_2 + c) : (a + b_1)$ because P is the center of mass of $[QR]$.

EXAMPLE 4. *In the figure below, compute $EF : FD$ and $BF : FG$.*



SOLUTION. We shall arrange the masses so that the point F is the center of mass. So we start by assigning weights to A and B to obtain a balance $[AB]$ at E: clearly, assigning mass 4 to B and 3 to A will accomplish this. Next, to balance $[AC]$ at G we need to assign mass $\frac{9}{7}$ to C. Finally, to balance $[BC]$ at D, we need another mass of $\frac{18}{35}$ at B, producing a total mass of $4 + \frac{18}{35}$ at B. The point F is now at the center of mass of the system! See the figure below:



From the above, it's easy to compute the desired ratios:

$$EF : FD = \tfrac{9}{5} : 7 = 9 : 35 \quad \text{and} \quad BF : FG = \tfrac{30}{7} : \tfrac{158}{35} = 75 : 79.$$

Exercises

1. In $\triangle ABC$, D is the midpoint of $[BC]$ and E is on $[AC]$ with $AE : EC = 1 : 2$. Let G be the intersection of segments $[BE]$ and $[AD]$ and find $AG : GD$ and $BG : GE$.

2. In $\triangle ABC$, D is on $[AB]$ with $AD = 3$ and $DB = 2$. E is on $[BC]$ iwht $BE = 3$ and $EC = 4$. Compute $EF : FA$.

3. In quadrilateral $ABCD$, E, F, G, and H are the trisection points of $[AB]$, $[BC]$, $[CD]$, and DA nearer A, C, C, and A, respectively. Show that $EFGH$ is a parallogram. (Show that the diagonals bisect each other.)

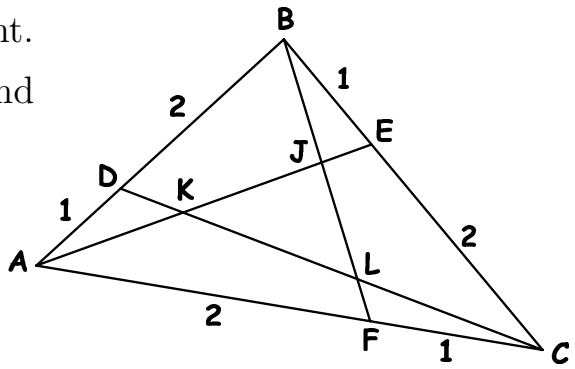
4. Let $[AD]$ be an altitude in $\triangle ABC$, and assume that $\angle B = 45°$ and $\angle C = 60°$. Assume that F is on $[AC]$ such that $[BF]$ bisects $\angle B$. Let E be the intersection of $[AD]$ and $[BF]$ and compute $AE : ED$ and $BE : EF$.

5. [11] In triangle ABC, point D is on $[BC]$ with $CD = 2$ and $DB = 5$, point E is on $[AC]$ with $CE = 1$ and $EA = 3$, $AB = 8$, and $[AD]$ and $[BE]$ intersect at P. Points Q and R lie on $[AB]$ so that $[PQ]$ is parallel to $[CA]$ and $[PR]$ is parallel to $[CB]$. Find the ratio of the area of $\triangle PQR$ to the area of $\triangle ABC$.

6. In $\triangle ABC$, let E be on $[AC]$ with $AE : EC = 1 : 2$, let F be on $[BC]$ with $BF : FC = 2 : 1$, and let G be on $[EF]$ with $EG : GF = 1 : 2$. Finally, assume that D is on $[AB]$ with C, D, G colinear. Find $CG : GD$ and $AD : DB$.

7. In $\triangle ABC$, let E be on $[AB]$ such that $AE : EB = 1 : 3$, let D be on $[BC]$ such that $BD : DC = 2 : 5$, and let F be on $[ED]$ such that $EF : FD = 3 : 4$. Finally, let G be on $[AC]$ such that the segment $[BG]$ passes through F. Find $AG : GC$ and $BF : FG$.

8. You are given the figure to the right.



   (a) Show that $BJ : JF = 3 : 4$ and $AJ : JE = 6 : 1$.

   (b) Show that
   $DK : KL : LC =$
   $EJ : JK : KA =$
   $FL : LJ : JB = 1 : 3 : 3$.

   (c) Show that the area of $\triangle JKL$ is one-seventh the area of $\triangle ABC$.

   (Hint: start by assigning masses 1 to A, 4 to B and 2 to C.)

9. Generalize the above result by replacing "2" by n. Namely, show that the area ratio

$$\text{area}\,\triangle JKL : \text{area}\,\triangle ABC = (n-1)^3 : (n^3 - 1).$$

   (This is a special case of **Routh's theorem** .)

[11] This is essentially problem #13 on the 2002 AMERICAN INVITATIONAL MATHEMATICS EXAMINATION (II).

Chapter 2

Discrete Mathematics

2.1 Elementary Number Theory

While probably an oversimplication, "number theory" can be said to be concerned with the mathematics of the ordinary whole numbers:

$$0, \pm 1, \pm 2, \ldots.$$

We shall, for convenience denote the set of whole numbers by $\mathbb{Z}$.

Notice that the famous **Fermat conjecture**[1] falls into this context, as it asserts that

> *For any integer $n \geq 3$, the equation*
>
> $$x^n + y^n = z^n$$
>
> *has no solution with $x, y, z \in \mathbb{Z}$ with $x, y, z \neq 0$.*

Of course, the assertion is false with $n = 1$ or 2 as, for instance, $3^2+4^2 = 5^2$.

[1] which was proved by Andrew Wiles in 1995

2.1.1 The division algorithm

Very early on, students learn the arithmetic of integers, namely, that of *addition, subtraction, multiplication*, and *division.* In particular, students learn (in about the fifth or sixth grade) that a positive integer a can be divided into a non-negative integer b, resulting in a **quotient** q and a remainder r:

$$\boxed{b = qa + r, \quad 0 \leq r < a.}$$

For instance, the following division of 508 by 28 should serve as an ample reminder.

$$\begin{array}{r} 18 \\ 28\,\overline{)\,508} \\ \underline{28} \\ 228 \\ \underline{224} \\ 4 \end{array}$$

In this case the quotient is 18 and the remainder is 4:

$$508 \;=\; 18 \cdot 28 + 4.$$

The fact that the above is always possible is actually a *theorem*:

THEOREM. (Division Algorithm) *Let $a, b \in \mathbb{Z}$, where $a > 0$, $b \geq 0$. Then there exist* **unique** *integers q and r such that*

$$b = qa + r, \quad \textit{where} \;\; 0 \leq r < a.$$

PROOF. Let S be the following subset of the set $\mathbb{Z}$ of integers:

$$S \;=\; \{b - xa \mid x \in \mathbb{Z} \text{ and } b - xa \geq 0\}.$$

Now let r be the smallest element of this set; note that $r \geq 0$, and let q be defined so that $r = b - qa$. Therefore, we already have $b = qa + r$. Notice that if $r \not\leq a$, then we may set $r' = r - a \geq 0$ and so

$$r' = r - a - a = b - qa - a = b - (q+1)\, a.$$

We see, therefore, that $r' \in S$; since $r' \geq 0$ this contradicts our choice of r in the first place!

Next, we shall show that the quotient and remainder are unique. Therefore, assume that

$$b = qa + r = q'a + r', \quad \text{where} \quad 0 \leq r,\, r' < a.$$

Therefore we conclude that $(q - q')\, a = r' - r$. Since $0 \leq r', r < a$ we see that $|r' - r| < a$ and so $|(q - q')\, a| = |r' - r| < a$ which clearly forces $q - q' = 0$. But then $r' = r$ and we're done![2]

In the above, if $r = 0$, and so $b = qa$, we say that a **divides** b and write $a \,|\, b$.

If $a,\, b \in \mathbb{Z}$ and not both are 0, we say that the integer d is the **greatest common divisor** of a and b if

(i) $d > 0$

(ii) $d \,|\, a$ and $d \,|\, b$,

(iii) if also $d' \,|\, a$ and $d' \,|\, b$ and if $d' > 0$, then $d' \leq d$.

EXAMPLE. In small examples, it's easy to compute the greatest common divisor of integers. For example, the greatest common divisor of 24 and 16 is easily seen to be 4. In examples such as this, the greatest

[2]The assumption in the theorem that a and b are both non-negative was made only out of convenience. In general, the division algorithm states that for two integers $a,\, b \in \mathbb{Z}$, with $a \neq 0$, there exist unique integers q and r such that

$$b = qa + r, \quad \text{where } 0 \leq r < |a|.$$

common divisor is typically obtained by factoring the given numbers into prime factors. However, there is an even more efficient approach, based on the "Euclidean trick" and on the "Euclidean algorithm."

THEOREM. (The Euclidean Trick) *Let a, $b \in \mathbb{Z}$, not both zero. Then the greatest common divisor d of a and b exists. Furthermore, d has the curious representation as*

$$d \;=\; sa + tb,$$

for suitable integers s and t.

PROOF. Consider the set

$$S \;=\; \{xa + yb \,|\, x,\, y \in \mathbb{Z} \text{ and } xa + yb > 0\},$$

and let d be the smallest integer in S (so $d > 0$), and let $d = sa + tb$. Since the greatest common divisor of $|a|$ and $|b|$ is clearly the same as the greatest common divisor of a and b, we may as well just assume that a and b are both positive. Apply the division algorithm and divide d into both a and b:

$$a = q_1 d + r_1, \quad b = q_2 d + r_2, \quad 0 \le r_1,\, r_2 < d.$$

But then $r_1 = a - q_1 d = a - q_1(sa + tb) = (1 - q_1)\, a - q_1 tb$, we see that if $r_1 > 0$, then $r_1 \in S$, which is impossible since $r_1 < d$, and d was taken to be the *smallest* element of S. Therefore, we must have that $r_1 = 0$, which means that $d \,|\, a$. Similarly, $r_2 = 0$ and so $d \,|\, b$. If d' were another positive integer which divides a and b, then $a = md'$ and $b = nd'$, and so $d = sa + tb = s(md') + t(nd') = (sm + tn)\, d'$ which clearly forces $d' \,|\, d$ and so $d' \le d$.

NOTATION: We shall denote the greatest common divisor of a and b by $\gcd(a, b)$.

COROLLARY. *If $d = \gcd(a, b)$ and if d' is any integer satisfying $d' \,|\, a$ and $d' \,|\, b$, then also $d' \,|\, d$.*

PROOF. This is easy! There exist integers s and t with $sa + tb = d$;

given that d' divides both a and b, then obviously d' divides the sum $sa + tb = d$, i.e., $d' \mid d$ also.

We shall present the following without a formal proof. The interested reader should to able to trace through the steps.

THEOREM. (The Euclidean Algorithm) *Let a and b be integers, and assume that $a > 0$. Perform the following divisions:*

$$b = q_1 a + r_1, \quad 0 \leq r_1 < a.$$

If $r_1 = 0$ then $a \mid b$ and so, in fact $a = \gcd(a, b)$. If $r_1 > 0$, divide r_1 into a:

$$a = q_2 r_1 + r_2, \quad 0 \leq r_2 < r_1.$$

If $r_2 = 0$ then one shows easily that $r_1 = \gcd(a, b)$. If $r_2 > 0$, we divide r_2 into r_1:

$$r_1 = q_3 r_2 + r_3, \quad 0 \leq r_3 < r_2.$$

If $r_3 = 0$, then $r_2 = \gcd(a, b)$. If $r_3 > 0$, we continue as above, eventually obtaining $\gcd(a, b)$ as the last nonzero remainder in this process. Furthermore, retracing the steps also gives the "multipliers" s and t satisfying $sa + tb = \gcd(a, b)$.

EXAMPLE. To compute $\gcd(84, 342)$ we can do this by factoring: $84 = 6 \cdot 14$ and $342 = 6 \cdot 57$ from which we get $\gcd(84, 342) = 6$. However, if we apply the Euclidean algorithm, one has
$342 = 4 \cdot 84 + 6$,
$84 = 16 \cdot 6 + 0$.
Therefore, again, $6 = \gcd(84, 342)$. However, we immediately see from the first equation that $6 = 1 \cdot 342 - 4 \cdot 84$, so we can take $s = 1$ and $t = -4$.

Let a and b be integers. We say that the positive integer l is the **least common multiple** of a and b, if

(i) l is a multiple of both a and b,

(ii) If l' is a positive multiple of both a and b then $l \leq l'$.

We denote the least common multiple of a and b by $\text{lcm}(a, b)$.

Assume that a and b are integers satisfying $\gcd(a, b) = 1$. Then we say that a and b are **relatively prime**. We say that the integer $p > 1$ is **prime** if the only positive divisors of p are 1 and p itself. Note that if p is prime and if a is any integer not divisible by p, then clearly p and a are relatively prime.

LEMMA. *Assume that a and b are relatively prime integers and that the integer $a \mid bc$ for some integer c. Then, in fact, $a \mid c$.*

PROOF. We have that for some integers s, $t \in \mathbb{Z}$ that $sa + tb = 1$. Therefore $sac + tbc = c$. Since $a \mid bc$, we have $bc = qa$ for some integer q, forcing

$$c = sac + tbc = (sc + tq)a$$

which says that $a \mid c$, as required.

LEMMA. *Assume that a and b are relatively prime integers, both dividing the integer l. Then $ab \mid l$.*

PROOF. We have that $l = bc$ for a suitable integer c. Since $a \mid l$ we have that $a \mid bc$; apply the above lemma to conclude that $a \mid c$, i.e., $c = ar$ for some integer r. Finally, $l = bc = bar$ which says that $ab \mid l$.

THEOREM. *Given the integers a, $b \geq 0$,* $\text{lcm}(a, b) = \dfrac{ab}{\gcd(a, b)}$.

PROOF. Let $d = \gcd(a, b)$ and set $l = \dfrac{ab}{d}$. Clearly l is a multiple of both a and b. Next, if s and t are integers such that $sa + tb = d$, then $s \cdot \dfrac{a}{d} + t \cdot \dfrac{b}{d} = 1$, proving that $a' = \dfrac{a}{d}$ and $b' = \dfrac{b}{d}$ are relatively prime. From this we may conclude that at least one of the pairs (d, a') or (d, b') is relatively prime. Assume that $\gcd(d, a') = 1$ and let $d' = \gcd(a', b)$.

Then $d' \mid a'$ and $d' \mid b$ and so clearly $d' \mid d$. But then d' divides both a' and d, forcing $d' \mid \gcd(d, a')$, i.e., $d' = 1$. That is to say, a' and b are relatively prime. Therefore if l' is any multiple of a and b then l' is a multiple of a' and b; since a' and b are relatively prime, we have, by the above lemma, that $a'b \mid l'$. In other words, $l \mid l'$, proving that $l = \operatorname{lcm}(a, b)$.

Exercises

1. Assume that a and b are integers and that $d > 0$ is an integer dividing both a and b. Show that if for some integers s, $t \in \mathbb{Z}$ we have $d = sa + tb$, $d = \gcd(a, b)$.

2. Assume that a and b are integers and that there exist integers s, $t \in \mathbb{Z}$ such that $sa + tb = 1$. Show that a and b are relatively prime.

3. Find $\gcd(1900, 399)$, $\operatorname{lcm}(1900, 399)$. Find an explicit representation
$\gcd(1900, 399) = 1900s + 399t,\ s,\ t \in \mathbb{Z}$.

4. Find $\gcd(2100, 399)$, $\operatorname{lcm}(2100, 399)$. Find an explicit representation
$\gcd(2100, 399) = 2100s + 399t,\ s,\ t \in \mathbb{Z}$.

5. Assume that n is a positive integer and that $a,\ b \in \mathbb{Z}$ with $\gcd(a, n) = \gcd(b, n) = 1$. Prove that $\gcd(ab, n) = 1$.

6. Assume that p is a prime, a and b are integers and that $p \mid ab$. Use the Euclidean trick to show that either $p \mid a$ or $p \mid b$.

7. Assume that a and b are relatively prime and that $a \mid bc$ for some integer c. Prove that $a \mid c$.

8. Show that for all integers $n \geq 0$, $6 \mid n(n+1)(2n+1)$.[3]

[3]Of course, this is obvious to those who know the formula:

$$\sum_{k=1}^{n} k^2 \ = \ \frac{n(n+1)(2n+1)}{6}.$$

9. Let p be a prime and show that for all integers $h,\ 1 \leq h \leq p-1$, $p \Big| \binom{p}{h}$. Conclude that for any integers x and y, the numbers $(x+y)^p$ and $x^p + y^p$ have the same remainder when divided by p.

10. Show that the converse of Exercise 9 is also true. Namely, if n is a positive integer such that for all integers $h,\ 1 \leq h \leq n-1$, $n \Big| \binom{n}{h}$, then n is prime. (Hint: Assume that n isn't prime, and let p be a prime divisor of n. Show that if p^r is the highest power of p dividing n, then p^{r-1} is the highest power of p dividing $\binom{n}{p}$.)

11. Assume that n, m are positive integers and k is an exponent such that $n \,|\, (m^k - 1)$. Show that for any non-negative integer h, $n \,|\, (m^{hk} - 1)$.

12. Assume that you have two measuring vessels, one with a capacity of a liters and one of a capacity of b liters. For the sake of specificity, assume that we have an 8-liter vessel and a 5-liter vessel. Using these vessels we may dip into a river and measure out certain amounts of water. For example, if I wish to measure exactly 3 liters of water I could fill the 8-liter vessel, and then from this fill the 5-liter vessel; what remains in the 8-liter vessel is exactly 3 liters of water.

    (a) Using the 8-liter vessel and 5-liter vessel, explain how to measure out exactly 1 liter of water.

    (b) Assume that we return to the general situation, viz., where we have an a liter vessel and a b-liter vessel. Explain how to measure out exactly d liters of water, where $d = \gcd(a, b)$.

13. Let a and b be integers, both relatively prime to the positive integer n. Show that ab is also relatively prime to n.

14. Here's a cute application of the Euclidean Algorithm. Let a and b be positive integers and let $q_k,\ r_k,\ k = 1, 2, \ldots$, be the sequence of integers determined as in the Euclidean Algorithm (page 59). Assume that r_m is the first zero remainder. Then[4]

[4]The expression given is often called a **simple finite continued fraction**.

$$\frac{b}{a} = q_1 + \cfrac{1}{q_2 + \cfrac{1}{q_3 + \cfrac{1}{\ddots \quad \ddots \\ q_{m-1} + \cfrac{1}{q_m}}}}.$$

15. For any positive integer n, let $\mathbb{U}_n$ be the set of all integers relatively prime to n. Now let m, n be relatively prime positive integers and show that
$$\mathbb{U}_{mn} = \mathbb{U}_m \cap \mathbb{U}_n.$$

16. Let $n > 1$ be an integer and define the so-called *Euler* ϕ-**function** (or Euler's **totient** function) by setting

$$\phi(n) \;=\; \#\text{ of integers } m,\ 1 \le m < n \text{ which are relatively prime with } n.$$

Now prove the following.

(a) If p is prime, then $\phi(p) = p - 1$.

(b) If p is prime, and if e is a positive integer, then $\phi(p^e) = p^{e-1}(p-1)$.

(c) If m and n are relatively prime, $\phi(mn) = \phi(m)\phi(n)$. (Hint: Try this line of reasoning. Let $1 \le k < mn$ and let r_m, r_n be the remainders of k by dividing by m and n, respectively. Show that if $\gcd(k, mn) = 1$, then $\gcd(r_m, m) = \gcd(r_n, n) = 1$. Conversely, assume that we have integers $1 \le r_m < m$ and $1 \le r_n < n$ with $\gcd(r_m, m) = \gcd(r_n, n) = 1$. Apply the Euclidean trick to obtain integers s, s_m, s_n, t, t_n, t_m satisfying

$$sm + tn = 1, \quad s_m r_m + t_m m = 1, \quad s_n r_n + t_n n = 1.$$

Let $k = smr_n + tnr_m$, and let k_{mn} be the remainder obtained by dividing k by mn. Show that $1 \le k_{mn} < mn$ and that $\gcd k_{mn}, mn) = 1$. This sets up a correspondence between the positive integers less than mn and relatively prime to mn and the pairs of integers less than m and n and relatively prime to m and n, respectively.)

(d) Show that for any positive integer n, $\phi(n) \geq \sqrt{n/2}$. (Hint: prove that for every prime power p^e, where e is a positive integer, $\phi(p^e) \geq p^{e/2}$, unless $p = 2$ and $e = 1$. What happens in this case?)

See the footnote.[5]

17. Given the positive integer n, and the positive divisor $d \,|\, n$ show that the number of integers k satisfying $1 \leq k < n$ with $\gcd(k, n) = d$ is $\phi\left(\frac{n}{d}\right)$. Conclude that

$$\sum_{d \,|\, n} \phi(d) = n,$$

where the above sum is over the positive divisors of n.

18. Let q and n be positive integers. Show that

$$\begin{array}{l}\#\text{ of integers } m,\ 1 \leq m < qn \\ \text{ which are relatively prime with } n\end{array} \quad = \quad q\phi(n).$$

19. Suppose that x is a positive integer with $x = qn + r,\ 0 \leq r < n$. Show that

$$q\phi(n) \leq \begin{array}{l}\#\text{ of integers } m,\ 1 \leq m < x \\ \text{ which are relatively prime with } n\end{array} \leq (q+1)\phi(n).$$

20. Conclude from Exercises 18 and 19 that

$$\lim_{x \to \infty} \frac{\left(\begin{array}{l}\#\text{ of integers } m,\ 1 \leq m < x \\ \text{ which are relatively prime with } n\end{array}\right)}{x} \quad = \quad \frac{\phi(n)}{n}.$$

[5] Euler's ϕ-function has an interesting recipe, the proof of which goes somewhat beyond the scope of these notes (it involves the notion of "inclusion-exclusion"). The formula says that for any integer $n > 1$,

$$\phi(n) = n \prod_{p \,|\, n} \left(1 - \frac{1}{p}\right),$$

where the product is taken over prime divisors p of n. A main ingredient in proving this is the result of Exercise 17, above. Note that this formula immediately implies that $\phi(mn) = \phi(m)\phi(n)$ when m and n are relatively prime.

2.1.2 The linear Diophantine equation $ax + by = c$

Suppose that a, b, c are integers and suppose that we wish to find all possible solutions of the **linear Diophantine equation** $ax + by = c$. First of all we need a condition on a, b, c in order to guarantee the existence of a solution.

THEOREM. *The linear Diophantine equation $ax + by = c$ has a solution if and only* $\gcd(a, b) \mid c$.

PROOF. Set $d = \gcd(a, b)$ and assume that $c = kd$ for some integer k. Apply the Euclidean trick to find integers s and t with $sa + tb = d$; multiply through by k and get $a(sk) + b(tk) = kd = c$. A solution is therefore $x = sk$ and $y = tk$. Conversely, assume that $ax + by = c$. Then since $d \mid a$ and $d \mid b$, we see that $d \mid (ax + by)$, i.e., $d|c$, proving the theorem.

As the above indicates, applying the Euclidean algorithm will yield a solution of the Diophantine equation $ax + by = c$. We would like now to show how to obtain the **general solution** of this equation, that is to find a recipe for generating all possible solutions. Let's start with a fixed solution (x_0, y_0) and let (x, y) be another solution. This immediately implies that $ax_0 + by_0 = c$, $ax + by = c$ and so $a(x_0 - x) = b(y - y_0)$. Setting $d = \gcd(a, b)$ we have

$$\frac{a}{d}(x_0 - x) = \frac{b}{d}(y - y_0).$$

Next, we note that since $\frac{a}{d}$ and $\frac{b}{d}$ are relatively prime, then by Exercise 7 on page 61 we have that $\frac{a}{d}$ divides $y - y_0$, say $y - y_0 = \frac{a}{d}t$ for some integer t. But then $\frac{a}{d}(x_0 - x) = \frac{b}{d} \cdot \frac{a}{d}t$, forcing $x_0 - x = \frac{b}{d}t$. In other words, starting with a fixed solution (x_0, y_0) of the Diophantine equation $ax + by = c$ we know that any other solution (x, y) must have the form

$$x = x_0 - \frac{b}{d}t, \quad y = y_0 + \frac{a}{d}t, \quad t \in \mathbb{Z} \tag{2.1}$$

Finally, we see (by substituting into the equation) that the above actually is a solution; therefore we have determined *all* solutions of the given Diophantine equation. We summarize.

THEOREM. *Given the linear Diophantine equation $ax + by = c$ where c is a multiple of $d = \gcd(a, b)$, and given a particular solution (x_0, y_0), the general solution is given by*

$$x = x_0 - \frac{b}{d}t, \quad y = y_0 + \frac{a}{d}t, \quad t \in \mathbb{Z}.$$

EXAMPLE. Consider the Diophantine equation $2x + 3y = 48$.

(i) Find all solutions of this equation.

(ii) Find all **positive** solutions, i.e., all solutions (x, y) with x, $y > 0$.

SOLUTION. First of all, a particular solution can be found by simple inspection: clearly $(x, y) = (24, 0)$ is a solution. Next, since 2 and 3 are relatively prime we conclude from the above theorem that the general solution is given by

$$x = 24 - 3t, \quad y = 2t, \quad t \in \mathbb{Z}.$$

Next, if we seek only positive solutions then clearly $t > 0$ and $24 - t > 0$. This reduces immediately to $0 < t < 24$, which is equivalent with saying that $1 \leq t \leq 23$. That is, the positive solutions are described by writing

$$x = 24 - 3t, \quad y = 2t, \quad t \in \mathbb{Z}, \quad 1 \leq t \leq 23.$$

EXERCISES

1. Find all integer solutions of the Diophantine equation $4x + 6y = 100$. Also, find all positive solutions.

2. Find all solutions of $15x + 16y = 900$, with $x,\ y \geq 0$.

3. Suppose that someone bought a certain number of 39-cent pens and a certain number of 69-cent pens, paying \$11.37 for the total. Find the number of 39-cent pens and the number of 69-cent pens purchased.

4. I recently purchased a number of DVDs at 6¥each and a number of DVDs at 7¥each, paying 249¥for the total. Find the number of 6¥DVDs and the number of 7¥DVDs assuming that I purchased approximately the same number of each.

5. Solve $15x - 24y = 3,\ x, y \geq 0$.

6. Farmer Jones owes Farmer Brown \$10. Both are poor, and neither has any money, but Farmer Jones has 14 cows valued at \$184 each and Farmer Jones has a large collection of pigs, each valued at \$110. Is there a way for Farmer Jones to pay off his debt?

7. A **Pythagorean triple** is a triple (a, b, c) of positive integers such that $a^2 + b^2 = c^2$. Therefore, $(3, 4, 5)$ is an example of a Pythagorean triple. So is $(6, 8, 10)$. Call a Pythagorean triple (a, b, c) **primitive** if a, b, and c share no common factor greater than 1. Therefore, $(3, 4, 5)$ is a primitive Pythagorean triple, but $(6, 8, 10)$ is not.

   (a) Assume that s and t are positive integers such that

   (i) $t < s$,

   (ii) s and t are relatively prive, and

   (iii) one of s, t is odd; the other is even.

   Show that if $x = 2st,\ y = s^2 - t^2,\ z = s^2 + t^2$, then (x, y, z) is a Pythagorean triple.

   (b) Show that every Pythagorean triple occurs as in (a), above.

8. This problem involves a system of Diophantine equations.[6] Ed and Sue bike at equal and constant rates. Similarly, they jog at equal and constant rates, and they swim at equal and constant rates. Ed

[6]Essentially Problem #3 from the 2008 American Invitational Mathematics Examination.

covers 74 kilometers after biking for 2 hours, jogging for 3 hours, and swimming for 4 hours, while Sue covers 91 kilometers after jogging for 2 hours, swimming for 3 hours, and biking for 4 hours. Their biking, jogging, and swimming rates are all whole numbers in kilometers per hour. Find these rates.

2.1.3 The Chinese remainder theorem

CONGRUENCE AND THE INTEGERS MODULO n. If n is a positive integer, and if a and b are integers, we say that a **is congruent to** b **modulo** n and write $a \equiv b (\mathrm{mod}\ n)$ if $n \,|\, (a-b)$. Next, we write $\mathbb{Z}_n = \{0_n, 1_n, 2_n, \ldots, (n-1)_n\}$ with the understanding that if b is any integer, and if $b = qn + r$, where $0 \leq r < n$, then $b_n = r_n$. Sometimes we get lazy and just write $\mathbb{Z}_n = \{0, 1, 2, \ldots, n-1\}$ without writing the subscripts if there is no possibility of confusion. As an example, we see that $\mathbb{Z}_6 = \{0, 1, 2, 3, 4, 5\}$ with such further stipulations as $8 = 2$, $22 = 4$, $-5 = 1$. The integers modulo n can be added (and multiplied) pretty much as ordinary integers, we just need to remember to reduce the answer modulo n.

EXAMPLE. We can write out the sums and products of integers modulo 6 conveniently in tables:

$+$	0	1	2	3	4	5
0	0	1	2	3	4	5
1	1	2	3	4	5	0
2	2	3	4	5	0	1
3	3	4	5	0	1	2
4	4	5	0	2	3	4
5	5	0	1	2	3	4

$\cdot$	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	1	2	3	4	5
2	0	2	4	0	2	4
3	0	3	0	3	0	3
4	0	4	2	0	4	2
5	0	5	4	3	2	1

The following story[7] conveys the spirit of the Chinese Remainder Theorem:

[7] Apparantly due to the Indian mathematician Brahmagupta (598–670).

> An old woman goes to market and a horse steps on her basket and crushes the eggs. The rider offers to pay for the damages and asks her how many eggs she had brought. She does not remember the exact number, but she remembered that when she had taken them out two at a time, there was one egg left. The same happened when she picked them out three, four, five, and six at a time, but when she took them seven at a time they came out even. What is the smallest number of eggs she could have had?

The solution of the above is expressed by a system of congruences: if m is the number of eggs that the old woman had, then

$$\begin{aligned} m &\equiv 1 (\text{mod } 2) \\ m &\equiv 1 (\text{mod } 3) \\ m &\equiv 1 (\text{mod } 4) \\ m &\equiv 1 (\text{mod } 5) \\ m &\equiv 1 (\text{mod } 6) \\ m &\equiv 0 (\text{mod } 7) \end{aligned}$$

Expressed in terms of integers modulo n for various n, we can express the above as

$$m_2 = 1_2; \quad m_3 = 1_3; \quad m_4 = 1_4; \quad m_5 = 1_5; \quad m_6 = 1_6; \quad m_7 = 0_7.$$

Note first that there is some redundancy in the above problem. Namely, notice that if $m_4 = 1_4$, then surely $m_2 = 1_2$. Indeed,

$$\begin{aligned} m_4 = 1_4 &\implies 4 \,|\, (4-1) \\ &\implies 2 \,|\, (4-1) \\ &\implies m_2 = 1_2. \end{aligned}$$

In exactly the same way we see that the condition $m_6 = 1_6$ implies that

$m_3 = 1_3$. Therefore, we are really faced with the task of finding the smallest integer m satisfying

$$m_4 = 1_4, \quad m_5 = 1_5, \quad m_6 = 1_6, \quad m_7 = 0_7.$$

The first question that should occur to the reader is "why is there any solution m to the above?" As we'll see, this will be the point of emphasis in the Chinese Remainder Theorem.

THEOREM. (Chinese Remainder Theorem) *Let a and b be positive integers, and set $d = \gcd(a, b)$. Let x and y be any two integers satifying $x_d = y_d$. Then there is always an integer m such that*

$$m_a = x_a, \quad m_b = y_b.$$

Furthermore, if $l = \operatorname{lcm}(a, b)$ then any other solution m' is congruent to m modulo l.

PROOF. First of all since $x_d = y_d$ we know that $d \mid (x - y)$; assume that $x - y = zd$, for some integer z. Next, let s and t be integers satisfying $sa + tb = d$, from this we obtain

$$sza + tzb = zd = x - y.$$

From this we see that $x - sza = y + tzb$; we now take m to be this common value: $m = x - sza = y + tzb$ from which it is obvious that $m_a = x_a$ and $m_b = y_b$.

Finally, if m' is another solution, then we have $m' \equiv m (\text{mod } a)$ and $m' \equiv m (\text{mod } b)$ and so $m' - m$ is a multiple of both a and b. Therefore $l \mid (m' - m)$ and so $m' \equiv m (\text{mod } l)$ proving the theorem.

We'll consider a couple of examples.

EXAMPLE 1. Solve the simultaneous congruences

$$\begin{aligned} m &\equiv 14 (\text{mod } 138) \\ m &\equiv 23 (\text{mod } 855). \end{aligned}$$

Applying the Euclidean algorithm yields

$$\begin{aligned} 855 &= 6 \cdot 138 + 27 \\ 138 &= 5 \cdot 27 + 3 \\ 27 &= 9 \cdot 3 + 0, \end{aligned}$$

and so $d = \gcd(138, 855) = 3$; furthermore the above shows that

$$3 = 138 - 5 \cdot 27 = 138 - 5(835 - 6 \cdot 138) = 31 \cdot 138 - 5 \cdot 855$$

(and so $s = 31$ and $t = -5$). Also, since $14 \equiv 23 (\text{mod } 3)$ we conclude that the above congruences can be solved. Indeed, $14 - 23 = -3 \cdot 3$ (so $z = -3$) and so a solution is $m = x - sza = 14 + 31 \cdot 138 \cdot 3 = 12,848$. Finally, we can prove that 12,848 is actually the *least positive integer solution* of the above congruences above, as follows. To do this, apply the Chinese Remainder Theorem to conclude that if m' is any other solution, and if $l = \text{lcm}(138, 855) = 39,330$, then $m' \equiv 12,848 (\text{mod } 39,330)$. This is clearly enough!

EXAMPLE 2. Find the least positive integer solution of

$$\begin{aligned} m &\equiv 234 (\text{mod } 1832) \\ m &\equiv 1099 (\text{mod } 2417). \end{aligned}$$

This one is technically more involved. However, once one recognizes that 2417 is a prime number, we see immediately that 2417 and 1832 are relatively prime and so at least we know that a solution exists. Now comes the tedious part:

$$\begin{aligned} 2417 &= 1 \cdot 1832 + 585 \\ 1832 &= 3 \cdot 585 + 77 \\ 585 &= 7 \cdot 77 + 46 \\ 77 &= 1 \cdot 46 + 31 \\ 46 &= 1 \cdot 31 + 15 \\ 31 &= 2 \cdot 15 + 1 \\ 15 &= 15 \cdot 1 + 0 \end{aligned}$$

Therefore, $d = \gcd(1832, 2417) = 1$; working backwards through the above yields

$$157 \cdot 1832 - 119 \cdot 2417$$

(so $s = 157$ and $t = -119$). We have $234 - 1099 = -865 = z$ and so a solution is given by $m = x - sza = 234 + 157 \cdot 865 \cdot 1832 = 248,794,994$. Finally, any other solution m' will be congruent to 248,794,994 modulo $l = \text{lcm}(1832, 2417) = 1832 \cdot 2417 = 4,427,944$. We therefore reduce 248,794,994 modulo l using the division algorithm:

$$248,794,994 \;=\; 56 \cdot 4,427,944 = 830,130,$$

and so the least integer solution is $m = 830,130$.

EXAMPLE 3. In this example we indicate a solution of three congruences. From this, the student should have no difficulty in solving more than three congruences, including the lead problem in this subsection. Find the least positive solution of the congruences

$$\begin{aligned} m &\equiv 1 (\text{mod } 6) \\ m &\equiv 7 (\text{mod } 15) \\ m &\equiv 4 (\text{mod } 19). \end{aligned}$$

First, we have

$$1 \cdot 15 - 2 \cdot 6 \;=\; 3$$

from which we conclude that $3 = \gcd(6, 15)$. Next, we have $7 - 1 = 2 \cdot 3$, and so

$$2 \cdot 15 - 2 \cdot 2 \cdot 6 \;=\; 2 \cdot 3 = 7 - 1;$$

this tells us to set $m_1 = 1 - 2 \cdot 2 \cdot 6 = 7 - 2 \cdot 15 = -23$. We have already seen that all solutions will be congruent to -23 modulo $l_1 = \text{lcm}(6, 15) = 30$. Therefore, the least positive solution will be $m_1 = 7$ (which could have probably more easily been found just by inspection!). Note that if m is integer satisfying $m \equiv 7 (\text{mod } 30)$, then of course we also have $m \equiv 7 (\text{mod } 6)$ and $m \equiv 7 (\text{mod } 15)$, and so $m \equiv 1 (\text{mod } 6)$ and $m \equiv 7 (\text{mod } 15)$. Therefore, we need to find the least positive integer solution of

$$\begin{aligned} m &\equiv 7 (\text{mod } 30) \\ m &\equiv 4 (\text{mod } 19). \end{aligned}$$

In this case, $7 \cdot 30 - 11 \cdot 19 = 1$ and so $3 \cdot 7 \cdot 30 - 3 \cdot 11 \cdot 19 = 3 = 7 - 4$ which tells us to set $m = 7 - 3 \cdot 7 \cdot 30 = 4 - 3 \cdot 11 \cdot 19 = -623$. Any other solution will be congruent to -623 modulo $30 \cdot 19 = 570$; apply the division algorithm

$$-623 = 2 \cdot 570 + 517.$$

It follows, therefore, that the solution we seek is $m = 517$.

We conclude this section with a simple corollary to the Chinese Remainder Theorem; see Exercise 16c on page 63.

Corollary to Chinese Remainder Theorem. *Let m and n be relatively prime positive integers, Then $\phi(mn) = \phi(m)\phi(n)$.*

Proof. Let $a,\ b$ be positive integers with $1 \leq a < m,\ 1 \leq b < n$, and $\gcd(a, m) = \gcd(b, n) = 1$. By the Chinese Remainder Theorem, there is a unique integer k, with $1 \leq k < mn$ satisfying $k_m = a,\ k_n = b$. Clearly $\gcd(k, mn) = 1$. Conversely, if the positive integer k is given with $1 \leq k < mn$, and $\gcd(k, mn) = 1$, then setting $a = k_m,\ b = k_n$ produces integers satisfying $1 \leq a < m,\ 1 \leq b < n$ and such that $\gcd(a, m) = \gcd(b, n) = 1$.

Exercises

1. Let n be a positive integer and assume that $a_1 \equiv b_1 (\text{mod } n)$ and that $a_2 \equiv b_2 (\text{mod } n)$. Show that $a_1 + b_1 \equiv a_2 + b_2 (\text{mod } n)$ and that $a_1 b_1 \equiv a_2 b_2 (\text{mod } n)$.

2. Compute the least positive integer n such that $n \equiv 12, 245, 367 (\text{mod } 11)$. (Hint: this is *very* easy! Don't try a direct approach.)

3. Compute the least positive integer n such that $n \equiv 12, 245, 367 (\text{mod } 9)$.

4. Find the least positive integer solution of the congruences

$$\begin{aligned} m &\equiv 7(\text{mod } 10) \\ m &\equiv 17(\text{mod } 26). \end{aligned}$$

5. Find the least positive integer solution of the congruences

$$\begin{aligned} m &\equiv 7(\text{mod } 10) \\ m &\equiv 5(\text{mod } 26) \\ m &\equiv 1(\text{mod } 12). \end{aligned}$$

6. Solve the problem of the woman and the eggs, given at the beginning of this section.

7. If A and B are sets, one defines the **Cartesian product** of A and B by setting

$$A \times B = \{(a, b) \mid a \in A \text{ and } b \in B\}.$$

Now suppose that the positive integers m and n are relatively prime, and define the function

$$f : \mathbb{Z}_{mn} \to \mathbb{Z}_m \times \mathbb{Z}_n \text{ by } f(x_{mn}) = (x_m, x_n) \in \mathbb{Z}_m \times \mathbb{Z}_n.$$

Using the Chinese remainder theorem, show that the function f is one-to-one and onto.

8. [8] The integer N is written as

$$N = 102030x05060y$$

in decimal (base 10) notation, where x and y are missing digits. Find the values of x and y so that N has the largest possible value and is also divisible by both 9 and 4. (Hint: note that $N \equiv -1 + x + y(\text{mod } 9)$ and $N \equiv y(\text{mod } 4)$.)

[8] This is problem #5 on the January 10, 2008 ASMA (American Scholastic Mathematics Association) senior division contest.

2.1.4 Primes and the fundamental theorem of arithmetic

We have already defined a **prime** as a positive integer p greater than 1 whose only positive divisors are 1 and p itself. The following result may seem a bit obvious to the naive reader. What I want, though, is for the reader to understand the nature of the proof.[9]

LEMMA. *Any positive integer $n > 1$ has at least one prime factor.*

PROOF. Denoting by $\mathbb{N}$ the set of positive integers, we define the set

$$C \;=\; \{n \in \mathbb{N} \,|\, n > 1 \text{ and } n \text{ has no prime factors } \}.$$

Think of C as the set of "criminals;" naturally we would like to show that $C = \varnothing$, i.e., that there are no criminals. If $C \neq \varnothing$, then C has a smallest element in it; call it c_0 (the "least criminal"). Since c_0 cannot itself be prime, it must have a non-trivial factorization: $c_0 = c_0' c_0''$, where $1 < c_0', c_0'' < c_0$. But then, $c_0', c_0'' \notin C$ and hence aren't criminals. In particular, c_0' has a prime factor, *which is then a factor of* c_0. So c_0 wasn't a criminal in the first place, proving that $C = \varnothing$, and we're done!

Using the above simple result we can prove the possibly surprising result that there are, in fact, infinitely many primes. This was known to Euclid; the proof we give is due to Euclid:

THEOREM. *There are infinitely many primes.*

PROOF. (Euclid) Assume, by way of contradiction that there are only finitely primes; we may list them:

$$p_1,\, p_2,\, \ldots, p_n.$$

Now form the positive integer $n = 1 + p_1 p_2 \cdots, p_n$. Note that none of the primes $p_1, p_2, \ldots, p_n$ can divide n. However, because of the above lemma we know that n must have a prime divisor $p \neq p_1, p_2, \ldots, p_n$.

[9] I will formalize this method of proof in the next section.

Therefore the original list of primes did not contain *all* of the primes. This contradiction proves the result.

Knowing that there are infinitely many primes, one may ask a slighly more subtle question, namely whether the infinite series

$$\sum_{\text{primes } p} \left(\frac{1}{p}\right) = \frac{1}{2} + \frac{1}{3} + \frac{1}{5} + \cdots + \frac{1}{31} + \cdots$$

converges or diverges. One can show that this series actually diverges, which shows that the prime numbers are relatively densely packed within the set of positive integers.

There are many unsolved conjectures related to prime numbers; we'll just state two such here. The first is related to **twin primes** which are prime pairs of the form p, $p+2$, where both are primes. The first few twin primes are 3, 5, 5, 7, 11, 13, and so on. The so-called "Twin Prime" conjecture which states that there are an infinite number of twin primes. The next is the **Goldbach conjecture** which states that any even integer greater than 2 is the sum of two primes. Neither of these conjectures has been proved.

Using the above method of "criminals"[10] one eventually arrives at the important **Fundamental Theorem of Arithmetic**:

THEOREM. (Fundamental Theorem of Arithmetic) *Any positive integer $n > 1$ has a unique factorization into primes. In other words*

(i) there exist primes $p_1 < p_2 < \cdots < p_r$ and exponents $e_1, e_2, \ldots, e_r$ such that

$$n = p_1^{e_1} p_2^{e_2} \cdots p_r^{e_r}.$$

(ii) The factorization above is unique in that if $n = q_1^{f_1} q_2^{f_2} \cdot q_s^{f_s}$ then $s = r$, $p_1 = q_1, p_2 = q_2, \ldots, p_r = q_r$ and $e_1 = f_1, e_2 = f_2, \ldots, e_r = f_r$.

Now let a and b be positive integers greater than 1. Write

[10] My surrogate for *mathematical induction*

$$a = p_1^{e_1} p_2^{e_2} \cdots p_r^{e_r}, \quad b = p_1^{f_1} p_2^{f_2} \cdots p_r^{f_r}$$

be the prime factorization of a and b where some of the exponents might be 0. For each $i = 1, 2, \ldots, r$, let $m_i = \min\{e_i, f_i\}$ and let $M_i = \max\{e_i, f_i\}$. The following should be clear:

$$\gcd(a, b) = p_1^{m_1} p_2^{m_2} \cdots p_r^{m_r}, \quad \operatorname{lcm}(a, b) = p_1^{M_1} p_2^{M_2} \cdots p_r^{M_r}.$$

From the above we see that we have two rather different methods of finding the greatest common divisor and least common multiple of two positive integers. The first is the Euclidean algorithm, which we encountered on page 59, and the second is based on the Fundamental Theorem of Arithmetic above. On the surface it would appear that the latter method is much easier than the former method—and for small numbers this is indeed the case. However, once the numbers get large then the problem of factoring into primes becomes considerably more difficult than the straightforward Euclidean algorithm.

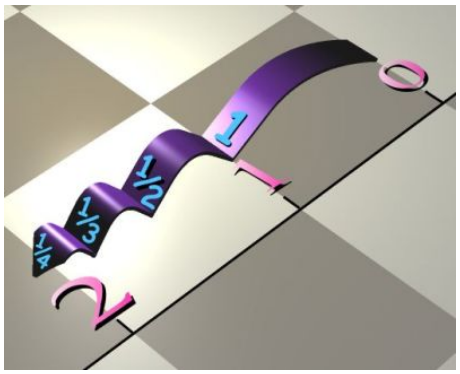
Exercises

1. Find the prime factorizations of the numbers
   (a) 12500
   (b) 12345
   (c) 24227
2. Find the factorization of the numbers $p^3(p-1)^2(p+1)(p^2+p+1)$, where $p = 2, 3, 5, 7$.
3. Compute the gcd and lcm of the following pairs of numbers
   (a) 2090 and 1911
   (b) 20406 and 11999
   (c) $2^{10} + 1$ and $2^{10} - 1$.
4. Show that if p is a prime, then $p+1$ and p^2+p+1 must be relatively prime. Find integers s and t such that $s(p+1) + t(p^2+p+1) = 1$.

5. Show that there exist unique positive integers x and y satisfying $x^2 + 84x + 2008 = y^2$. Find these integers.[11]

6. For each positive integer n, define

$$H(n) = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}.$$



Prove that if $n \geq 2$, then $H(n)$ is **not** an integer.

(Hint: Let k be the largest integer such that $2^k \leq n$, and let M be the least common multiple of the integers $1, 2, \ldots, 2^k - 1$, $2^k + 1, \ldots, n$. What happens when you multiply $H(n)$ by M?)

7. Here's an interesting system of "integers" for which the Fundamental Theorem of Arithmetic fails. Define the set

$$\mathbb{Z}[\sqrt{-5}] \;=\; \{a + b\sqrt{-5} \,|\, a,\; b \in \mathbb{Z}\}.$$

Define primes as on page 60,[12] and show that

$$3 \cdot 7 \;=\; (1 + 2\sqrt{-5})(1 - 2\sqrt{-5}) \;=\; (4 + \sqrt{-5})(4 - \sqrt{-5})$$

give three distinct prime factorizations of 21. In otherwords, the uniqueness aspect of the Fundamental Theorem of Arithmetic fails to hold in this case.

8. In this exercise we outline another proof that there exist infinitely many primes. To this end, define the n-th **Fermat number** F_n, by setting $F_n = 2^{2^n} + 1,\; n = 0, 1, 2, \ldots,$.

   (a) Show that $\prod_{m=0}^{n-1} = F_n - 2,\; n = 1, 2, \ldots$ (Induction!)

   (b) Conclude from part (a) that the Fermat numbers F_m and F_n are relatively prime whenever $m \neq n$.

[11] Essentially Problem #4 from the 2008 American Invitational Mathematics Examination.

[12] Actually, in more advanced treatments, one distinguishes between the notion of a "prime" and the notion of an "irreducible," with the latter being defined more or less as on page 60 (I'm trying to avoid a systematic discussion of "units"). On the other hand, a number p is called **prime** if whenever $p \,|\, ab$, then $p \,|\, a$ or $p \,|\, b$. In the above exercise the numbers given are all irreducibles but, of course, aren't prime.

(c) Conclude from part (b) that there must be infinitely many primes.

9. Here's yet another proof that there are infinitely many primes[13] We start with the simple observation that for any integer $n \geq 2$, n and $n+1$ share no prime factors. Therefore, the product $n(n+1)$ must contain at least two distinct prime factors. We now generate a sequence of integers as follows. Let

$$\begin{aligned} n_1 &= 2 \cdot 3 \\ n_2 &= n_1(n_1+1) = 42 \\ n_3 &= n_2(n_2+1) = 42 \cdot 43 = 1806 \\ &\vdots \end{aligned}$$

What is the minimum number of distinct prime factors contained in n_k?

10. For any positive integer n, let $\tau(n)$ be the number of divisors (including 1 and n) of n. Thus $\tau(1) = 1$, $\tau(2) = 2$, $\tau(3) = 2$, $\tau(4) = 3$, $\tau(10) = 4$, etc. Give a necessary and sufficient condition for $\tau(n)$ to be odd.

11. Continuation of Exercise 10. For each positive integer n, set

$$S(n) = \tau(1) + \tau(2) + \cdots + \tau(n).$$

Let a be the number of integers $n \leq 2000$ for which $S(n)$ is even. Compute a.[14]

2.1.5 The Principle of Mathematical Induction

In the previous section we showed that every integer n has at least one prime factor essentially by dividing the set $\mathbb{N}$ into the two subsets: the set of all integers n which have a prime factor, and set of those which do not. This latter set was dubbed the set of "criminals" for the sake

[13] See Filip Saidak, A NEW PROOF OF EUCLID'S THEOREM, Amer. Math. Monthly, Vol. 113, No. 9, Nov., 2006, 937–938.

[14] This is a modification of Problem #12 of the American Invitational Mathematics Examination, 2005 (I).

of color. The proof rested on the fact that this set C of criminals must have a **least element**, which meant that any positive integer m which is less than any element of C cannot be a criminal.

Before formalizing the above, let's take up an example of a somewhat different nature. Consider the proposition that, for any $n \in \mathbb{N}$, one has

$$1^2 + 2^2 + 3^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6}.$$

Naturally, for each such n, either the above statement is true or false. This allows us to divide $\mathbb{N}$ into two subsets: the subset G (for "good guys") of integers $n \in \mathbb{N}$ for which the above statement is true, and the set C (for "criminals") for which the above statement is false. Obviously

$$\mathbb{N} \;=\; G \cup C, \quad \text{and} \quad G \cap C = \varnothing.$$

Also — and this is important — note that $1 \in G$, i.e., if $n = 1$, then the above statement is easily verified to be true. Put differently, 1 is not a criminal; it's a good guy!

In order to prove that the above statement is true **for all** $n \in \mathbb{N}$, we need only show that $C = \varnothing$. Thus, let m be the **least element** of C, and note that since $1 \notin C$ we have that $m - 1 \in G$: that is to say the above statement is valid with $n = m - 1$. Watch this:

$$\begin{aligned}
1^2 \;+\; 2^2 + 3^2 + \cdots + m^2 &= 1^2 + 2^2 + 3^2 + \cdots (m-1)^2 + m^2 \\
&= \frac{(m-1)m(2(m-1)+1)}{6} + m^2 \quad \text{(This is the key step!)} \\
&= \frac{1}{6}(2m^3 - 3m^2 + m + 6m^2) \quad \text{(This is just algebra.)} \\
&= \frac{1}{6}(m(m^2 + 3m + 1) = \frac{m(m+1)(2m+1)}{6} \quad \text{(A little more algebra.)}
\end{aligned}$$

Let's have a look at what just happened. We started with the assumption that the integer m is a criminal, the least criminal in fact, and then observed in the end that $1^2 + 2^2 + 3^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6}$, meaning that m **is not a criminal**. This is clearly a contradiction! What caused this contradiction is the fact that there was an element in C, so the only way out of this contradiction is to conclude that $C = \varnothing$.

Therefore every element $n \in \mathbb{N}$ is in G, which means that the above statement is true for all positive integers n.

Let's formalize this a bit. Assume that for each $n \in \mathbb{N}$ we assign a property $P(n)$ to this integer, which may be true or false. In the previous section, the relevant propery was

$$P(n): \quad n \text{ has at least one prime factor.}$$

In the example just discussed,

$$P(n): \quad 1^2 + 2^2 + 3^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6}.$$

The point is that once we have a property assigned to each $n \in \mathbb{N}$, we may consider the set $G \subseteq\in \mathbb{N}$ of all integers n for which $P(n)$ is true, and the set (the criminals) of all integers n for which $P(n)$ is false. In trying to establish that $C = \varnothing$, we may streamline our argument via

Principle of Mathematical Induction. *Let $\mathbb{N}$ denote the set of positive integers, and assume that for each $n \in \mathbb{N}$ we have a property $P(n)$. Assume that*

(i) $P(a)$ is true, for some $a \in \mathbb{N}$. (This "starts" the induction.)

(ii) Whenever $P(m)$ is true for all $a \leq m < n$, (the so-called **inductive hypothesis***) then $P(n)$ is also true.*

Then $P(n)$ is true for all $n \geq a$.

Proof. Let C be the set of all integers $\geq a$ for which $P(n)$ is false. We shall prove that $C = \varnothing$, which will imply that $P(n)$ is true for all $n \in \mathbb{N}$. By hypothesis (i) above, we see that $a \notin C$; therefore, if we take n to be the **least element** of C, then $n \neq a$. Therefore, for any positive integer m with $a \leq m < n$, $P(m)$ must be true. By hypothesis (ii) above, we conclude that, in fact, $P(n)$ must be true, which says that $n \notin C$. This contradiction proves that $C = \varnothing$, and the proof is complete.

At first blush, it doesn't appear that the above principle accomplishes much beyond what we were already able to do. However, it

does give us a convenient language in which to streamline certain arguments. Namely, when we consider an integer n for which $P(m)$ is true for all $m < n$, we typically simply say,

By induction, $P(m)$ is true for all $m < n$.

Let's see how to use this language in the above two examples.

EXAMPLE 1. *Any integer $n \geq 2$ has at least one prime factor.*

PROOF. We shall prove this by induction on $n \geq 2$. Since 2 is a prime factor of itself, we see that the induction starts. Next, assume that n is a given integer. If n is prime then, of course, there's nothing to prove. Otherwise, n factors as $n = ab$, where a and b are positive integers satisfying $2 \leq a, b < n$. By induction a has a prime factor, and hence so does n. Therefore, by the principle of mathematical induction we conclude that every integer $n \geq 2$ has a prime factor and the proof is complete.

EXAMPLE 2. *For any integer $n \geq 1$ one has*

$$1^2 + 2^2 + 3^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6}.$$

PROOF. We prove this by mathematical induction. The above is clearly true for $n = 1$, and so the induction starts. Next, let n be a given integer. By induction we assume that the above recipe is valid for all positive integers $m < n$. We compute:

$$\begin{aligned} 1^2 + 2^2 + 3^2 + \cdots + n^2 &= 1^2 + 2^2 + 3^2 + \cdots + (n-1)^2 + n^2 \\ &= \frac{n(n-1)(2n-1)}{6} + n^2 \quad \text{(by induction)} \\ &= \frac{n(n+1)(2n+1)}{6} \end{aligned}$$

and the proof is complete.

EXERCISES

1. Prove the following:

   (i) $1+3+5+\cdots+(2n-1)=n^2 \quad (n=1,2,\ldots)$

   (ii) $1^3+2^3+3^3+\cdots+n^3=\frac{1}{4}n^2(n+1)^2 \quad (n=1,2,\ldots)$

   (iii) $\dfrac{1}{1\cdot 3}+\dfrac{1}{3\cdot 5}+\cdots\dfrac{1}{(2n-1)(2n+1)}=\dfrac{n}{2n+1} \quad (n=1,2,\ldots).$
   (Do you really need mathematical induction? Try partial fractions!)

   (iv) $1^2+\left(\dfrac{1}{2}\right)^2+\left(\dfrac{1}{3}\right)^2+\cdots+\left(\dfrac{1}{n}\right)^2<2-\dfrac{1}{n} \quad (n=2,3,\ldots)$

2. As in Exercise 6 on page 78 we define, for any positive integer n,

$$H(n)=1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{n}.$$

   Show that for any integer $m\geq 0$, that $H(2^m)\geq\dfrac{m+2}{2}$.

3. Let n be a positive integer.

   (a) Prove that if k is an integer with $0\leq k\leq n$, $\dbinom{n}{k}=\dbinom{n-1}{k}+\dbinom{n-1}{k-1}$. (This doesn't require induction.)

   (b) Prove that if S is a set with n elements, and if $0\leq k\leq n$, then there are $\binom{n}{k}$ subsets of S with k elements. (Use induction.)

4. Prove that for all $n\geq 1.$, $1^3+2^3+\cdots n^3=(1+2+3+\cdots+n)^2$.

5. Prove that for all $n\geq 1$, and for all $x\geq 0$, that $(1+x)^n>1+nx$. (Is induction really needed?)

6. Prove the classical inequality

$$\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}\geq n^2$$

   whenever x_1, x_2, $\ldots x_n>0$ and $x_1+x_2+\cdots x_n=1$. (Hint: using induction, note first that you can arrive at the inequality

$$\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}+\frac{1}{x_{n+1}} \geq \frac{n^2}{1-x_{n+1}}+\frac{1}{x_{n+1}}.$$

Next, you need to argue that because $0 < x_{n+1} < 1$,

$$\frac{n^2}{1-x_{n+1}}+\frac{1}{x_{n+1}} \geq (n+1)^2;$$

this is not too difficult. Incidently, when does equality occur in the above inequality?)

7. Prove that for all integers $n \geq 1$, $2\sum_{j=1}^{n} \sin x \cos^{2j-1} x \;=\; \sin 2nx$.

8. Prove that for all integers $n \geq 0$, $\sin x \prod_{j=0}^{n} \cos 2^j x \;=\; \dfrac{\sin\left(2^{n+1}x\right)}{2^{n+1}}$.

9. Prove that for all integers $n \geq 0$, that $\sum_{j=1}^{n} \sin(2j-1)x \;=\; \dfrac{1-\cos 2nx}{2\sin x}$.

10. (This is a bit harder.) Prove the partial fraction decomposition

$$\frac{1}{x(x+1)(x+2)\cdots(x+n)} = \frac{1}{n!}\sum_{k=0}^{n}(-1)^k\binom{n}{k}\frac{1}{x+k},$$

where n is a non-negative integer.

11. [15] We shall use mathematical induction to prove that all positive integers are equal. Let $P(n)$ be the proposition

$P(n)$: "If the maximum of two positive integers is n then the integers are equal."

[15] Due to T.I. Ramsamujh, THE MATHEMATICAL GAZETTE, Vol. 72, No. 460 (Jun., 1988), p. 113.

Clearly $P(1)$ is true. Assuming that $P(n)$ is true, assume that u and v are positive integers such that the maximum of u and v is $n+1$. Then the maximum of $u-1$ and $v-1$ is n, forcing $u-1=v-1$ by the validity of $P(n)$. Therefore, $u=v$. What's wrong with this argument?

12. If A is a finite subset of real numbers, let $\pi(A)$ be the product of the elements of A. If $A=\emptyset$, set $\pi(A)=1$. Let $S_n=\{1,2,3,\ldots,n\},\ n\geq 1$ and show that

    (a) $\displaystyle\sum_{A\subseteq S_n}\frac{1}{\pi(A)}=n+1$, and that

    (b) $\displaystyle\sum_{A\subseteq S_n}\frac{(-1)^{|A|}}{\pi(A)}=0$

13. If A is a finite subset of real numbers, let $\sigma(A)$ be the sum of the elements of A. Let $n\geq 1$, and set $S_n=\{1,2,3,\ldots,n\}$, as above. Show that

    (a)[16] $\displaystyle\sum_{A\subseteq S_n}\frac{\sigma(A)}{\pi(A)}=(n^2+2n)-(n+1)\left(1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{n}\right)$, and that

    (b) $\displaystyle\sum_{A\subseteq S_n}\frac{(-1)^{|A|}\sigma(A)}{\pi(A)}=-\frac{1}{n}$

2.1.6 Fermat's and Euler's theorems

We start with a potentially surprising observation. Namely we consider integers a not divisible by 7 and consider powers a^6, reduced modulo 7. Note that we may, by the division algorithm, write $a=7q+r$, where since a is not divisible by 7, then $1\leq r\leq 6$. Therefore, using the binomial theorem, we get

[16] This is Problem #2 on the 20th USA Mathematical Olympiad, April 23, 1991. It's really not that hard!

$$a^6 = \sum_{k=0}^{6} \binom{6}{k} (7q)^k r^{6-k} \equiv r^6 (\text{mod } 7).$$

This reduces matters to only six easily verifiable calculations:

$$1^6 \equiv 1(\text{mod } 7),\ 2^6 \equiv 1(\text{mod } 7),\ 3^6 \equiv 1(\text{mod } 7),$$

$$4^6 \equiv (-3)^6 \equiv 1(\text{mod } 7),\ 5^6 \equiv (-2)^6 \equiv 1(\text{mod } 7),\ 6^6 \equiv (-1)^6 \equiv 1(\text{mod } 7).$$

In other words, for any integer a not divisible by 7, we have $a^{p-1} \equiv 1(\text{mod } 7)$.

In order to generalize the above result, we shall first make the following observation, namely that if x and y are arbitrary integers, and if p is a prime number, then using exercise 9 on page 62 we get

$$\begin{aligned} (x+y)^p &\equiv \sum_{k=0}^{p} \binom{p}{k} x^k y^{p-k} \\ &\equiv x^p + y^p (\text{mod } p). \end{aligned}$$

That is to say, for any integers x and y and any prime number p, we have

$$\boxed{(x+y)^p \equiv x^p + y^p (\text{mod } p)}.$$

THEOREM. (Fermat's Little Theorem) *Let p be a prime number. Then for all integers a not divisible by p we have*

$$a^{p-1} \equiv 1 (mod\ p).$$

PROOF. There are a number of proofs of this fact;[17] perhaps the most straightforward is based on the Binomial Theorem together with the

[17]It is interesting to note that while Fermat first observed this result in a letter in 1640, the first known complete proof was not given until 1736 by Leonard Euler.

above observation. Note first that it suffices to assume that $a \geq 1$; we shall argue by induction on a. Note that if $a = 1$ the result is clearly valid. Next, assuming that $a > 1$, then by induction we may assume that $(a-1)^p \equiv (a-1) (\text{mod } p)$. From this we proceed:

$$\begin{aligned} a^p &\equiv ((a-1)+1)^p \\ &\equiv (a-1)^p + 1^p \ \text{ (by the above result)} \\ &\equiv a - 1 + 1 \ \text{ (by induction)} \\ &\equiv a \, (\text{mod } p), \end{aligned}$$

which completes the proof.

There is a striking generalization of Fermat's Little Theorem, as follows. I won't prove this here as the most natural proof of this is within the context of group theory. Anyway, recall the Euler ϕ-function (see Exercise 16 on page 63), defined by setting

$\phi(n) \ = \ \#$ of integers $m, \ 1 \leq m < n$ which are relatively prime with n.

This obviously says, in particular that if p is prime then $\phi(p) = p - 1$.

THEOREM. (Euler's Theorem) *Let n be any positive integer. Then for any integer a with* $\gcd(a, n) = 1$ *we have*

$$a^{\phi(n)} \equiv 1 (mod \ n).$$

Note that Euler's Theorem obviously contains Fermat's Little Theorem as a corollary.

EXERCISES

1. Compute the units digit of $(23)^{987}$
2. Compute the least positive integer solution of $n \equiv 123^{139} (\text{mod } 7)$.
3. Compute the least positive integer solution of $n \equiv 506^{10^6} (\text{mod } 11)$.

4. Let p be a prime number. The integers a and b are said to be **multiplicative inverses modulo** p if $ab \equiv 1 (\text{mod } p)$. Using the Euclidean trick, prove that if p doesn't divide a, then a has a multiplicative inverse modulo p.

5. Find the multiplicative inverse of 2 modulo 29.

6. Find the multiplicative inverse of 3 modulo 113.

7. Prove Wilson's Theorem:

$$(p-1)! \equiv -1 (\text{mod } p),$$

where p is a prime. (Hint; pair each divisor of $(p-1)!$ with its inverse modulo p; of course, this requires the result of exercise 4, above.)

8. The **order** of the integer a modulo the prime p is the *least* positive integer n such that $a^n \equiv 1 (\text{mod } p)$. Show that $n \,|\, p-1$. (Hint: show that if $d = \gcd(n, p-1)$, then $a^d \equiv 1 (\text{mod } p)$.)

9. As we saw from Fermat's little theorem, if p is prime and if a is an integer not divisible by p, then $a^{p-1} \equiv 1 \,(\text{mod}\, p)$. What about the converse? That is, suppose that n is a positive integer and that for every integer a relatively prime to n we have $a^{n-1} \equiv 1 \,(\text{mod}\, n)$. Must n then be prime? Looking for a counter example takes some time, leading one to (almost) believe this converse. However, suppose that we were to find a candidate integer n and found that **for every prime divisor** p **of** n, that $p-1 \,|\, n-1$. Show that n satisfies the above.[18]

10. Here's a very surprising application of Euler's Theorem, above.[19] Define the sequence $a_1, a_2, \ldots$, by setting $a_1 = 2$, $a_2 = 2^{a_1}$, $a_3 = 2^{a_2}, \ldots$. Then for any integer n, the sequence $a_1, a_2, \ldots$, eventually becomes constant (mod n). The proof proceeds by induction on n and can be carried out along the following lines.

[18] Such an integer is called a **Carmichael number**, the first such being $n = 561$, which is why the converse to Fermat's little theorem can appear true! It is known that there are, in fact, infinitely many Carmichael numbers, which means that there are infinitely many counter examples to the converse of Fermat's little theorem.

[19] I'm indebted to my student, Nelson Zhang, for pointing out this exercise, commenting also that this is Problem #3 on the 1991 USA Olympiad contest. The hints given above are the result of our discussion.

(a) Since $\phi(n) < n$ for all n, we see that the sequence a_1, a_2, ..., eventually becomes constant modulo $\phi(n)$.

(b) Write $n = 2^r k$, where k is an odd integer. Since a_1, a_2, ..., eventually becomes constant modulo $\phi(n)$, it also eventually becomes constant modulo $\phi(k)$.

(c) Conclude from Euler's Theorem (87) that a_1, a_2, ..., eventually becomes constant modulo k.

(d) Argue that a_1, a_2, ..., eventually becomes constant modulo 2^r and hence eventually becomes constant modulo n.

2.1.7 Linear congruences

A linear congruence is of the form $ax \equiv b(\text{mod n})$, where a, b, n are integers, $n > 0$, and x is regarded as unknown. In order to solve this equation, we would hope that a would have an inverse modulo n. In other would if there exists an integer a' such that $a'a \equiv 1(\text{mod } n)$, then we can solve the above congruence by multiplying through by a':

$$x \equiv a'b \,(\text{mod } n).$$

Next, if a and n are relatively prime, then we can employ the Euclidean trick and write

$$sa + tn = 1,$$

for suitable integers s and t. But this says already that

$$sa = 1 - tn \equiv 1(\text{mod } n),$$

i.e., that $a' = s$ is the desired inverse of a modulo n.

EXAMPLE. Solve the congruence $5x \equiv 14(\text{mod } 18)$.

SOLUTION. We employ the Euclidean algorithm:

$$\begin{aligned} 18 &= 3 \cdot 5 + 3 \\ 5 &= 1 \cdot 3 + 2 \\ 3 &= 1 \cdot 2 + 1. \end{aligned}$$

Now work backwards and get

$$2 \cdot 18 - 7 \cdot 5 = 1.$$

This says that the inverse of 5 modulo 18 is -7. Therefore we see that the solution of the above is

$$x \equiv -7 \cdot 14 \equiv (-7)(-4) \equiv 28 \equiv 10 (\text{mod } 18).$$

EXERCISE

1. Solve the linear congruences
   (a) $17x \equiv 4 (\text{mod } 56)$
   (b) $26x \equiv 7 (\text{mod } 15)$
   (c) $18x \equiv 9 (\text{mod } 55)$

2.1.8 Alternative number bases

In writing positive integers, we typically write in **base 10**, meaning that the digits represent multiples of powers of 10. For instance, the integer 2,396 is a compact way of writing the sum

$$2,396 = 6 \cdot 10^0 + 9 \cdot 10^1 + 3 \cdot 10^3 + 2 \cdot 10^2.$$

In a similar way, decimal numbers, such as 734.865 likewise represent sums of (possibly negative) powers of 10:

$$734.865 = 5 \cdot 10^{-3} + 6 \cdot 10^{-2} + 8 \cdot 10^{-1} + 4 \cdot 10^0 + 3 \cdot 10^1 + 7 \cdot 10^2.$$

The coefficients are called the (decimal) **digits**.

Arguably the second-most popular number base is 2, giving **binary numbers** (or binary **representations** of numbers). In this case the binary digits include only "0" and "1". As an example, we can convert a binary number such as 1001101 into its equivalent decimal number by computing the relevant powers of 2:

$$1001101 = 1 \cdot 2^0 + 0 \cdot 2^1 + 1 \cdot 2^2 + 1 \cdot 2^3 + 0 \cdot 2^4 + 0 \cdot 2^5 + 1 \cdot 2^6 = 77.$$

Another way of expressing this fact is by writing $77_2 = 1001101$, meaning that the binary representation of the decimal number 77 is 1001101.

EXAMPLE 1. Find the binary representation of the decimal number 93.
SOLUTION. First notice that the highest power of 2 less than or equal to 93 is 2^6. Next, the highest power of two less than or equal to $93-2^6$ is 2^4. Continuing, the highest power of 2 less than or equal to $93-2^6-2^4$ is 2^3. Eventually we arrive at $93 = 2^6 + 2^4 + 2^3 + 2^2 + 1$, meaning that $93_2 = 1011101$.

EXAMPLE 2. Find the binary representation of 11111. Note first that if n is the number of binary digits required, then after a moment's thought one concludes that $n - 1 \leq \log_2 11111 < n$. Since $\log_2 11111 = \dfrac{\ln 11111}{\ln 2} \approx 13.44$, we conclude that 11111 will require 14 binary digits. That is to say, $11111 = 2^{13} +$ lower powers of 2. Specifically, one shows that

$$11111 = 2^{13} + 2^{11} + 2^9 + 2^8 + 2^6 + 2^5 + 2^2 + 2 + 1.$$

That is to say, $11111_2 = 10101101100111$.

As one would expect, there are *b-ary* representations for any base. For example a **trinary** representation would be a representation base 3, and the number n of trinary digits needed to represent m would satisfy $n - 1 \leq \log_3 m \leq n$.

EXAMPLE 3. The representation of 11111 in trinary would require 9 trinary digits since $\log_3 11111 \approx 8.48$. Specifically,

$$11111 \;=\; 3^8 + 2 \cdot 3^7 + 2 \cdot 3^4 + 3^2 + 3 + 2,$$

which says that $11111_3 = 12\,002\,0112$.

In computer science numbers are sometimes representation in hexadecimal notation (base 16); the "digits" used are 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F. Therefore $17_{16} = 11,\ 15_{16} = \mathrm{F}, 206_{16} = \mathrm{CE}$

EXERCISES

1. Compute representations of 1435
   (a) in binary;
   (b) in ternary;
   (c) in quarternary (4-*ary*)
   (d) in hexadecimal
2. Compute representations of 10,000
   (a) in binary;
   (b) in ternary;
   (c) in quarternary (4-*ary*)
   (d) in hexadecimal
3. The largest known Mersenne prime[20] is the number $2^{43,112,609} - 1$. Compute the number of decimal digits needed to represent this huge prime number. Compute the number of binary digits (trivial) and the number of ternary digits needed for its representation.
4. Here's a bit of a challenge. Represent the decimal .1 ($= \frac{1}{10}$) in binary. What makes this a bit of a challenge is that in binary, the decimal representation is an infinite repeating decimal (or should I say "bi-cimal"?). As a hint, note that $10_2 = 1010$. Now do a long division into 1.[21]

[20] As of August, 2008; this is a prime of the form $2^p - 1$, where p is prime.
[21] The answer is $.000\overline{1100}$.

2.1.9 Linear recurrence relations

Many, if not most reasonably serious students have heard of the **Fibonacci sequence**[22] the first few terms of which are

$$1,\ 1,\ 2,\ 3,\ 5,\ 8,\ 13,\ 21,\ 34, \ldots$$

Even one who hasn't had much exposure to mathematics can easily guess the successive term in the sequence given the previous two terms as it is clear that this term is the sum of the previous two terms. Put more mathematically, if u_n denotes the n-th term, then one has the **linear difference equation**

$$u_{n+2} = u_{n+1} + u_n,\ n = 1, 2, \ldots,\quad u_0 = 1,\ u_1 = 1.$$

More elementary sequences come from the familiar **arithmetic** and **geometric** sequences. Arithmetic sequences are generated by difference equations of the form $u_{n+1} = u_n + d,\ n = 0, 1, 2, \ldots$, where d is a constant. Geometric sequences come from the difference equation $u_{n+1} = ku_n,\ n = 0, 1, 2, \ldots$. The general term for the arithmetic and geometric sequences can be easily solved for in terms of u_0:

$$\text{Arithmetic: } u_{n+1} = u_n + d,\ n = 1, 2, \ldots \Longrightarrow u_n = u_0 + nd.$$

$$\text{Geometric: } u_{n+1} = ku_n,\ n = 1, 2, \ldots \Longrightarrow u_n = k^n u_0.$$

The above three difference equations are linear in the sense that none of the unknown terms u_n occur to powers other than 1. A very famous nonlinear recurrence relation is the so-called **logistic recurrence equation** (or "Logistic map"), given by a relation of the form

$$u_{n+1} = k(1 - u_n)u_n,\ n = 0, 1, 2, \ldots.$$

[22] which made a cameo appearance in the movie, *The Da Vinci Code*.

For certain values of k, the above sequence can exhibit some very strange—even chaotic—behavior!

The **general homogeneous linear difference equation** of **order** k has the form

$$u_{n+k} = a_1 u_{n+k-1} + a_2 u_{n+k-2} + \cdots + a_k u_n, \quad n = 0, 1, 2, \ldots$$

Of fundamental importance is the associated **characteristic polynomial**

$$C(x) = x^k - a_1 x^{k-1} - a_2 x^{k-2} - \cdots - a_k.$$

The **charasteristic equation** finds the zeros of the **characteristic polynomial**:

$$x^k - a_1 x^{k-1} - a_2 x^{k-2} - \cdots - a_k = 0.$$

Given the monic[23] polynomial

$$C(x) \;=\; x^k - a_1 x^{k-1} - a_2 x^{k-2} - \cdots - a_k,$$

with real coefficients, and if $\mathbf{u} = (u_n)$ is a sequence, we shall denote by $C(\mathbf{u})$ the sequence $\mathbf{u}' = (u'_n)_{n\geq 0}$ where

$$u'_n \;=\; u_{n+k} - a_1 u_{n+k-1} - a_2 u_{n+k-2} - \cdots - a_k u_n.$$

Therefore, the task of solving a linear difference equation is to solve

$$C(\mathbf{u}) = \mathbf{v},$$

where $\mathbf{v} = (v_n)_{n\geq 0}$ is a given sequence. If $\mathbf{v} = \mathbf{0}$ (the sequence all of whose terms are 0) we call the difference equation **homogeneous**. We shall be primarily concerned with homogeneous difference equations;

[23] "Monic" simply means that the leading coefficient is 1.

note, however that the difference equations leading to arithmetic sequences ($u_{n+1} - u_n = d,\ n = 0, 1, 2, \ldots$) are not homogeneous. We'll treat generalizations of the arithmetic sequences in Section 2.1.9, below.

We shall now separate the homogeneous and inhomogeneous cases:[24]

Homogeneous difference equations

We shall consider a few commonly-occuring cases.

Linear. Given the monic polynomial $C(x)$ we are trying to solve $C(\mathbf{u}) = 0$ for the unknown sequence $\mathbf{u} = (u_0, u_1, u_2, \ldots)$. Assume that the polynomial is linear: $C(x) = x - k$, for some real constant k; thus the difference equation assumes the form

$$u_{n+1} = ku_n, \quad n = 0, 1, 2, \ldots \tag{2.2}$$

This says that each successive term is k times the preceding term; this is the definition of a **geometric sequence** with **ratio** k. Clearly, then the solution is

$$u_n = k^n A, \quad n = 0, 1, 2, \ldots \tag{2.3}$$

where A is an arbitrary constant. The solution given in equation (2.3) above is called the **general solution** of the first-order difference equation (2.2). The **particular solution** is then obtained by specifying a particular value for A.

[24]The reader having studied some linear differential equations will note an obvious parallel!

Quadratic—distinct factors over the reals. Next, assume that our polynomial $C(x)$ is quadratic; $C(x) = x^2 - ax - b$, where $a, b \in \mathbb{R}$. Thus, we are trying to solve the **second-order** homogeneous difference equation

$$u_{n+2} = au_{n+1} + bu_n,\ n = 0, 1, 2, \ldots \tag{2.4}$$

Assume furthermore that $C(x)$ factors into two **distinct** real linear factors:

$$C(x) = (x - k_1)(x - k_2),\ k_1 \neq k_2 \in \mathbb{R}.$$

In this case it turns out that we both $u_n = k_1^n A_1,\ n = 0, 1, 2, \ldots$ and $u_n = k_2^n A_2,\ n = 0, 1, 2, \ldots$, where $A_1, A_2 \in \mathbb{R}$ are both solutions of (2.4). This is verified by direct substitution: if $u_n = k_1^n A_1,\ n = 0, 1, 2, \ldots$, then

$$\begin{aligned} u_{n+2} - au_{n+1} - bu_n &= k_1^{n+2} A_1 - ak_1^{n+1} A_1 - bk_1^n A_1 \\ &= k_1^n A_1 (k_1^2 - ak_1 - b) \\ &= k_1^n A_1 (k_1 - k_1)(k_1 - k_2) = 0. \end{aligned}$$

This proves that $u_n = k_1^n A_1,\ n = 0, 1, 2, \ldots$ is a solution. Likewise, $u_n = k_2^n A_2,\ n = 0, 1, 2, \ldots$ is another solution. However, what might seem surprising is that the sum

$$u_n = k_1^n A_1 + k_2^n A_2,\ n = 0, 1, 2, \ldots \tag{2.5}$$

of these solutions is also a solution of (2.4). Again, this is proved by a direct substitution:

$$\begin{aligned} u_{n+2} - au_{n+1} - bu_n &= k_1^{n+2} A_1 + k_2^{n+2} A_2 - a(k_1^{n+1} A_1 + k_2^{n+1} A_2) - b(k_1^n A_1 + k_2^n A_2) \\ &= k_1^{n+2} A_1 - ak_1^{n+1} A_1 - bk_1^n A_1 + k_2^{n+2} A_2 - ak_2^{n+1} A_2 - bk_2^n A_2 \\ &= k_1^n A_1 (k_1^2 - ak_1 - b) + k_2^n A_2 (k_2^2 - ak_2 - b) \\ &= k_1^n A_1 (k_1 - k_1)(k_1 - k_2) + k_1^n A_2 (k_2 - k_1)(k_2 - k_2) = 0 + 0 = 0. \end{aligned}$$

Finally, one can show that *any* solution of (2.4) is of the form given in (2.5). We won't belabor these details any further.

EXAMPLE 1. Solve the second-order linear homogeneous difference equation

$$u_{n+2} = u_{n+1} + 2u_n \; n = 0, 1, 2, \ldots$$

given that $u_0 = 0$ and $u_1 = 1$.

SOLUTION. Note first that writing down the first few terms of the sequence is easy:

$$\begin{aligned} u_2 &= u_1 + 2u_0 = 1 + 0 = 1 \\ u_3 &= u_2 + 2u_1 = 1 + 2 = 3 \\ u_4 &= u_3 + 2u_2 = 3 + 2 = 5 \\ u_5 &= u_4 + 2u_3 = 5 + 6 = 11 \end{aligned}$$

and so on. In other words, the first few terms of the sequence look like

$$u_n = 0,\ 1,\ 1,\ 3,\ 5,\ 1,\ \ldots.$$

What we're trying to find, however, is a recipe for the general term. Since the characteristic polynomial of this difference equation is $C(x) = x^2 - x - 2 = (x+1)(x-2)$, we conclude by equation (2.5) that the solution must look like

$$u_n = A_1 2^n + A_2(-1)^n, \; n = 0, 1, 2, \ldots$$

where A_1 and A_2 are constants. However, since $u_0 = 0$ and $u_1 = 1$ we obtain

$$\begin{aligned} 0 &= u_0 = A_1 2^0 + A_2(-1)^0 = A_1 + A_2 \\ 1 &= u_1 = A_1 2^1 + A_2(-1)^1 = 2A_1 - A_2 \end{aligned}$$

all of which implies that $A_1 = \frac{1}{3},\ A_2 = -\frac{1}{3}$. The particular solution of the above linear difference equation is therefore

$$u_n = \frac{2^n}{3} - \frac{(-1)^n}{3},\ n = 0, 1, 2, \ldots$$

Quadratic—repeated factor over the reals.

Here we assume that our polynomial $C(x)$ is quadratic with a multiple factor: $C(x) = x^2 - 2kx - k^2 = (x-k)^2$, where $k \in \mathbb{R}$. As in the above case, one solution has the form $u_n = Ak^n\ n = 0, 1, 2, \ldots$. However, a second solution has the form $u_n = Bnk^n,\ n = 0, 1, 2, \ldots$. We check this by direct substitution:

$$\begin{aligned} u_{n+2} - 2ku_{n+1} + k^2 u_n &= B(n+2)k^{n+2} - 2kB(n+1)k^{n+1} + nBk^2k^n \\ &= Bk^{n+2}((n+2) - 2(n+1) + n) = 0. \end{aligned}$$

Likewise, one than then show that the sum of these solutions is a solution of the second-order homogeneous difference equation:

$$u_n = Ak^n + Bnk^n,\ n = 0, 1, 2, \ldots.$$

Quadratic—irreducible. In this case we consider the second-order linear homogeneous difference equation whose characteristic equation is irreducible (over the reals). Thus the discriminant of the characteristic polynomial is negative (and has complex conjugate zeros). A simple example of such would be the difference equation

$$u_{n+2} \;=\; -u_{n+1} - u_n,\ n = 0, 1, 2, \ldots,$$

since the characteristic polynomial $C(x) = x^2 + x + 1$ is irreducible over the real numbers.

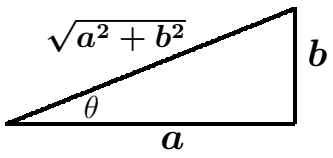
Assume now that we have the second-order homogeneous linear difference equation (2.4) has characteristic polynomial with two *complex* zeros $a + bi$ and $a - bi$, where a, $b \in \mathbb{R}$, and $b \neq 0$. Using the same argument in as in the previous section, we may conclude that a **complex** solution of (2.4) is

$$u_n = A(a + bi)^n,\ n = 0, 1, 2, \ldots,$$

where A is any real constant. However, since the coefficients in the equation (2.4) are *real* one may conclude that the **real** and **imaginary** parts of the above complex solution are also solutions. Therefore, we would like to find the real and imaginary parts of the powers $(a + bi)^n,\ n \geq 0$. To do this we write the complex number $a + bi$ in **trigonometric form**. We start by writing

$$a + bi = \sqrt{a^2 + b^2}\left(\frac{a}{\sqrt{a^2 + b^2}} + \frac{bi}{\sqrt{a^2 + b^2}}\right).$$

Next let θ be the angle represented below:



Therefore, $a + bi = \cos\theta + i\sin\theta$, from which one concludes[25] that

$$(a + bi)^n \;=\; (\cos\theta + i\sin\theta)^n = \cos n\theta + i\sin n\theta.$$

[25]This is usually called *DeMoivre's Theorem*, and can be proved by a repeated application of the addition formulas for sine and cosine.

That is to say, the real and imaginary parts of $(a+bi)^n$ are $\cos n\theta$ and $\sin n\theta$, where θ is as above. From this, one finally concludes that the solution of (2.4) in the present case has the form

$$u_n \;=\; A\cos n\theta + B\sin n\theta,\;\; n = 0, 1, 2, \ldots,$$

where A and B are real constants.

It's time to come up for air and look at an example.

EXAMPLE 2. Solve the second-order homogeneous difference equation

$$u_{n+2} = -u_{n+1} - u_n,\;\; n = 0, 1, 2, \ldots, \tag{2.6}$$

where $u_0 = 1,\;\; u_1 = 1.$

SOLUTION. The characteristic polynomial $C(x) = x^2+x+1$ which has zeros $\dfrac{-1+i\sqrt{3}}{2}$ and $\dfrac{-1-i\sqrt{3}}{2}$. We write the first complex number in trigonometric form

$$\frac{-1+i\sqrt{3}}{2} \;=\; \cos\frac{2\pi}{3} + i\sin\frac{2\pi}{3},$$

from which it follows that

$$\left(\frac{-1+i\sqrt{3}}{2}\right)^n \;= \cos\frac{2\pi n}{3} + i\sin\frac{2\pi n}{3}.$$

From this it follows that the general solution is given by

$$u_n \;=\; A\cos\frac{2\pi n}{3} + B\sin\frac{2\pi n}{3},\;\; n = 0, 1, 2, \ldots.$$

However, given that $u_0 = 0,\ u_1 = 1$, we get

$$\begin{aligned} 0 &= A \\ 1 &= A\cos\frac{2\pi}{3} + B\sin\frac{2\pi}{3} = -\frac{A}{2} + \frac{\sqrt{3}\,B}{2} \end{aligned}$$

Therefore $A = 0$ and $B = \dfrac{2}{\sqrt{3}}$, forcing the solution to be

$$u_n = \frac{2}{\sqrt{3}}\sin\frac{2\pi n}{3},\ n = 0, 1, 2, \ldots.$$

Higher-degree characteristic polynomials.

We won't treat this case systematically, except to say that upon factoring the polynomial into irreducible linear and quadratic factors, then one can proceed as indicated above (see Exercise 14). Additional complications result with higher-order repeated factors which we don't treat here.

Higher-order differences

In Section 2.1.9 we treated only the so-called **homogeneous** linear difference equations. An **inhomogeneous** linear difference equation has the general form

$$C(\mathbf{u}) = \mathbf{v},$$

where $C(x)$ is a monic polynomial, $\mathbf{v} = (v_n)_{n\geq 0}$ is a given sequence and where $\mathbf{u} = (u_n)_{n\geq 0}$ is the unknown sequence.

We have already encountered such an example above, in the example on page 312 giving an arithmetic sequence:

$$u_{n+1} - u_n = d,\ n = 0, 1, 2, \ldots.$$

We won't treat inhomogeneous linear difference equations in any detail except for a very special case, namely those having **constant**

higher-order differences. The **arithmetic sequences** have constant first-order differences; if d is this difference then we have $u_{n+1} - u_n = d$, $n = 0, 1, 2, \ldots$. Suppose next that the second-order differences are constant: this means that the difference of the difference is constant, written as

$$(u_{n+2} - u_{n+1}) - (u_{n+1} - u_n) = d, \ n = 0, 1, 2, \ldots.$$

In other words,

$$u_{n+2} - 2u_{n+1} + u_n = d, \ n = 0, 1, 2, \ldots.$$

Writing more compactly and in terms of the characteristic polynomial, we have

$$C(\mathbf{u}) = \mathbf{d}, \ n = 0, 1, 2, \ldots, \text{where} \ \ C(x) = (x-1)^2,$$

and where $\mathbf{d}$ is the constant sequence d, d, ...

Constant third-order differences with constant difference d would be expressed as

$$((u_{n+3}-u_{n+2})-(u_{n+2}-u_{n+1}))-((u_{n+2}-u_{n+1})-(u_{n+1}-u_n)) = d, \ n = 0, 1, 2, \ldots,$$

i.e.,

$$u_{n+3} - 3u_{n+2} + 3u_{n+1} - u_n = d, \ n = 0, 1, 2, \ldots.$$

Again, a compact representation of this difference equation is

$$C(\mathbf{u}) = \mathbf{d}, \ n = 0, 1, 2, \ldots, \text{where} \ \ C(x) = (x-1)^3.$$

Continuing along these lines we see that a sequence with finite k-th order differences can be expressed via

$$C(\mathbf{u}) = \mathbf{d}, \ n = 0, 1, 2, \ldots, \text{where} \ \ C(x) = (x-1)^k. \tag{2.7}$$

Such difference equations can be solved in principle; in fact the general solution of (2.7) can be expressed as a polynomial. We shall summarize as a theorem below.[26]

THEOREM. *A sequence $u_0, u_1, u_2, \ldots$ is expressible as a polynomial of degree k in n if and only if its k-th order differences are constant.*

PROOF. Assume that the k-th order differences of the sequence $u_0, u_1, u_2, \ldots$ are constant. We shall prove by induction on k that u_n is expressible as a polynomial of degree k in n. So assume that $k > 1$ and that the result is valid whenever we have constant m-th order differences, where $m < n$ is a positive integer.

We set $v_0 = u_1 - u_0$, $v_1 = u_2 - u_1$, $v_2 = u_3 - u_2$, ..., then we have a sequence whose $(k-1)$-st order differences are constant. By induction, we have a representation of the form

$$v_n = b_{k-1}n^{k-1} + b_{k-2}n^{k-2} + \cdots + b_1 n + b_0,$$

for suitable constants b_0, b_1, b_2, ..., b_{k-1}.

Next — and this is the relatively difficult part — let a_1, a_2, ..., a_k be the unique solution of the linear equations represented in matrix form:

$$\begin{bmatrix} \binom{1}{1} & \binom{2}{2} & \binom{3}{3} & \cdots & \binom{k-1}{k-1} & \binom{k}{k} \\ 0 & \binom{2}{1} & \binom{3}{2} & \cdots & \binom{k-1}{k-2} & \binom{k}{k-1} \\ 0 & 0 & \binom{3}{1} & \cdots & \binom{k-1}{k-3} & \binom{k}{k-2} \\ \vdots & & & & \vdots & \vdots \\ & & & & \binom{k-1}{1} & \binom{k}{2} \\ 0 & 0 & 0 & \cdots & 0 & \binom{k}{1} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_k \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_{k-1} \end{bmatrix}.$$

Having solved the above, one then verifies that the equation

$$a_k(n+1)^k + a_{k-1}(n+1)^{k-1} + \cdots + a_1(n+1)$$

$$= a_k n^k + a_{k-1}n^{k-1} + \cdots + a_1 n + b_{k-1}n^{k-1} + b_{k-2}n^{k-2} + \cdots b_1 n + b_0.$$

[26] As an alternative to using the theorem, note that if a sequence $\mathbf{u} = (u_n)$ has constant k-th order differences, then, in fact, $\mathbf{u}$ satisfies the **homogeneous** difference equation $C(\mathbf{u}) = 0$, where $C(x) = (x-1)^{k+1}$. One can now proceed along the lines of the "repeated factor" case, given above.

Finally, we set $a_0 = u_0$ and use the fact that $u_{n+1} = u_n + v_n$ to check that

$$u_n = a_k n^k + a_{k-1} n^{k-1} + \cdots + a_1 n + a_0,$$

and we are finished, as we have successfully proved that the terms of the sequence $u_0, u_1, \vdots$, are expressible as a polynomial of degree k.

As for the converse, we assume that the sequence $u_0, u_1, u_2, \ldots$ is be expressible as a polynomial of degree k in n:

$$u_n = a_k n^k + a_{k-1} n^{k-1} + \cdots + a_1 n + a_0;$$

we shall show by induction on k that the k-th order differences are constant. To this end, let

Note next that the first order differences are

$$\begin{aligned} v_n &= u_{n+1} - u_n \\ &= (a_k(n+1)^k + a_{k-1}(n+1)^{k-1} + \cdots + a_0) \\ &\quad -(a_k n^k + a_{k-1} n^{k-1} + \cdots + a_0) \\ &= \text{ polynomial in } n \text{ of degree at most } k-1. \end{aligned}$$

By induction, the sequence $(v_n)_{n\geq 0}$ has constant $(k-1)$-st differences. But then it follows immediately that the sequence $(u_n)_{n\geq 0}$ must have constant k-th order differences, and we are done!

EXAMPLE 3. Solve the inhomogeneous linear difference equation

$$u_{n+2} - 2u_{n+1} + u_n = 1,\ n = 0, 1, 2, \ldots,\ u_0 = 2,\ u_1 = 4.$$

SOLUTION. The difference equation says that the second-order differences are constant and equal to 1; this implies that the sequence must be quadratic, say

$$u_n = an^2 + bn + c,\ n = 0, 1, 2, \ldots.$$

Note first that we can solve for the leading coefficient a by substituting the polynomial $an^2 + bn + c$ into the above difference and noting that the linear terms $(bn + c)$ have zero second-order differences and hence don't contribute. This gives

$$a(n+2)^2 - 2a(n+1)^2 + an^2 = 1,$$

which quickly reduces to $2a = 1$, so $a = \frac{1}{2}$. Next, we find b and c by using the initial conditions:

$$\begin{aligned} c &= 2 \\ \frac{1}{2} + b + c &= 4. \end{aligned}$$

This quickly leads to $b = \frac{3}{2},\ c = 2$ and so the solution is given by

$$u_n = \tfrac{1}{2}n^2 + \tfrac{3}{2}n + 2,\ n = 0, 1, 2, \ldots.$$

EXERCISES

1. Let $(u_n)_{n\geq 0}$ be an arithmetic sequence. Prove that the sequence $(e^{u_n})_{n\geq 0}$ is a geometric sequence.

2. Let $(u_n)_{n\geq 0}$ be a geometric sequence with $u_n > 0$ for all n. Prove that $(\log u_n)_{n\geq 0}$ is an arithmetic sequence.

3. Consider the "counting sequence" 1, 2, 3, ...

   (a) Represent this sequence as the solution of an inhomogeneous first-order linear difference equation.

   (b) Represent this sequence as the solution of a homogeneous second-order linear difference equation. Find the general solution of this second-order difference equation.

4. Solve the linear difference equation $u_{n+1} = -2u_n,\ n = 0, 1, 2, \ldots$, where $u_0 = 2$

5. Solve the second-order difference equation $u_{n+2} = -4u_{n+1} + 5u_n,\ n = 0, 1, 2, \ldots$ where $u_0 = 1 = u_1$.

6. Solve the second-order difference equation $u_{n+2} = -4u_{n+1} - 4u_n,\ n = 0, 1, 2, \ldots$ where $u_0 = 1,\ u_1 = 0$.

7. Solve the **Fibonacci** difference equation $u_{n+2} = u_{n+1} + u_n,\ n = 0, 1, 2, \ldots$ where $u_0 = u_1 = 1$.

8. Let $F(n),\ n = 0,\ 1,\ 2,\ \ldots$ be the Fibonacci numbers. Use your result from Exercise #7 to compute the number of digits in $F(1000000)$. (Hint: use $\log_{10}$ and focus on the "dominant term.")

9. Consider the "generalized Fibonacci sequence," defined by $u_0 = 1,\ u_1 = 1$, and $u_{n+2} = au_{n+1} + bu_n,\ n \geq 0$; here a and b are positive real constants.

   (a) Determine the conditions on a and b so that the generalized Fibonacci sequence remains bounded.

   (b) Determine conditions on a and b so that $u_n \to 0$ as $n \to \infty$.

10. The **Lucas numbers** are the numbers $L(n),\ n = 0,\ 1,\ 2,\ \ldots$ where $L(0) = 2,\ L(1) = 1,$ and where (just like the Fibonacci numbers) $L(n+2) = L(n+1) + L(n),\ n \geq 0$. Solve this difference equation, thereby obtaining an explicit formula for the Lucas numbers.

11. Let $F(n),\ L(n),\ n \geq 0$ denote the Fibonacci and Lucas numbers, respectively. Show that for each $n \geq 1,\ L(n) = F(n+1)+F(n-1)$.

12. Solve the second-order difference equation $u_{n+2} = -4u_n,\ n = 0, 1, 2, \ldots$ where $u_0 = 1 = u_1$.

13. Solve the second-order difference equation $u_{n+2} = 2u_{n+1} - 2_n,\ n = 0, 1, 2, \ldots,$
    $u_0 = 0,\ u_1 = 2.$

14. Solve the third-order difference equation $u_{n+3} = -3u_{n+2} + u_{n+1} + u_n,\ n = 0, 1, 2, \ldots,$
    $u_0 = 1, u_1 = 1, u_2 = -1.$

15. Solve the inhomogeneous linear difference equation

$$u_{n+2} - 2u_{n+1} + u_n = 2,\ n = 0, 1, 2, \ldots,\ u_0 = 2,\ u_1 = 6,\ u_2 = 12.$$

16. Solve the inhomogeneous linear difference equation

$$u_{n+3} - 3u_{n+2} + 3u_{n+1} - u_n = 2,\ n = 0, 1, 2, \ldots,$$

$$u_0 = 0,\ u_1 = 4,\ u_2 = 10,\ u_3 = 20.$$

17. Given the sequence $u_0, u_1, u_2, \ldots$, note that the first few k-th order differences are

first-order: $u_n - u_{n-1}$

second-order: $(u_n - u_{n-1}) - (u_{n-1} - u_{n-2}) = u_n - 2u_{n-1} + u_{n-2}$

third-order: $((u_n - u_{n-1}) - (u_{n-1} - u_{n-2})) - ((u_{n-1} - u_{n-2}) - (u_{n-2} - u_{n-3}))$
$= u_n - 3u_{n-1} + 3u_{n+2} - u_{n-3}$

Find a general formula for the k-order differences and prove this formula.

18. As we have seen, the sequence $u_n = n^k$ has constant k-th order differences. Therefore,

$$\sum_{l=0}^{k} \binom{k}{l} (-1)^l u_{n-l} = \sum_{l=0}^{k} \binom{k}{l} (-1)^l (n-l)^k = \text{ constant },$$

i.e., is independent of n.

(a) Conclude from this that one has the curious combinatorial identity: if $r < k$, then

$$\sum_{l=0}^{k} \binom{k}{l} (-1)^l l^r \ = \ 0.$$

(Hint: Show that for each such r the above expression is the coefficient of n^{k-r} in the constant polynomial $\sum_{l=0}^{k} \binom{k}{l} (-1)^l (n-l)^k$.)

(b) Using part (a) show that

$$\sum_{l=0}^{k} \binom{k}{l} (-1)^l l^k \;=\; (-1)^k k!$$

(Hint: this can be shown using induction[27].)

(c) Conclude that if $C(x) = (x-1)^k$, a the solution of $C(\mathbf{u}) = \mathbf{d}$, where $\mathbf{d} = d,\, d,\, \ldots$ is written as

$$u_n = an^k + \text{ lower-degree terms in } n,$$

then $a = \dfrac{d}{k!}$.

19. Let $F_1 = 1$, $F_2 = 1$, $F_3 = 2, \ldots$ be the Fibonacci sequence. Show that one has the curious identity

$$\frac{x}{1 - x - x^2} \;=\; \sum_{k=1}^{\infty} F_k x^k.$$

[27]Here's how:

$$\begin{aligned}
\sum_{l=0}^{k} \binom{k}{l} (-1)^l l^k &= \sum_{l=0}^{k} \left[\binom{k-1}{l} + \binom{k-1}{l-1} \right] (-1)^l l^k \\
&= \sum_{l=0}^{k} \binom{k-1}{l} (-1)^l l^k + \sum_{l=0}^{k} \binom{k-1}{l-1} (-1)^l l^k \\
&= \sum_{l=0}^{k-1} \binom{k-1}{l} (-1)^l l^k - \sum_{l=0}^{k-1} \binom{k-1}{l} (-1)^l (l+1)^k \\
&= -\sum_{l=0}^{k-1} \binom{k-1}{l} (-1)^l \sum_{m=0}^{k-1} \binom{k}{m} l^m \\
&= -\sum_{m=0}^{k-1} \sum_{l=0}^{k-1} \binom{k}{m} \binom{k-1}{l} (-1)^l l^m \\
&= -\sum_{m=0}^{k-1} \binom{k}{m} \sum_{l=0}^{k-1} \binom{k-1}{l} (-1)^l l^m \\
&= -\binom{k}{k-1} \sum_{l=0}^{k-1} \binom{k-1}{l} (-1)^l l^{k-1} \qquad \text{(we've used (a))} \\
&= -k \cdot (-1)^{k-1} (k-1)! = (-1)^k k! \qquad \text{(induction)}
\end{aligned}$$

(Just do the long multiplication showing that $(1 - x - x^2)\left(\sum_{k=1}^{\infty} F_k x^k\right) = x$. This says that the rational function $\dfrac{x}{1 - x - x^2}$ is a **generating function** for the Fibonacci sequence.)

20. A sequence a_1, a_2, a_3, ..., of real numbers is called a **harmonic sequence** if for each $n \geq 1$, a_{n+1} is the harmonic mean of a_n and a_{n+2} (see Exercise 9 of page 42). Show that a given sequence a_1, a_2, ... is a harmonic sequence if and only if all $a_i \neq 0$ and the sequence of reciprocals $\dfrac{1}{a_1}$, $\dfrac{1}{a_2}$, $\dfrac{1}{a_3}$ is an arithmetic sequence.

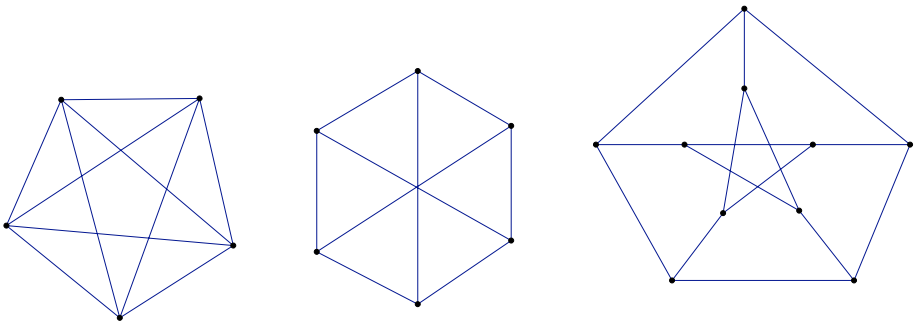
2.2 Elementary Graph Theory

In this section we shall consider one of the most important topics in contemporary discrete mathematics—that of a **graph**. This concept has a huge variety of applications and has become especially important to the relatively new discipline of **management science**.

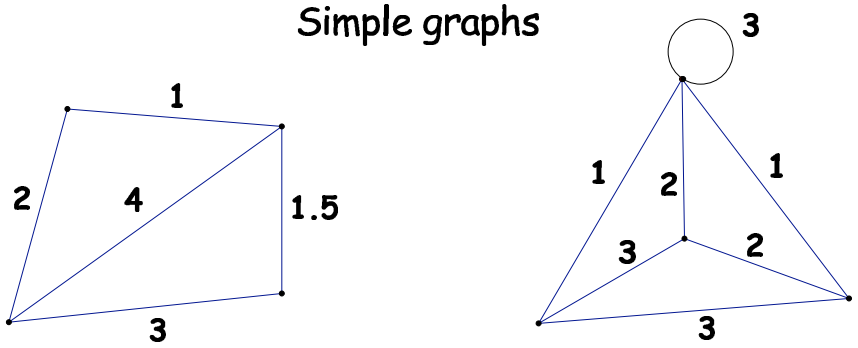
Mathematically, a graph is easy enough to define. It consists of a set V of **vertices** and a numerical relationship between pairs of vertices (sort of a "distance" or "cost" function). Namely, between any two vertices v_i and v_j is a non-negative real number c_{ij} such that it is always true that $c_{ij} = c_{ji}$. If $c_{ij} \neq 0$ we call $\{v_i, v_j\}$ an **edge**. Put intuitively, the cost of getting from vertex v_i to v_j is the same as the cost of getting from vertex v_j to v_i. In other words, the matrix $C = [c_{ij}]$ is a *symmetric matrix*, called the **adjacency matrix**.[28] This matrix is called the **adjacency matrix** of the graph.

If the costs c_{ij} satisfy $c_{ii} = 0$ for all indices i, and c_{ij} is always 0 or 1, then we call the graph a **simple graph**; otherwise we call the graph a **weighted graph**. Perhaps the pictures below will clarify this.

[28] If this matrix isn't symmetric, then the graph is called a **directed graph**; we'll study those briefly in Subsection 2.2.3.



Simple graphs



Non-Simple graphs

Other definitions are as follows. An edge is called a **loop** if it joins a vertex to itself (see the above figure). Let v_i and v_j be vertices in a graph. We say that v_i and v_j are **adjacent** if there is an edge joining v_i and v_j (that is if the cost $c_{ij} > 0$). Also,

A **walk** in a graph is a sequence of linked edges .

A **trail** in a graph is a sequence of linked edges such that no edge appears more than one.

A **path** in a graph is a walk with no repeated vertices.

A **circuit** in a graph is a trail that begins and ends at the same vertex.

A **cycle** in a graph is a path which begins and ends at the same vertex.

If any two vertices of a graph can be joined by a path, then the graph is called **connected**.

2.2.1 Eulerian trails and circuits

Suppose that a postman is charged with delivering mail to residences in a given town. In order to accomplish this in an efficient manner he

would ideally choose a route that would allow him to avoid walking the same street twice. Thus, if the town is represented by a simple graph whose edges represent the streets, then the problem is clearly that of finding a trail in the graph which includes every edge: such a trail is called an **Eulerian trail**. If the postman is to begin and end at the same vertex, then what is sought is an **Eulerian circuit**. General problems such as this are called **routing problems**.

CLASSIC EXAMPLE. In the ancient city of Königsberg (Germany) there were seven bridges, arranged in a "network" as depicted in the figure below:

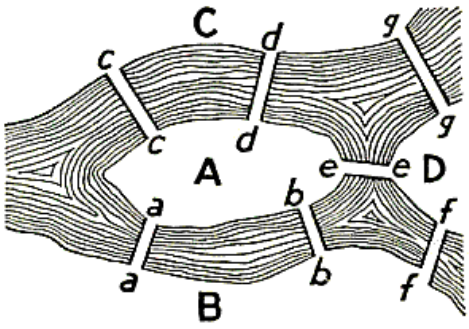
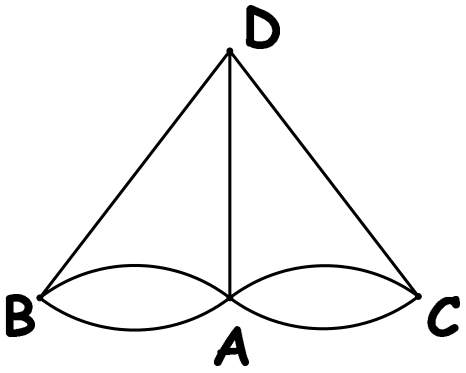


FIGURE 98. *Geographic Map: The Königsberg Bridges.*

A prize was offered to anyone who could determine a route by which each of the bridges can be traversed once and then return to the starting point.

A casual inspection of the above layout of bridges shows that this can be represented by a graph having four vertices and seven edges, as in the graph to the right.



From the above, we see that the adjacency matrix for the seven bridges of Königsberg with labeling $A = 1$, $B = 2$, $C = 3$, and $D = 4$ is given by

$$A = \begin{bmatrix} 0 & 2 & 2 & 1 \\ 2 & 0 & 0 & 1 \\ 2 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

DEFINITION. *The* **degree** *of a vertex v in a graph is the number of edges on this vertex. A loop on a vertex is counted twice in computing the degree of a vertex.*

Notice that if we are given the adjacency matrix, then the sum of the elements of the i-th row is the degree of the vertex i.

THEOREM. *Let G be a finite graph with adjacency matrix A. Then the number of walks of length 2 from vertex v_i to vertex v_j is the (i, j) entry of A^2. More generally, the number of walks of length k from vertex v_i to vertex v_j is the (i, j) entry of A^k.*

A moment's thought is also enough to be convinced of the following theorem:

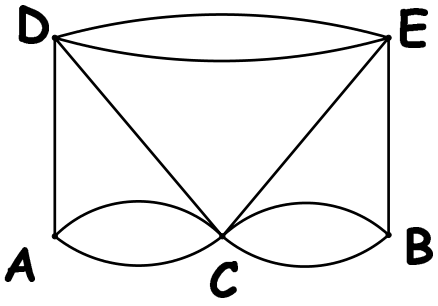
THEOREM. (Euler's Theorem) *Let G be a graph.*

(i) If the graph has any vertices of **odd** *degree, then G cannot contain an Eulerian circuit.*

(ii) If the graph has more than two vertices of **odd** *degree, then G cannot contain an Eulerian trail.*

As a result of Euler's theorem, we see that the bridges of Königsberg problem has no solution!

EXAMPLE 1. The picture to the right depicts a graph G below with exactly two vertices of odd degree, one at vertex A and one at vertex B. The reader should have no difficulty in concluding that G has no Eulerian circuits but does have an Eulerian trail from A to B (or from B to A).



Notice that if we add the degrees of all the vertices in a graph, then every edge get counted twice; this already proves the following.

Theorem. (Euler's Degree Theorem) *The sum of the degrees of the vertices equals twice the number of edges in the graph.*

As a result, one has

Corollary. *The number of vertices of odd degree must be an even number.*

The above results are **negative** in the sense that they tell us when it's impossible to construct Eulerian circuits or Eulerian trails. We shall give an algorithm which allows us to find an Eulerian circuit in a graph all of whose degrees are even.

Fleury's algorithm for finding an Eulerian circuit

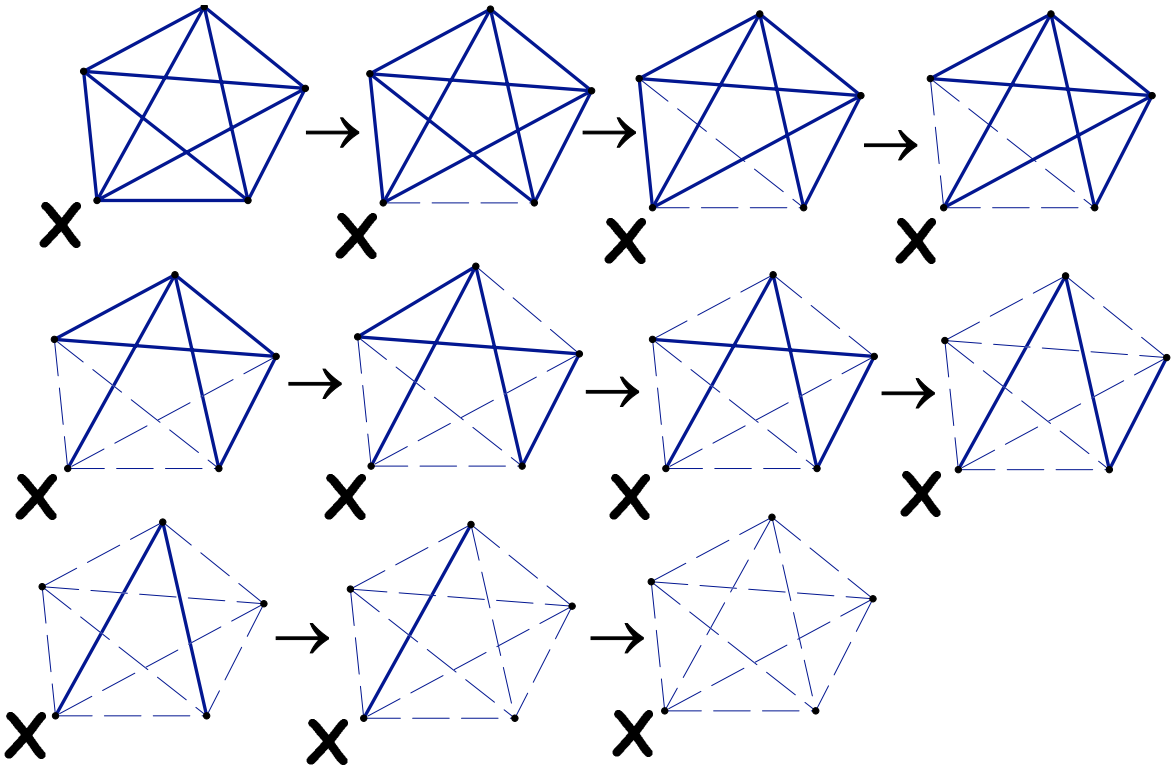
Assume that we are given the graph G all of whose vertex degrees are even. In tracing a trail in G, after having traveled along an edge E, we shall remove this edge (we have "burned our bridges behind us").

Step 1. Pick a vertex X.

Step 2. Move from X to an adjacent vertex Y along the edge E unless removing E disconnectes the graph. (There may be several choices. Also, if there is only one choice, you need to take this choice!)

Step n. Return finally to X.

The above algorithm is depicted in the following sequence. The dotted edges represent the removed edges.



EXERCISES

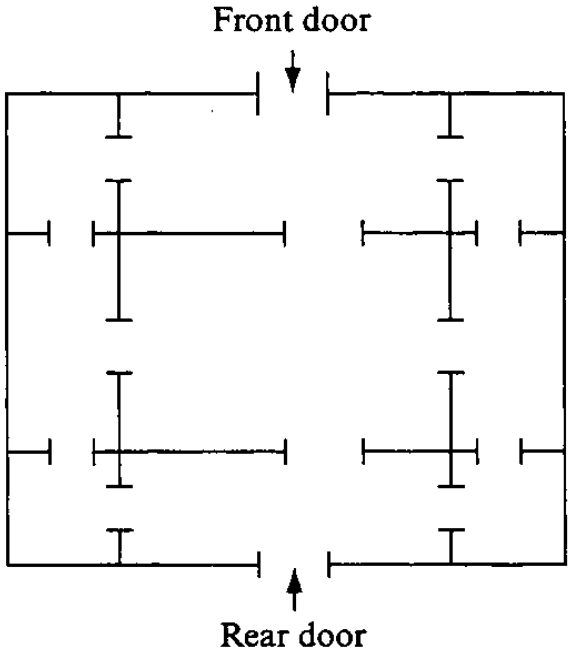
1. Sketch a graph whose adjacency matrix is

$$A = \begin{bmatrix} 0 & 1 & 2 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 2 & 0 \\ 2 & 0 & 2 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 2 & 2 \\ 0 & 2 & 1 & 2 & 0 & 1 \\ 1 & 0 & 1 & 2 & 1 & 2 \end{bmatrix}$$

   How many paths of length 2 are there from vertex v_2 to vertex v_4?

2. The following floor plan shows the ground level of a new home. Is it possible to enter the house through the front door and exit through the rear door, going through each internal doorway exactly once?

Model this problem with a suitable graph and give a reason for your answer.

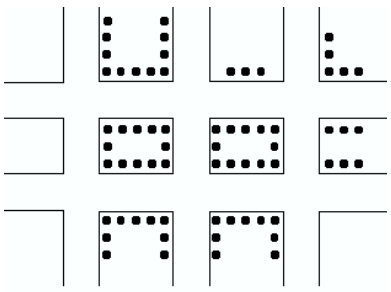


3. Consider the graph G having adjacency matrix

$$A = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 2 & 1 \\ 1 & 0 & 0 & 2 & 1 \\ 1 & 2 & 2 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix}$$

   (a) Draw the graph.
   (b) Explain why G has an Eulerian circuit.
   (c) Find one.

4. The map to the right illustrates a portion of a postal carrier's delivery route. The dots indicate mailboxes into which mail must be delivered. Find a suitable graph to represent the carrier's route. Is there an Eulerian circuit? Is there an Eulerian trail?



5. We consider the following family of simple graphs O_n, $n = 1, 2, \ldots$, defined as follows.

   Vertices: The set of vertices is the set $\{\pm 1, \pm 2, \ldots, \pm n\}$.

   Edges: The vertex i is adjacent to the vertex j precisely when $|i| \neq |j|$.

   (a) Draw the graphs O_1, O_2, O_3.

   (b) What is the degree of every vertex in O_n?

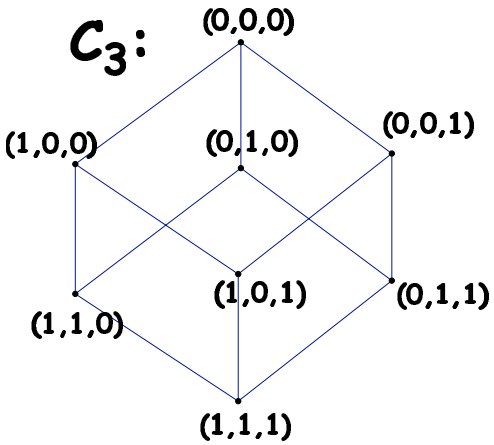
   (c) Is there an Eulerian circuit in O_n, $n > 1$?

6. We consider the family of graphs C_n, $n = 1, 2, \ldots$ defined as follows.

   Vertices: The set of vertices is the set of **binary sequences** $v = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)$, where each $\epsilon_i = 0$ or 1.

   Edges: The vertex v is adjacent to the vertex w precisely when the binary sequences defining v and w differ in exactly one place.

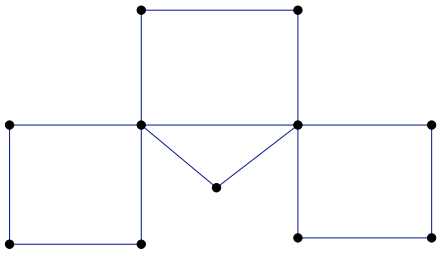
   The graph C_3 is indicated to the right.



   (a) What is the degree of each vertex in C_n, $n \geq 1$?

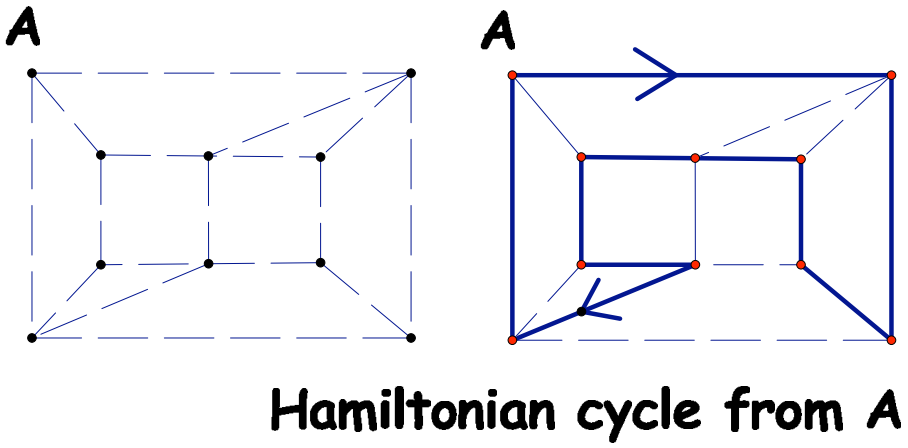
   (b) How many paths are there from the vertex $(0, 0, \ldots, 0)$ to the vertex $(1, 1, \ldots, 1)$?

7. Does the graph to the right have an Eulerian circuit? If so, find it.



2.2.2 Hamiltonian cycles and optimization

In the previous subsection we were largely concerned with the problem of moving around a graph in such a way that each edge is traversed exactly once. The present subsection is concerned with the "dual" problem, namely that of moving around a graph in such a way that each vertex is visited exactly once. Such a walk is called a **Hamiltonian path**. If we return to the original vertex, the walk is called a **Hamiltonian cycle**. The following figure depicts a graph and a Hamiltonian cycle in the graph:

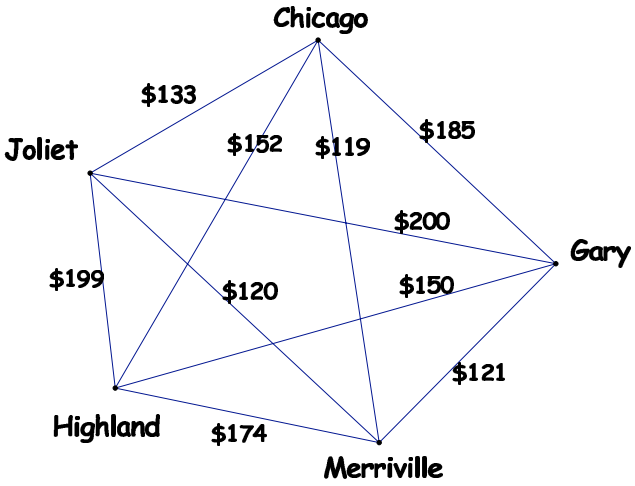


Curiously, unlike the question of the existence of Eulerian circuits, there is no definitive simple criterion for the existence of Hamiltonian cycles. Known results typically involve a lower bound on the degree of each vertex.[29] See the exercises below for a few additional examples.

[29]For example, a 1952 theorem of Paul Dirac says that a graph with n vertices has a Hamiltonian cycle provided that each vertex has degree $\geq n/2$. Øystein Ore generalized this result to graphs (with $n \geq 3$ vertices) such that for each pair of non-adjacent vertices the sum of their degrees is $\geq n$.

Of more significance than just finding a Hamiltonian cycle in a simple graph is that of finding a Hamiltonian cycle of **least total weight** in a weighted graph. Such is the nature of the **traveling salesman problem**. We start with a simple example.

EXAMPLE 1. A salesman needs to visit five cities in the American Midwest: Chicago, Gary, Joliet, Merriville, and Highland. The cost of travel between the cities is depicted in the graph to the right.[30]



We display the costs in tabular form. It will be convenient to use the letters A, B, C, D, and E to represent the cities. Notice that since the matrix of entries is symmetric, there is no need to fill in all of the entries.

	A = Chicago	B = Gary	C = Merriville	D = Highland	E = Joliet
Chicago	*	$185	$119	$152	$133
Gary		*	$121	$150	$200
Merriville			*	$174	$120
Highland				*	$199
Joliet					*

Assuming that the salesman will begin and end his trip in Chicago, what it the **optimal**, i.e., cheapest route for him to take? That is, which Hamilton cycle will afford the least total weight?

In order to answer this question, a few observations are in order. First of all, a **complete graph** is one in which every pair of distinct vertices are joined by an edge. Thus, the above graph is a (weighted) complete graph. Next, it is obvious that in a complete graph with n vertices, there are exactly $(n-1)!$ Hamiltonian cycles starting from a given vertex. In the present example there are $4! = 24$ Hamiltonian

[30]The numbers are taken from Example 2, page 201 of Excursions in Modern Mathematics, Fourth Edition, by Peter Tannenbaum and Robert Arnold.

cycles.

In order to find the Hamiltonian cycle of minimal weight, we shall resort to the **Brute-Force Method**, that is we shall form a complete list of the Hamiltonian cycles and their weights, choosing the one of minimal weight as the solution of our problem. There is one final simplification, namely, if the complete graph with vertices $\{v_1, v_2, \ldots, v_n\}$ is weighted, then the weight of the Hamiltonian cycle $(v_1, v_2, \ldots, v_n, v_1)$ clearly has the same weight as the "reverse cycle" $(v_1, v_n, v_{n-1}, \ldots, v_2, v_1)$. Therefore the Brute Force Method will require us to compare the weights of $\frac{1}{2}(n-1)!$ Hamiltonian cycles.

We now list the weights of the Hamiltonian cycles in the above graph, highlighting the cycle of minimal weight.

cycle	**weight**	**reverse cycle**
ABCDEA	$185 + 121 + 174 + 199 + 133 = \$812$	AEDCBA
ABCEDA	$185 + 121 + 120 + 199 + 152 = \$777$	ADECBA
ABDCEA	$185 + 150 + 174 + 120 + 133 = \$762$	AECDBA
ABDECA	$185 + 150 + 199 + 120 + 119 = \$773$	ACEDBA
ABECDA	$185 + 200 + 120 + 174 + 152 = \$831$	ADCEBA
ABEDCA	$185 + 200 + 199 + 174 + 119 = \$877$	ACDEBA
ACBDEA	$119 + 121 + 150 + 199 + 133 = \$722$	AEDBCA
ACBEDA	$119 + 121 + 200 + 199 + 152 = \$791$	ADEBCA
ADBCEA	$152 + 150 + 121 + 120 + 133 = \$676$	**AECBDA**
ADBECA	$152 + 150 + 200 + 120 + 119 = \$741$	ACEBDA
AEBCDA	$133 + 200 + 121 + 174 + 152 = \$780$	ADCBEA
AEBDCA	$133 + 200 + 150 + 174 + 119 = \$776$	ACDBEA

As a result of the above computations we see that the minimal cost is for the salesman to visit the cities in the order

Chicago $\longrightarrow$ Highland $\longrightarrow$ Gary $\longrightarrow$ Merriville $\longrightarrow$ Joliet $\longrightarrow$ Chicago,

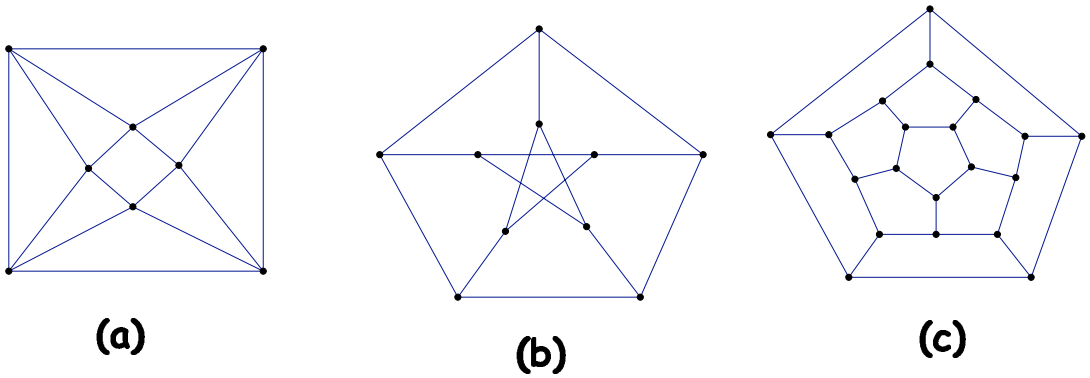
which results in a total cost of \$676. In the next subsection we shall consider a few algorithms which can be used to determine "good" Hamiltonian cycles if not the optimal Hamiltonian cycle.

The above is an example of the **Traveling Salesman Problem**—

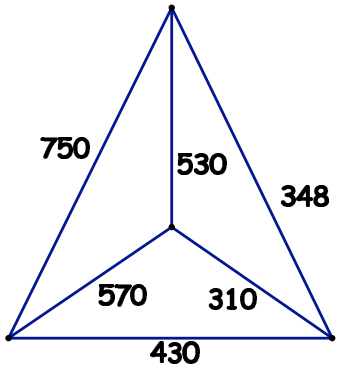
often abbreviated TSP—and is one of fundamental importance in Management Science. It is also related to the so-called **P = NP** problem (one of the Millennium problems)[31] in that a general good (i.e., efficient) solution of TSP would in fact prove that P = NP.

EXERCISES

1. Two of the three graphs below have a Hamiltonian cycle. Determine which two and in each case find a Hamiltonian cycle.



2. Find a Hamiltonian cycle of minimal weight in the graph to the right.



3. Let G be a complete graph having six vertices. Suppose that we label each edge with either a 0 or a 1. Prove that in this graph there must exist either

   (a) three vertices among whose edges are all labeled "0," or

   (b) three vertices among whose edges are all labeled "1."[32]

[31] See www.claymath.org/millinnium.

[32] This is an elementary example of "Ramsey Theory." In general, the **Ramsey number** of a complete graph with n vertices is the maximum number k such an arbitrary labeling of the edges (with 0s and1s) of the graph will result in a subgraph with k vertices having all the edge labels 0 or

TSP: The nearest-neighbor algorithm

As indicated above, the brute-force method will always find the optimal solution, but the amount of computing time required may be astronomical (which is hardly optimal!). In this and the following sections we shall consider two very simple algorithms which don't necessarily find the optimal solution but they are known to quickly find "good" solutions.

The **Nearest-Neighbor** algorithm starts with a vertex in the weighted graph, and then proceeds to move to the "nearest neighbor" without prematurely returning to a previous vertex.

Example. In attempting to construct the cheapest route starting from and returning to Chicago, we proceed as follows

1. Move from Chicago to Merriville; the cost of $119 is the cheapest among all costs involving travel from Chicago.
2. Move from Merriville to Joliet $120; this is the cheapest cost (other than $119, which puts us back in Chicago).
3. Move from Joliet to Highland at a cost of $199.
4. Move from Highland to Gary at a cost of $150.
5. Return to Chicago at a cost of $185.

The total cost of the above Hamiltonian route is $773, which while not optimal was an easy route to obtain.

Exercises

1. Consider the weighted graph with vertices A, B, C, D, and E, having weights assigned as follows

all the edge labels 1. The Ramsey number of the complete graph with six vertices is 3. In fact, one way the above problem is often described is as follows:

> Show that among six people there must be either three mutual friends or three mutual strangers.

	A	B	C	D	E
A	*	20	23	19	17
B		*	20	25	18
C			*	24	19
D				*	23
E					*

Use the Nearest-Neighbor algorithm to find a Hamiltonian cycle starting at vertex A. What is the total weight of this Hamiltonian cycle?

2. Use the Nearest-Neighbor algorithm to find a Hamiltonian cycle starting at vertex A. What is the resulting total weight of this cycle?

	A	B	C	D	E	F
A	*	4.7	5.1	3.6	1.1	0.8
B		*	0.6	8.2	5.7	5.2
C			*	8.1	5.9	5.6
D				*	3.2	3.1
E					*	1.5
F						*

3. There is a variation of the Nearest-Neighbor Algorithm which increases the computation time by a factor of the number of vertices of the weighted graph. This might seem stiff, but this added time pales by comparison with the time required to carry out the Brute-Force method. Namely, for each vertex of the weighted graph compute the Hamiltonian cycle constructed by the Nearest-Neighbor Algorithm, and then take the Hamiltonian cycle of least total weight. This is called the **Repetitive Nearest-Neighbor** algorithm. Do this for the above weighted graph consisting of travel among the given five Midwestern cities.

TSP: The cheapest-link algorithm

There is a alternative algorithm—the **Cheapest-Link** algorithm which efficiently computes a relatively cheap Hamiltonian cycle in a weighted graph. This is easy to describe, as follows.

In the weighted graph start by choosing the edge of minimal weight (the "cheapest link"). Next choose the next cheapest link, and so on.

As with the Nearest-Neighbor algorithm, we do not select any edges which would prematurely result in a cycle. Also, we need to avoid any edges which will result in more than two edges from a given vertex.

Example. We consider this algorithm on the Midwestern Cities graph.

1. Choose the {Chicago, Merriville} link as this is the cheapest among all links.
2. Choose the {Merriville, Joliet} link; this is the second cheapest at $120.
3. The third cheapest link is the {Gary, Merriville} link at $121; however, choosing this link will result in three edges issuing from Merriville. The fourth cheapest link is the {Chicago, Joliet} link at $133. However, this is also impossible as a premature cycle is formed. We settle for the {Gary, Highland} link at $150.
4. We choose the {Chicago, Highland} link at $152.
5. The only remaining choice given the constraints is the {Gary, Joliet} link at $200.

The above algorithm produces the Hamiltonian cycle

Chicago $\longrightarrow$ Merriville $\longrightarrow$ Joliet $\longrightarrow$ Gary $\longrightarrow$ Highland $\longrightarrow$ Chicago,

at a total (non-optimal) cost of $741.

The algorithm above are what are called **greedy algorithms** as at each stage they seek the optimal (i.e., cheapest) choice.

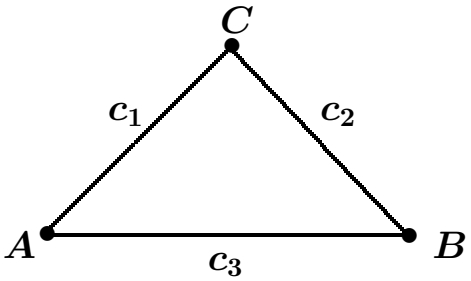
Exercises

1. Apply the Cheapest-Link algorithm to the graph indicated in the table in Exercise 1 on page 121.

2. Apply the Cheapest-Link algorithm to the graph indicated in the table in Exercise 2 on page 122.

2.2.3 Networks and spanning trees

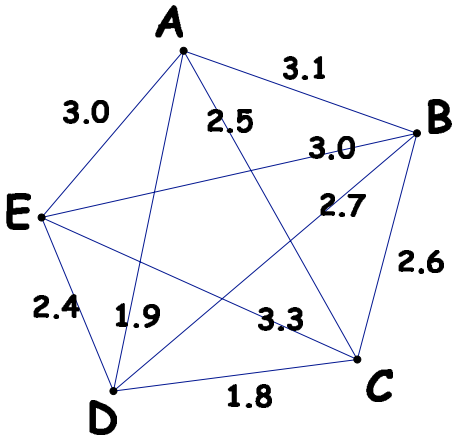
In this subsection we consider a problem similar to TSP but different in the sense that efficient **and** optimal solutions are possible. The basic idea is this: suppose, for example, that we have the weighted graph



We need for these three points to be "networked," i.e., in communication with each other, but without any redundancy. In other words, we don't need all three of the edges in the above graph because if A is networked with B, and B is networked with C then A is networked with C: there is a "transitivity" of networking. Therefore, the above idealized networking problem would be optimized by discarding the redundant (and most expensive) edge so that the sum of the remaining edge weights becomes a minimum.

Let us flesh out the above with a somewhat more detailed problem.

EXAMPLE 1. Assume that we have cities A, B, C, D, and E and that they can be networked according to the costs depicted in the weighted graph to the right.



What we are looking for is a network which will result in the the cities being interconnected but without any redundancy. Also, we are looking to do this with the least possible cost. The first condition simply states that we are looking for a "subgraph" of the above graph containing all of the vertices but without having any cycles in it. Such is called a

spanning tree of the graph.[33] The second says that we are looking for a **minimal-weight spanning tree**.

Before continuing with the above example, a few comments are in order. First of all, a given graph is called a **tree** if it is connected, has no multiple edges, and contains no cycles. Therefore, in particular, a tree is a simple graph. We shall prove a couple of results about trees.

LEMMA. *Let G be a tree, and let E be an edge in G. Then the removal of E results in a disconnected graph.*

PROOF. Let E be on the vertices v and w. If the removal of E doesn't disconnect G then there is a path from v to w without using the edge E. Since we can get from v to w via E, we clearly have a cycle in the graph G. Therefore, the removal of E must result in disconnecting G.

THEOREM. *Let G be a finite simple connected graph containing n vertices. Then G is a tree if and only if G has $n-1$ edges.*

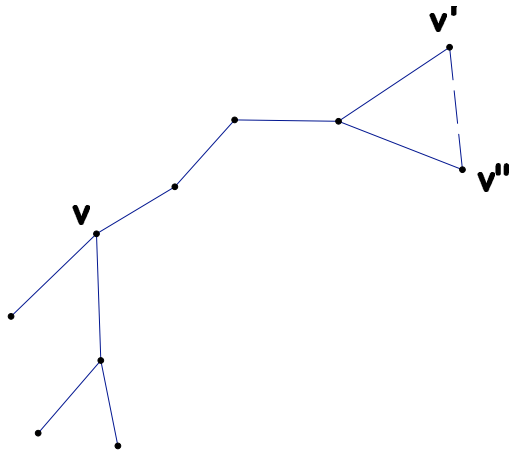
PROOF. Assume that G is a finite tree and fix a vertex v in G. For any vertex w in G denote by $d(v,w)$ (the **distance** from v to w) the length of the shortest path from v to w. Since G only has finitely many vertices, there must exist a vertex v' of **maximal** distance from v.

CLAIM: v' has only one edge on it, i.e., v' is an **end** in the tree G. Assume that $d(v,v') = d$ and let

$$v = v_0,\, v_1,\, v_2,\, \ldots,\, v_d = v'$$

[33]It is easy to see that any connected finite graph contains a spanning tree. Indeed, suppose that the tree T is a subgroup of the connected graph G having a maximal number of vertices. If these aren't all of the vertices of G, then by the connectivity of G one of the vertices of the tree must be adjacent to a new vertex in G. Adding this vertex (and the corresponding edge) creates a larger tree inside G, a contradiction. (Even if the graph has an infinite number of vertices, there still must exist a spanning tree. The proof, however, uses what's called **Zorn's Lemma** and is outside the scope of these notes.)

be a path from v to v', where each $\{v_{i-1}, v_i\}$ is an edge in G. Assume that v' is adjacent to another vertex v''. If a minimal length path from v to v'' must travel through v', then v'' must be of greater distance from v than is v'. This can't happen and so there must be a path from v to v'' which doesn't pass through v'. But with $\{v', v''\}$ being an edge, then we see that it is possible to construct a cycle through v', which is a contradiction.



We now may remove the vertex v' and the unique edge e on v from the graph G; what results is clearly a tree with $n-1$ vertices. Using induction, we may conclude that this tree must have $n-1-1 = n-2$ edges. If we replace the removed vertex and edge, we arrive at the conclusion that G itself must have $n-1$ edges.

Conversely, assume that G is a connected graph with n vertices and $n-1$ edges.

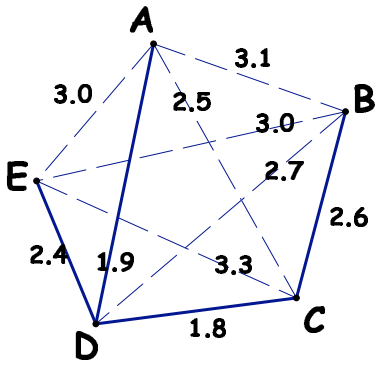
CLAIM: The graph G must contain an end vertex v. If not then each vertex of G must sit on at least two edges, and so

$$\#\text{ edges in } G \;=\; \tfrac{1}{2} \sum_{\substack{\text{vertices } v \\ \text{in } G}} (\#\text{ edges on } v) \;\geq n,$$

which is a contradiction. Therefore, G must contain an end vertex v.

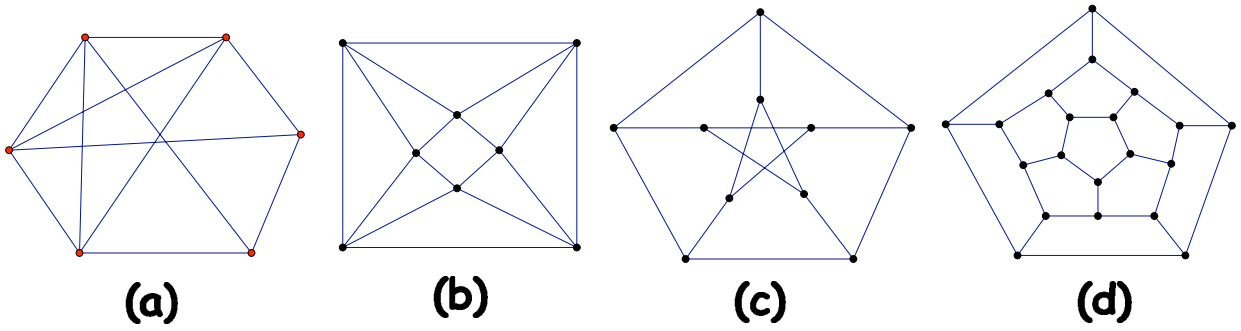
We now remove the end vertex v and the single edge containing v from the graph G. This results in a connected graph G' consisting of $n-1$ vertices and $n-2$ edges. Again using mathematical induction we conclude that G' must, in fact, be a tree. But then adding v and the single edge to G will certainly produce a tree, and we're done.

EXAMPLE 1, CONTINUED. We shall return to the above example only long enough to indicate a minimal-weight spanning tree. In the next subsection we shall indicate an efficient method to derive this optimal solution.



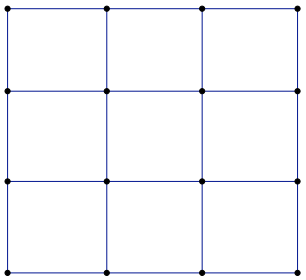
EXERCISES

1. Construct (by drawing) a spanning tree in each of the graphs depicted below.



2. Can you give a simple example of a graph which has no Hamiltonian cycle?

3. Indicate a Hamiltonian cycle in the graph to the right (if one exists).

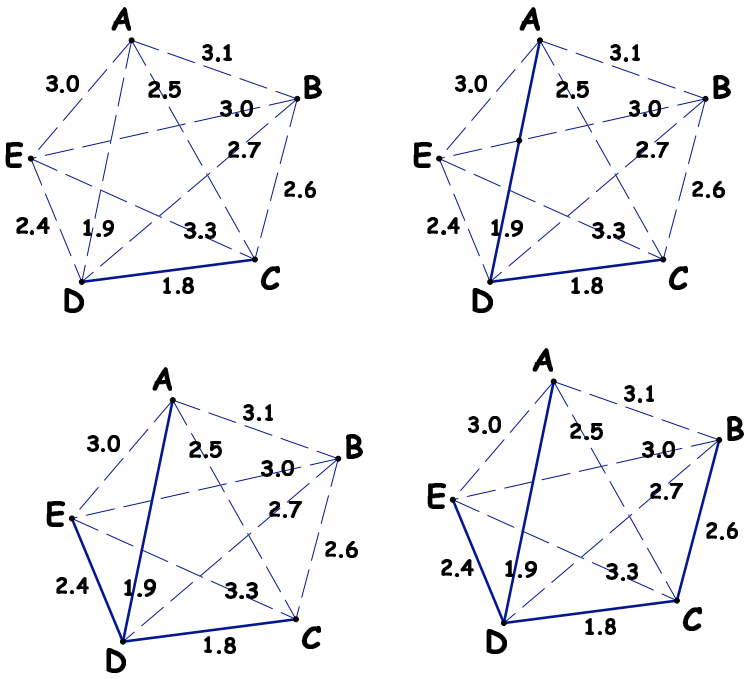


Kruskal's algorithm

Kruskal's algorithm is the same in spirit as the **Cheapest-Link algorithm** for finding minimal-weight Hamiltonian cycles. However, the surprising difference is that whereas the Cheapest Link algorithm doesn't always find the minimal-weight Hamiltonian cycle, Kruskal's algorithm will always find the minimal-weight spanning tree in a weighted graph.

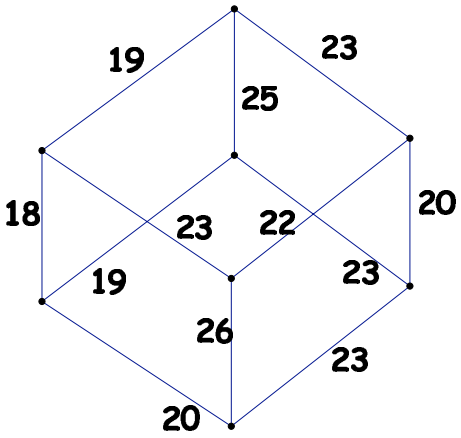
The algorithm is implemented by selecting in turn the edges of minimal weight—and hence is a greedy algorithm—disregarding any choice that creates a circuit in the graph. The algorithm ends when a spanning tree is obtained.

We indicate in steps how the minimal-weight spanning tree for the example on page 127 was obtained (notice that we couldn't choose the edge with weight 2.5, as this would create a cycle:



Exercises

1. Find a minimal spanning tree for the graph on the right.



2. The table to the right gives a describes a graph. An asterisk (*) indicates an edge of infinite weight. Use Kruskal's algorithm to find a minimal-weight spanning tree.

	A	B	C	D	E	F	G
A	*	5	8	7	*	*	*
B		*	5	*	4	5	*
C			*	2	2	*	3
D				*	*	*	2
E					*	3	1
F						*	3
G							*

3. (**Efficient upper and lower bounds for Hamiltonian cycles of minimal weight)** In this exercise we show how to obtain reasonable upper and lower bounds for the minimal weight of a Hamiltonian cycle in a weighted graph G.

   (a) **(Lower bound)** Notice that if we remove an edge from a Hamiltonian cycle we get a spanning tree. Therefore do this:

      i. Delete a vertex v and all the edges incident with v from the graph, call the resulting graph G_v.

      ii. Use Kruskal's algorithm to find a minimal spanning tree for G_v. Let the total weight of this tree be $\mathbf{W_v}$.

      iii. Replace the vertex v and two of the cheapest edges on v.

   Show that $\mathbf{W_v}+\mathbf{W} \leq$ total weight of a minimal-weight Hamiltonian cycle, where $\mathbf{W}$ denotes the sum of the weights of the two edges found in (iii), above. Thus we have efficiently ob-

tained a lower bound for the total weight of a minimal-weight Hamiltonian cycle.

(b) **(Upper bound)** Use one of the efficient methods above (Nearest-neighbor or cheapest-link algorithm) to find a Hamiltonian cycle. The weight is then an upper bound.

Prim's algorithm

Like Kruskal's algorithm, **Prim's algorithm** is an efficient method for finding a minimal-weight spanning tree in a weighted graph. We describe this as follows. Assume that the given weighted graph is G. For convenience, we shall initially regard all of the vertices and edges in G as colored black.

STEP 1. Pick an initial vertex v_1. Color this vertex red.

STEP 2. Find a vertex v_2 of minimal distance (weight) to v_1. Color the vertex v_2 and the edge $\{v_1, v_2\}$ red.

STEP 3. Choose a new vertex v_3 of minimal distance to either v_1 or v_2. Color the new vertex v_3 and the corresponding minimal-length edge red.

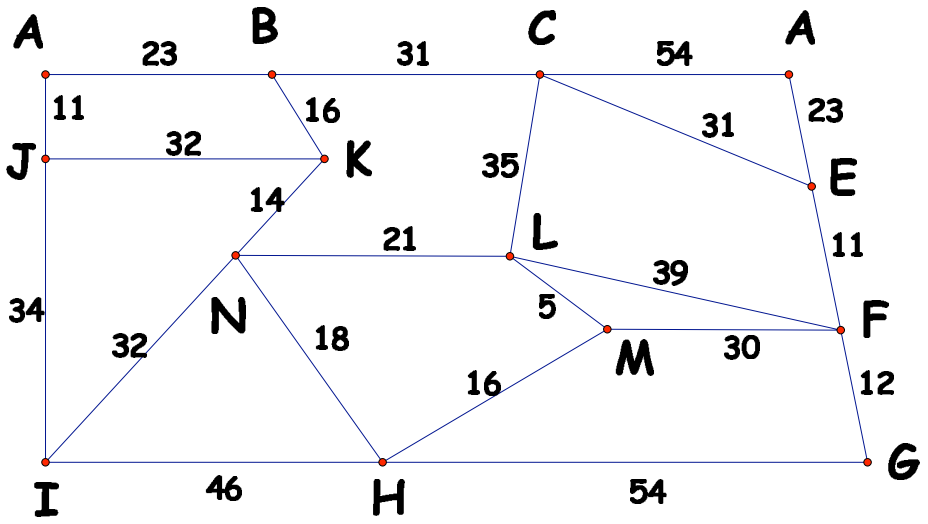
STEP n. Repeated application of the above will determine a red subtree of G with vertices $v_1, v_2, \ldots, v_{n-1}$. Find a black edge of minimal weight on one of the above $n-1$ vertices. Color this edge and the new vertex v_n which it determines red.

CONCLUSION. Continue until all vertices in G have been colored red; the resulting red graph is a minimal-weight spanning tree.

EXERCISES

1. Use Prim's algorithm to find minimal spanning trees in the first two exercises on page 129.

2. Use Prim's algorithm to find a minimal spanning tree in the graph below:



3. Use the methods of Exercise 3 on page 129 to find upper and lower bounds on the weight of a Hamiltonian cycle in the above graph.

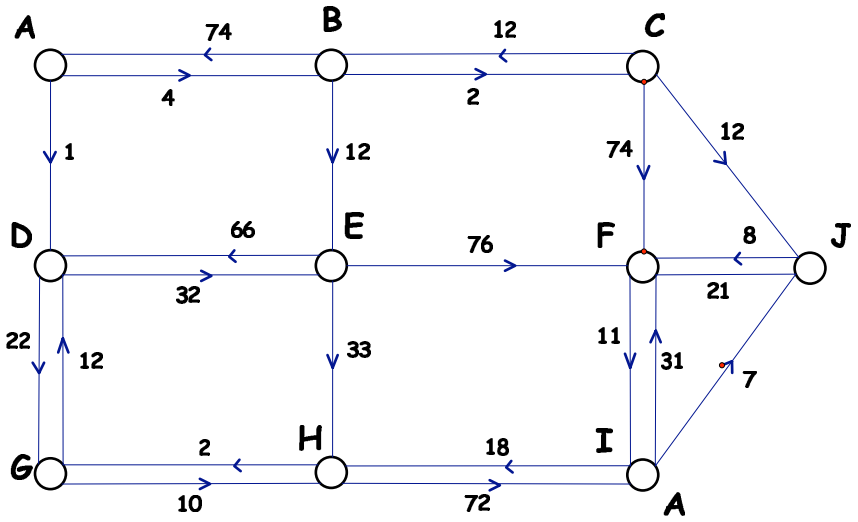
Weighted directed graphs; Dijkstra's algorithm

In many applications of graph theory, one notices that the cost of moving from vertex v_1 to the vertex v_2 might be different from the cost of moving from v_2 to the vertex v_1.[34] Such a graph is called a **weighted directed graph**. Of interest in this setting is to find the minimal weight (cost) in getting from an initial vertex—call it v_0—to some other vertex v.[35]

Below is depicted a weighted directed graph:

[34]For example the price of a airline ticket from Shanghai to Beijing is typically (slightly) less than the price of a ticket from Beijing to Shanghai.

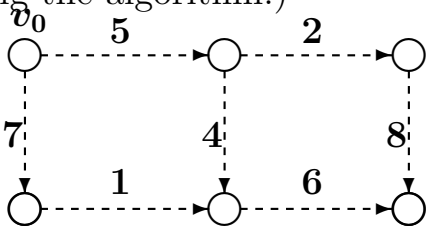
[35]This is sometimes called the **minimal connector problem**.



Of course, a weighted graph can be thought of as being directed where the weights are the same in both directions.

Dijkstra's algorithm[36] constructs in a graph G a **directed tree** starting from the vertex v_0 such that the minimal-weight path from v_0 to any other vertex v can be found by moving from v_0 to v along this tree. The description of the algorithm proceeds as follows. We shall assume, for convenience that all directed edges are initially drawn as "dotted directed edges." Also, each vertex shall initially carry a temporary label, to be replaced eventually with a permanent label (which will represent the minimal distance from the initial vertex v_0. (Caution: the temporary labels can change during the algorithm!)

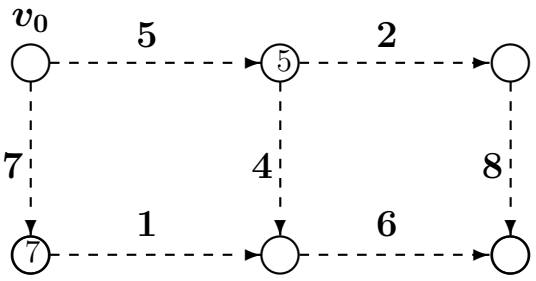
We'll use the graph to the right to illustrate Dijkstra's algorithm.



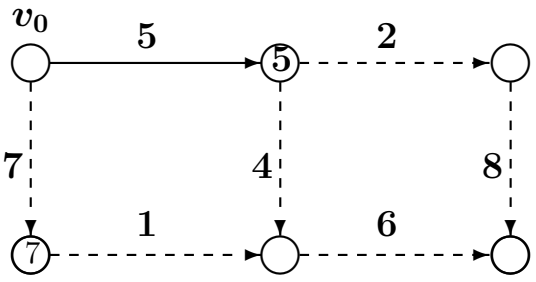
We now itemize the steps in Dijkstra's algorithm.

[36]There are a couple of really nice applets demonstrating Dijkstra's algorithm:
http://www.dgp.toronto.edu/people/JamesStewart/270/9798s/Laffra/DijkstraApplet.html
http://www-b2.is.tokushima-u.ac.jp/ ikeda/suuri/dijkstra/Dijkstra.shtml

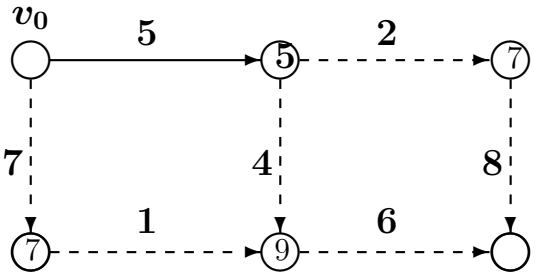
STEP 1. Find the vertices v in G such that (v_0, v) is a directed edge. Temporarily mark these vertices with their weighted distrances from v_0.



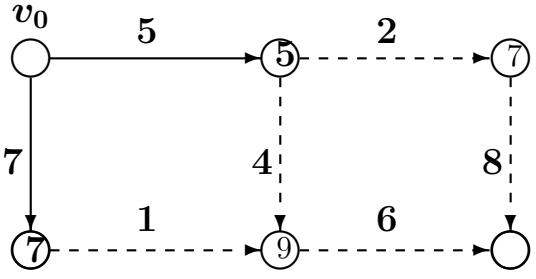
STEP 2. Fill in the edge connecting v_0 to the vertex v of minimal distance from v_0; the temporary label at v_1 is now a permanent label.



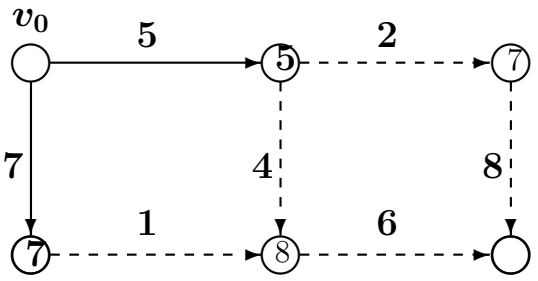
STEP 3. Find all new vertices connected to v_1; temporarily mark these vertices with their distances from v_0 through v_1.



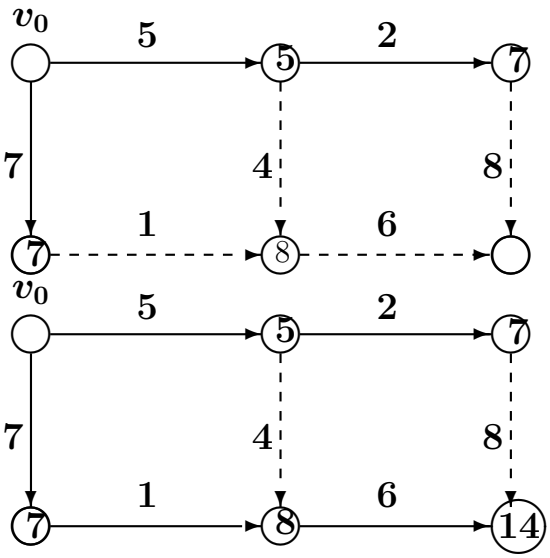
STEP 4. Select a vertex v_2 having a minimal weight label; color in the directed edge and make the label permanent. (Note that in the event that there is more than one vertex of minimal distance, the choice is arbitrary.)



STEP 5. Find all new vertices connected to v_2; mark these with their distances from v_0 (along solid directed edges) and through v_2. If such a vertex already has a temporary label, overwrite this label if the distance through v_2 is less than the existing label. (This is where a label can change! If there are no new vertices, go to the next step.)



STEP 6 AND BEYOND. Choose the vertex having the minimal temporary label. Color in the directed edge and make the label permanent. Keep repeating this process until all vertices have permanent labels; The darkened directed edges determine a directed tree through which minimal weight paths are determined.



EXERCISE.

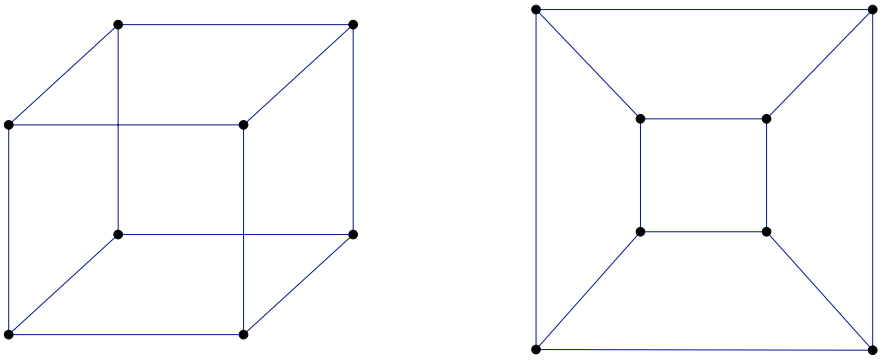
1. Use Dijkstra's algorithm to find a minimal-weight path from vertex A to vertex J in the graph on page 131.

2.2.4 Planar graphs

Two graphs G_1 and G_2 are **isomorphic** if there is a function

$$f: \text{ vertices of } G_1 \longrightarrow \text{ vertices of } G_2$$

such that $\{f(v_1), f(w_1)\}$ is an edge of G_2 exactly when $\{v_1, w_1\}$ is an edge of G_1. In other words, two graphs are isomorphic exactly when one is simply a redrawing of the other. A moment's thought reveals that the two graphs depicted below are isomorphic.

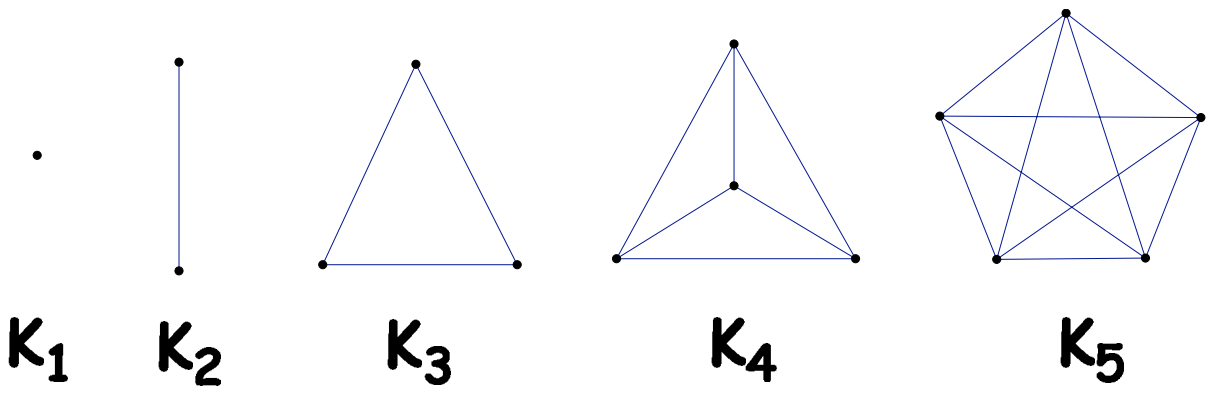


Assume that G_1 and G_2 are graphs and that

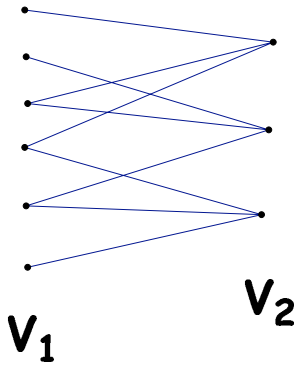
$$f: \text{ vertices of } G_1 \longrightarrow \text{ vertices of } G_2$$

determines an isomorphism between these graphs. If v_1 is a vertex of G_1, and if $v_2 = f(v_1)$, it should be instantly clear that v_1 and v_2 have the same degree. However, this condition isn't sufficient; see Exercise 1 on page 141.

There are two important families of graphs that warrant special consideration. The first is the family of **complete graphs** K_1, K_2, K_3, ... (see also page 118). The graph K_n is the simple graph (see page 109) having n vertices and such that every vertex is adjacent to every other vertex.

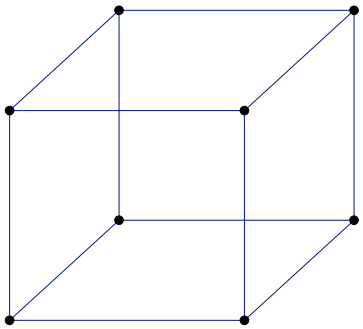


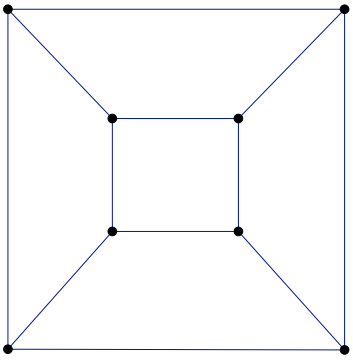
The next important family involves the so-called **bipartite graphs**. The simple graph G is called **bipartite** if its set V of vertices can be partitioned into two disjoint subsets $V = V_1 \cup V_2$ where there are no edges among the vertices of V_1 and there are no edges among the vertices of V_2.



The **complete bipartite graph** $K_{m,n}$, where m and n are positive integers, is the bipartite graph with vertices $V = V_1 \cup V_2$, $|V_1| = m$ and $|V_2| = n$ and where every vertex of V_1 is adjacent with every vertex of V_2 (and vice versa).

We turn now to the main topic of this section, that of **planar graphs**.[37] These are the graphs which can be "faithfully" drawn in the plane. By "faithful" we mean that the edges drawn between vertices will have no crossings in the plane. As a simple example, we consider below two versions of the graph of the cube: the first is how we usually imagine it in three-dimensional space, and the second is how we could draw it in the plane.

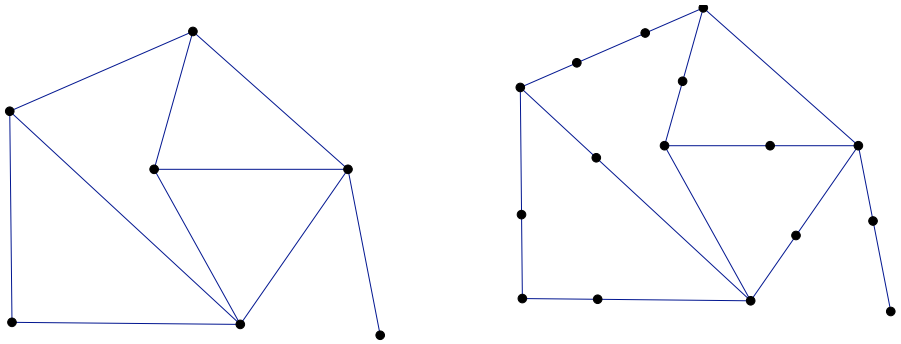




EXAMPLE 1. The complete graphs K_1, K_2, K_3, K_4 are obviously planar graphs. However, we shall see below that K_5 is not planar; in fact, none of the complete graphs K_n, $n \geq 5$ is planar. Also, the complete bipartite graph $K_{3,3}$ is also not planar (try it!). (We'll prove below that $K_{3,3}$ is not planar.)

[37] The topic of Planar graphs falls into the general category of "topological graph theory."

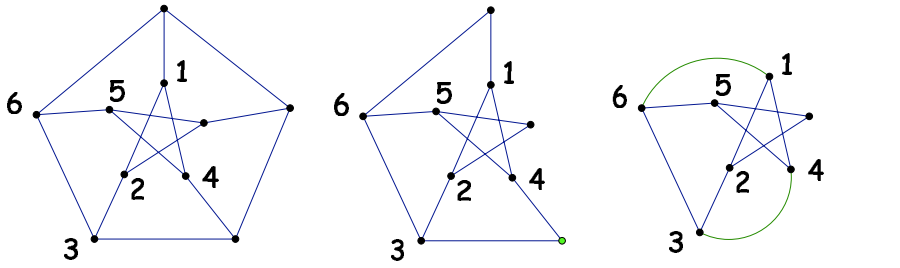
There are two fundamental theorems which give criteria for a graph to be planar. They're relatively deep results, so we won't give proofs here. The first result makes use of the notion of "homeomorphism" of graphs. Namely, two graphs are **homeomorphic** if one can be obtained from the other simply by adding vertices along existing edges. However, no new edges can be added!

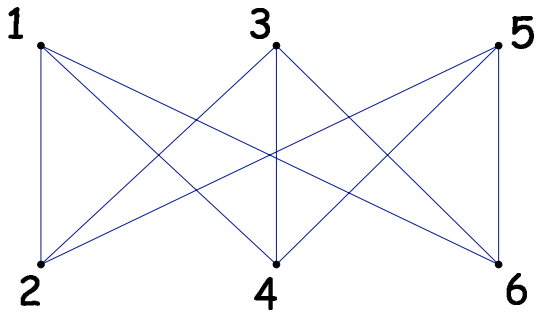


Homeomorphic graphs

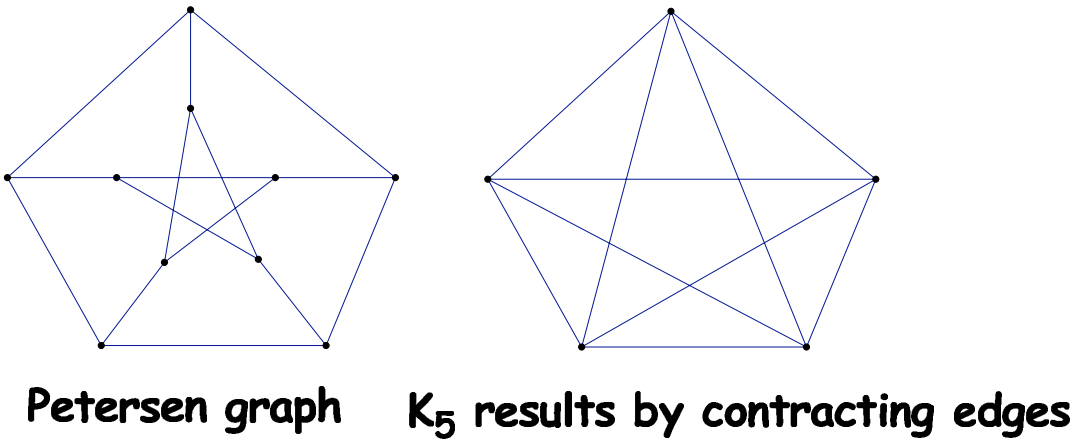
Theorem. (Kuratowski's Theorem) *A finite graph G is planar if and only if G has no subgraph homeomorphic to the complete graph K_5 on five vertices or the complete bipartite graph $K_{3,3}$.*

From Kuratowski's theorem we can deduce that the Petersen graph is not planar. Indeed, the sequence below shows that the Petersen graph has a subgraph which is homeomorphic with the complete bipartite graph $K_{3,3}$.





The next planarity condition is somewhat more useful but slightly more technical. First of all, a graph H is called a **minor** of the graph G if H is isomorphic to a graph that can be obtained by a number of edge contractions on a subgraph of G. Look at the so-called **Petersen graph**; it contains K_5 as a minor:



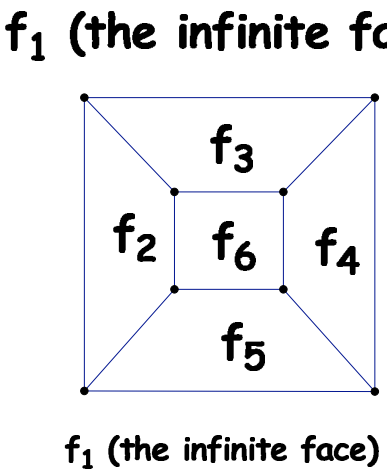
THEOREM. (Wagner's Theorem) *A finite graph G is planar if and only if it does not have K_5 or $K_{3,3}$ as a minor.*

As a result, we see that the Petersen graph is not planar.

Euler's formula and consequences

Once a planar graph has been drawn in the plane, it not only determines *vertices* and *edges*, it also determines **faces**. These are the 2-dimensional regions (exactly one of which is unbounded) which are bounded by the edges of the graph. The plane, together with a graph faithfully drawn in it is called a **planar map**. Thus, a planar map has the vertices and edges of the "embedded" graph G, it also has faces.

EXAMPLE 2. We look at the cube graph drawn in the plane. Notice that there are 6 naturally defined regions, or **faces**.

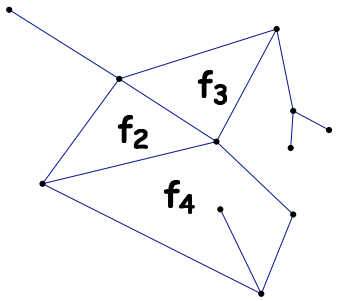


f1 (the infinite face)

EXAMPLE 3. Here is a more irregular planar graph with the faces indicated. Also, we have computed

$$\#\text{vertices} - \#\text{edges} + \#\text{faces} = 2;$$

this is a fundamental result.



v-e+f=11-13+4=2

If we compute $\#\text{vertices} - \#\text{edges} + \#\text{faces}$ for the planar map in Example 2 above, we also get 2. There must be something going on here! We start by defining the **Euler characteristic** of the planar map M by setting

$$\boldsymbol{\chi}(M) \;=\; \#\text{vertices} - \#\text{edges} + \#\text{faces}.$$

The surprising fact is that the number always produced by the above

is 2:

THEOREM. (Euler's Formula) *If M be a connected planar map, then* $\boldsymbol{\chi}(M) = 2$.

PROOF. Let T be a maximal spanning tree inside G; the existence of T was proved in the footnote on page 125. Note that since T has no cycles, there can only be one face: $f = 1$. Next, we know by the theorem on page 125 that $v = e + 1$. Therefore, we know already that $\chi(T) = v - e + f = 1 + 1 = 2$. Next, we start adding the edges of G to the tree T, noting that each additional edge divides an existing face in two. Therefore the expression $v - e + f$ doesn't change as e and f have both increased by 1, proving the result.[38]

COROLLARY. *For the simple planar map M, we have* $\mathbf{e} \leq 3\mathbf{v} - 6$.

PROOF. We may assume that M has at least three edges, for otherwise the underlying graph is a tree, where the result is easy. This easily implies that each face—including the infinite face—will be bounded by at least three edges. Next, notice that an edge will bound either a single face or two faces. If the edge e bounds a single face, then the largest connected subgraph containing e and whose edges also bound a single face is—after a moment's thought—seen to be a tree. Removing all edges of this tree and all vertices sitting on edges bounding a single face will result in removing the same number of vertices as edges. On the map M' which remains every edge bounds exactly two faces. Also, the number $\mathbf{f}$ of faces of M' is the same as the number of faces of the original map M. Let $\mathbf{v}'$, $\mathbf{e}'$ be the number of vertices and edges, respectively, of M'. Since every face of M' is bounded by at least three edges, and since every edge bounds exactly two faces of M' we infer that $3\mathbf{f} \leq 2\mathbf{e}'$. Therefore,

$$2 \;=\; \mathbf{v}' - \mathbf{e}' + \mathbf{f} \;\leq \mathbf{v}' - \mathbf{e}' + 2\mathbf{e}'/3 = \mathbf{v}' - \mathbf{e}'/3,$$

[38] In the most general setting, the **Euler characteristic** of a graph is a function of where it's faithfully drawn. For example, it turns out that the Euler characteristic of a graph faithfully drawn on the surface of a doughnut (a "torus") is always 0. See also the footnote on page 196.

From which it follows that $\mathbf{e}' \leq 3\mathbf{v}' - 6$. However, $\mathbf{e}' = \mathbf{e} - k$ and $\mathbf{v}' = \mathbf{v} - k$ for some fixed non-negative integer k from which we infer that $\mathbf{e} \leq 3\mathbf{v} - 6$.

Example 4. From the above result, we see immediately that the complete graph K_5 is not planar as it has $\binom{5}{2} = 10$ edges which is greater than $3\mathbf{v} - 6 = 9$.

If we have a planar bipartite graph, then the above result can be strengthened:

Corollary. *Let M be a simple planar map with no triangles. Then we have $\mathbf{e} \leq 2\mathbf{v} - 4$.*

Proof. As in the above proof, that each edge bounds two faces and that each face—including the infinite face—will be bounded by at least four edges (there are no triangles). This implies that $4\mathbf{f} \leq 2\mathbf{e}$. Therefore,

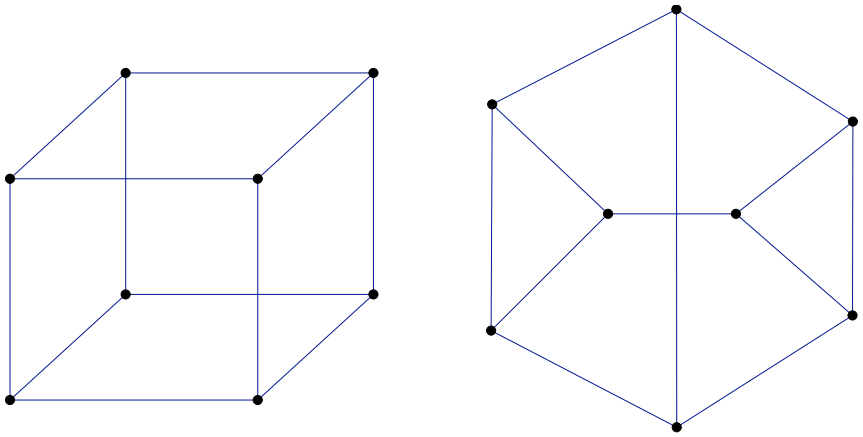
$$2 \;=\; \mathbf{v} - \mathbf{e} + \mathbf{f} \;\leq \mathbf{v} - \mathbf{e} + \mathbf{e}/2 = v - \mathbf{e}/2,$$

and so $\mathbf{e} \leq 2\mathbf{v} - 4$ in this case.

Example 5. From the above result, we see immediately that the complete bipartite graph $K_{3,3}$ is not planar. Being bipartite, it cannot have any triangles (see Exercise 5), furthermore, it has 9 edges which is greater than $2\mathbf{v} - 4 = 8$.

Exercises

1. Show that even though the degree of each vertex in both graphs below is 3, these graphs are **not** isomorphic.



2. Here's a slightly more sophisticated problem. Define the graphs G_1 and G_2, as follows. Start by letting n be a fixed positive integer.

   Vertices of G_1: These are the subsets of $\{1, 2, \ldots, n\}$.

   Edges of G_1: $\{A, B\}$ is an edge of G_1 exactly when

   $$|A \cap B| \;=\; \max\{|A| - 1, |B| - 1\}.$$

   (Notice that this says that either $A \subseteq B$ and $|B| = |A| + 1$ or that $B \subseteq A$ and that $|A| = |B| + 1$.)

   Vertices of G_2: These are the **binary sequences** $v = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)$, where each $\epsilon_i = 0$ or 1.

   Edges of G_2: $\{v, w\}$ is an edge of G_2 precisely when the binary sequences defining v and w differ in exactly one place. (This is the graph defined in Exercise 6 on page 116.)

   Show that the graphs G_1 and G_2 are isomorphic.

3. Assume that a graph G can be "faithfully" drawn on the surface of a sphere. Must this graph be planar?

4. Consider the "grid graph," constructed as follows. Let m and n be positive integers and in the coordinate plane mark the points having *integer coordinates* (k, l) such that $0 \leq k \leq m$ and $0 \leq n \leq m$. These are the vertices of the graph G. The edges in this graph connect the vertices separated by Euclidean distance 1. Show that this graph is bipartite.

5. Prove that any cycle in a bipartite graph must have even length. Conversely, if every cycle in a graph has even length, show that the graph must be bipartite.

6. How many edges are there in the complete bipartite graph $K_{m,n}$?

7. Let G be a finite simple graph (see page 109) of n vertices in which every vertex has degree k. Find a simple formula in terms for the number of edges in terms of n and k.

8. Let G be a planar graph. Prove that G must contain a vertex whose degree is at most 5.

9. Use the result of Exercise 8 to show that any planar graph can be 6-colored. That is to say, if G is a planar graph then using only six colors we can color the vertices of G in such a way that no two adjacent vertices have the same color.[39]

10. Prove that none of the complete graphs K_n, $n \geq 5$ is planar.

11. Let G be a planar graph and let M be the map it determines by an embedding in the plane. We define the **dual graph** G^* (relative to the map M) as follows. The vertices of G^* are the faces of M. Next, for each edge of G we draw an edge between the two faces bounded by this edge. (If this edge bounds a single face, then a loop is created.) Show (by drawing a picture) that even when every edge bounds two faces, then the dual graph might not be a simple graph even when G is a simple graph.

12. Let G be a planar graph, embedded in the plane, resulting in the map M. Let G^* be the dual graph relative to M. Let T be a spanning tree in G and consider the subgraph T^* of G^* to have all the vertices of G^* (i.e., all the faces of M) and to have those edges which corresponding to edges in G **but not in** T.

[39]Of course, the above result isn't "best possible." It was shown in 1976 by K. Apple and W. Haken that any planar map can 4-colored. For a nice online account, together with a sketch of a new proof (1997) by N. Robertson, D.P. Sanders, P.D. Seymour, and R. Thomas, see http://www.math.gatech.edu/~thomas/FC/fourcolor.html. Both of the above-mentioned proofs are computer aided.

It is not too difficult to prove that a planar graph can be 5-colored; see M. Aigner and G.M. Ziegler, *Proofs from the Book*, Third Edition, Springer, 2004, pages 200-201.

(a) Show that T^* is a spanning tree in G^*.

(b) Conclude that $v = e_T + 1$ and $f = e_{T^*} + 1$, where e_T is the number of edges in T and e_{T^*} is the number of edges in T^*.

(c) Conclude that $e_T + e_{T^*} = e$ (the number of edges in G).

(d) Conclude that $v + f = (e_T + 1) + (e_{T^*} + 1) = e + 2$, thereby giving another proof of Euler's theorem.

Chapter 3

Inequalities and Constrained Extrema

3.1 A Representative Example

The thrust of this chapter can probably be summarized through the following very simple example. Starting with the very simple observation that for real numbers x and y, $0 \leq (x-y)^2$. Expanding the right hand side and rearranging gives the following inequality:

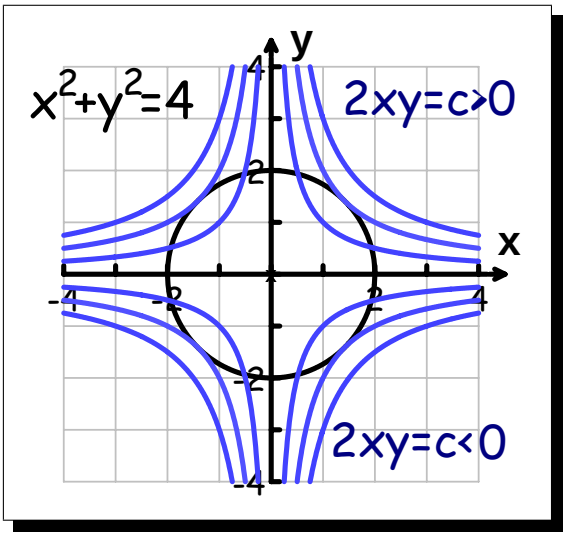
$$2xy \ \leq \ x^2 + y^2,$$

again valid for all x, $y \in \mathbb{R}$. Furthermore, it is clear that equality obtains precisely when $x = y$. We often refer to the as an **unconditional** inequality, to be contrasted from inequalities which are true only for certain values of the variable(s). This is of course, analogous to the distinction between "equations" and "identities" which students often encounter.[1]

We can recast the above problem as follows.

[1]By way of reminder, the **equality** $x^2 - x - 6 = 0$ admits a solution, viz., $x = -2,\ 3$, whereas the equality $x(x-2) = x^2 - 2$ is always true (by the distributive law), and hence is an **identity**.

PROBLEM. Given that $x^2 + y^2 = 4$, find the maximum value of $2xy$.



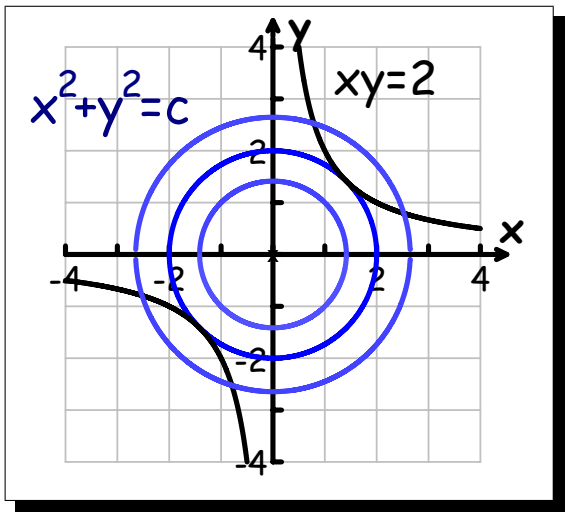
SOLUTION. If we are thinking in terms of the above-mentioned inequality $2xy \leq x^2 + y^2$, with equality if and only if $x = y$, then we see immediately that the maximum value of $2xy$ must be $x^2 + y^2 = 4$. However, it is instructive to understand this problem in the context of the graph to the right, where the "constraint curve" is the graph of $x^2 + y^2 = 4$ and we're trying to find the largest value of the constant c for which the graph $2xy = c$ meets the constraint curve.

From the above figure, it is clear that where $2xy$ obtains its maximum value will occur at a point where the graph is tangent to the circle with equation $x^2 + y^2 = 4$. As a result, this suggest that the solution can also be obtained using the methodology of differential calculus (indeed, it can!), but in this chapter we wish to stress purely algebraic techniques.

We can vary the problem slightly and ask to find the maximum value of xy given the same constraint $x^2 + y^2 = 4$. However, the maximum of xy is clearly $1/2$ the maximum of $2xy$ and so the maximum value of xy is 2.

In an entirely similar fashion we see that the minimum value of $2xy$ given $x^2 + y^2 = 4$ must be -2. This can be seen from the above figure. Even more elementary would be to apply the inequality $0 \leq (x+y)^2 \Rightarrow -2xy \leq x^2 + y^2$.

As a final variation on the above theme, note can that can interchange the roles of constraint and "objective function" and ask for the extreme values of $x^2 + y^2$ given the constraint $xy = 2$. The relevant figure is given to the right. Notice that there is no maximum of x^2+y^2, but that the minimum value is clearly $x^2 + y^2 = 4$, again occurring at the points of tangency.



EXERCISES.

1. Find the maximum of the function xy given the elliptical constraint $4x^2 + y^2 = 6$. Draw the constraint graph and the "level curves" whose equations are xy=constant.
2. Given that $xy = -5$, find the maximum value of the objective function $x^2 + 3y^2$.
3. Given that $xy = 10$, find the maximum value of the objective function $x + y$.
4. Suppose that x and y are positive numbers with $x + y = 1$. Compute the minimum value of $\left(1 + \frac{1}{x}\right)\left(1 + \frac{1}{y}\right)$.

3.2 Classical Unconditional Inequalities

Until further notice, we shall assume that the quantities x_1, x_2, ..., x_n are all positive. Define

ARITHMETIC MEAN:

$$\mathrm{AM}(x_1, x_2, \ldots, x_n) = \frac{x_1 + x_2 + \cdots + x_n}{n};$$

GEOMETRIC MEAN:

$$\mathrm{GM}(x_1, x_2, \ldots, x_n) = \sqrt[n]{x_1 x_2 \cdots x_n};$$

HARMONIC MEAN:

$$\mathrm{HM}(x_1, x_2, \ldots, x_n) = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}};$$

QUADRATIC MEAN:[2]

$$\mathrm{QM}(x_1, x_2, \ldots, x_n) = \sqrt{\frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n}}.$$

Note that if $a_1, a_2, \ldots$ is an arithmetic sequence, then a_n is the arithmetic mean of a_{n-1} and a_{n+1}. Likewise if $a_1, a_2, \ldots$ is a geometric sequence (and all $a_n > 0$), then a_n is the geometric mean of a_{n-1} and a_{n+1}.

A **harmonic sequence** is by definition the reciprocal of nonzero terms in an arithmetic sequence. Thus, the sequences

$$1, \ \frac{1}{2}, \ \frac{1}{3}, \ \ldots, \quad \text{and} \quad \frac{2}{3}, \ \frac{2}{5}, \ \frac{2}{7}, \ \ldots$$

are harmonic sequences. In general, if $a_1, a_2, \ldots$ is a harmonic sequence, then a_n is the harmonic mean of a_{n-1} and a_{n+1}.

One of our aims in this section is to prove the classical inequalities

$$\mathrm{HM} \leq \mathrm{GM} \leq \mathrm{AM} \leq \mathrm{QM}.$$

Before doing this in general (which will require mathematical induction), it's instructive first to verify the above in case $n = 2$.

Indeed, starting with $0 \leq (\sqrt{x} - \sqrt{y})^2$ we expand and simplify the result as

$$2\sqrt{xy} \leq x + y \Rightarrow \mathrm{GM}(x_1, x_2) \leq \mathrm{AM}(x_1, x_2).$$

Having proved this, note next that

$$\mathrm{HM}(x_1, x_2) = \left(\mathrm{AM}\left(\frac{1}{x_1} + \frac{1}{x_2}\right)\right)^{-1};$$

[2]Sometimes called the **root mean square**

since we have already shown that $\mathrm{GM}(x,y) \leq \mathrm{AM}(x,y)$ for $x,\, y \geq 0$, we now have

$$\mathrm{HM}(x_1,x_2) = \left(\mathrm{AM}\left(\frac{1}{x_1}+\frac{1}{x_2}\right)\right)^{-1} \leq \left(\mathrm{GM}\left(\frac{1}{x_1},\frac{1}{x_2}\right)\right)^{-1} = \mathrm{GM}(x_1,x_2).$$

Finally, note that since $2x_1x_2 \leq x_1^2+x_2^2$ (as proved in the above section),

$$(x_1+x_2)^2 = x_1^2 + 2(x_1x_2) + x_2^2 \leq 2(x_1^2+x_2^2).$$

Divide both sides of the above inequality by 4, take square roots and infer that $\mathrm{AM}(x_1,x_2) \leq \mathrm{QM}(x_1,x_2)$.

For a geometric argument showing HM $\leq$ GM $\leq$ AM, see Exercise 1, below.

We turn next to proofs of the above inequalities in the general case.

$\mathrm{AM}(x_1,\ldots,x_n) \leq \mathrm{QM}(x_1,\ldots,x_n)$: This is equivalent with saying that

$$\frac{(x_1+\cdots+x_n)^2}{n^2} \leq \frac{x_1^2+\cdots+x_n^2}{n},$$

which is equivalent with proving that

$$(x_1+\cdots+x_n)^2 \leq n(x_1^2+\cdots+x_n^2).$$

By induction, we may assume that

$$(x_1+\cdots+x_{n-1})^2 \leq (n-1)(x_1^2+\cdots+x_{n-1}^2).$$

Furthermore, note that for any real numbers x, y, we have $0 \leq (x-y)^2 = x^2+y^2-2xy \Rightarrow 2xy \leq x^2+y^2$. Armed with this, we proceed, as follows:

$$\begin{aligned}(x_1+\cdots+x_n)^2 &= (x_1+\cdots+x_{n-1})^2 + 2x_n(x_1+\cdots+x_{n-1}) + x_n^2\\ &\leq (n-1)(x_1^2+\cdots+x_{n-1}^2)\\ &\quad +(x_1^2+x_n^2)+\cdots+(x_{n-1}^2+x_n^2)+x_n^2\\ &= n(x_1^2+\cdots+x_n^2),\end{aligned}$$

which proves that AM $\leq$ QM. Notice that since, for **any** $x_1,\, x_2,\, \ldots,\, x_n$,

$$x_1+x_2+\cdots+x_n \leq |x_1|+|x_2|+\cdots+|x_n|,$$

then we see that $\mathrm{AM}(x_1, \ldots, x_n) \leq \mathrm{QM}(x_1, \ldots, x_n)$ is true without the assumption that all x_i are positive.

$\mathrm{GM}(x_1, \ldots, x_n) \leq \mathrm{AM}(x_1, \ldots, x_n)$: Let $C = \sqrt[n]{x_1 x_2 \cdots x_n}$. If all $x_i = C$, then

$$\sqrt[n]{x_1 x_2 \cdots x_n} = C = \frac{x_1 + \cdots + x_n}{n},$$

and we're done in this case. Therefore, we may assume that at least one of the x_is is less than C and that one is greater than C. Without loss of generality, assume that $x_1 > C$ and that $C > x_2$. Therefore, $(x_1 - C)(C - x_2) > 0$ and so $x_1 x_2 < C(x_1 + x_2) - C^2 \Rightarrow \dfrac{x_1 + x_2}{C} > \left(\dfrac{x_1}{C}\right)\left(\dfrac{x_2}{C}\right) + 1$. From this, we conclude

$$\begin{aligned} \frac{x_1 + x_2 + \cdots + x_n}{C} &> \frac{(x_1 x_2)/C + x_3 + \cdots + x_n}{C} + 1 \\ &\geq (n-1)\sqrt[n-1]{(x_1 x_2 \cdots x_n)/C^n} + 1 \quad \text{(using induction)} \\ &= (n-1) + 1 = n. \end{aligned}$$

That is to say, in this case we have

$$\frac{x_1 + x_2 + \cdots + x_n}{n} > C = \sqrt[n]{x_1 x_2 \cdots x_n},$$

concluding the proof that GM $\leq$ AM. (For a much easier proof, see Exercise 2 on page 160.)

$\mathrm{HM}(x_1, \ldots, x_n) \leq \mathrm{GM}(x_1, \ldots, x_n)$: From the above we get

$$\frac{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_3}}{n} \geq \sqrt[n]{\frac{1}{x_1}\frac{1}{x_2}\cdots\frac{1}{x_n}}\ ;$$

take reciprocals of both sides and infer that HM $\leq$ GM.

A generalization of AM $\leq$ QM is embodied in the very classical **Cauchy-Schwarz inequality**. We state this as a theorem.

THEOREM 1. (Cauchy-Schwarz Inequality) *Given* $x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n \in \mathbb{R}$, *one has*

$$(x_1 y_1 + x_2 y_2 + \cdots + x_n y_n)^2 \leq (x_1^2 + x_2^2 + \cdots + x_n^2)(y_1^2 + y_2^2 + \cdots + y_n^2).$$

PROOF. We define a quadratic function of x:

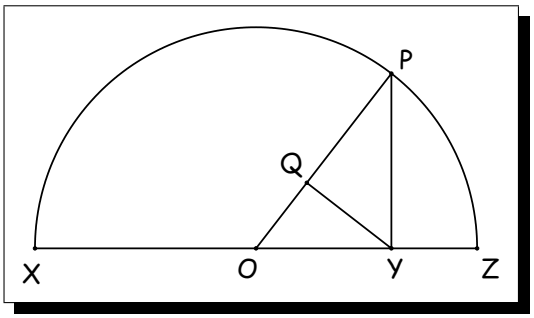
$$\begin{aligned} Q(x) &= (xx_1 - y_1)^2 + \cdots (xx_n - y_n)^2 \\ &= (x_1^2 + \cdots + x_n^2)x^2 - 2(x_1y_1 + \cdots x_ny_n)x + (y_1^2 + \cdots + y_n^2). \end{aligned}$$

Since $Q(x) \geq 0$, we see that the discriminant must be ≤ 0:

$$4(x_1y_1 + \cdots + x_ny_n)^2 - 4(x_1^2 + \cdots + x_n^2)(y_1^2 + \cdots + y_n^2) \leq 0,$$

so we're done! Pretty slick, eh?[3] See Exercise 4, below.

EXERCISES.



1. The diagram to the right depicts a semicircle with center O and diameter XZ. If we write $XZ = a + b$, identify $\text{AM}(a, b)$, $\text{GM}(a, b)$, and $\text{HM}(a, b)$ as lengths of segments in the diagram.

2. Suppose that a_1, a_2, a_3, ... is a sequence of non-negative terms such that for each index $i > 1$, $a_i = \text{QM}(a_{i-1}, a_{i+1})$. Show that a_1^2, a_2^2, a_3^2, ... is an arithmetic sequence.

3. Show that if a, $b > 0$ then

$$\frac{a+b}{2} \leq \frac{2}{3}\left(\frac{a^2 + ab + b^2}{a+b}\right).$$

4. Show how AM $\leq$ QM is a direct consequence of the Cauchy-Schwarz inequality.

5. State a necessary and sufficient condition for $\text{AM}(x_1, \ldots, x_n) = \text{QM}(x_1, \ldots, x_n)$.

[3]The Cauchy-Schwarz inequality can be generalized to complex numbers where it reads:

$$|x_1y_1 + x_2y_2 + \cdots + x_ny_n|^2 \leq (|x_1|^2 + |x_2|^2 + \cdots + |x_n|^2)(|y_1|^2 + |y_2|^2 + \cdots + |y_n|^2).$$

6. Find the maximum value of the objective function $x+y+z$ given that $x^2+y^2+z^2=4$. (Hint: use $\mathrm{AM}(x,y,z) \leq \mathrm{QM}(x,y,z)$.) Can you describe this situation geometrically?

7. Find the maximum value of the objective function $x^2+y^2+z^2$ given that $x+y+z=6$.

8. Suppose that x and y are positive numbers with $x+y=1$. Show that $\dfrac{1}{x}+\dfrac{1}{y} \geq 4$.

9. Suppose that x and y are positive numbers with $x+y=1$. Compute the minimum value of $\left(1+\dfrac{1}{x}\right)\left(1+\dfrac{1}{y}\right)$. (This was already given as Exercise 4 on page 147. However, doesn't it really belong in this section? Can you relate it to Exercise 8, above?)

10. Assume that $x_1, x_2, \ldots, x_n > 0$ and that $x_1+\cdots+x_n=1$. Prove that
$$\frac{1}{x_1}+\cdots+\frac{1}{x_n} \geq n^2.$$
(Hint: don't use mathematical induction!)

11. Let $n \geq 2,\ x, y > 0$. Show that
$$2\sum_{k=1}^{n-1} x^k y^{n-k} \leq (n-1)(x^n+y^n).$$
(This is somewhat involved; try arguing along the following lines.

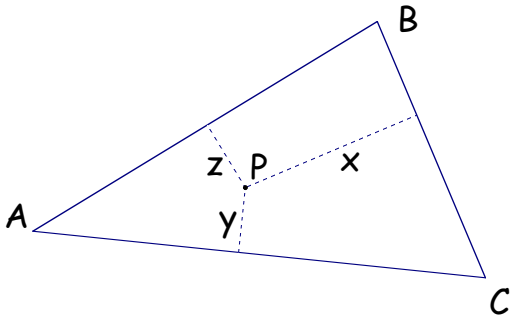
(i) Let $P(x,y) = (n-1)(x^n+y^n) - 2\sum_{k=1}^{n-1} x^k y^{n-k}$; note that $P(y,y)=0$ (i.e., $x=y$ is a zero of $P(x,y)$ regarded as a polynomial in x).

(ii) Show that $\dfrac{d}{dx}P(x,y)\Big|_{x=y} = 0$. Why does this show that $P(x,y)$ has at least a double zero at $x=y$?

(iii) Use Descartes Rule of Signs to argue that $P(x,y)$ has, for $x, y > 0$ **only** a double zero at $x=y$.

(iv) Show that this implies that $P(x, y) \geq 0$ when x, $y > 0$ with equality if and only if $x = y$.)

12. You are given $\triangle ABC$ and an interior point P with distances x to $[BC]$, y to $[AC]$ and z to $[AB]$ as indicated. Let $a = BC$, $b = AC$, and $c = AB$.



(a) Find the point P which minimizes the objective function

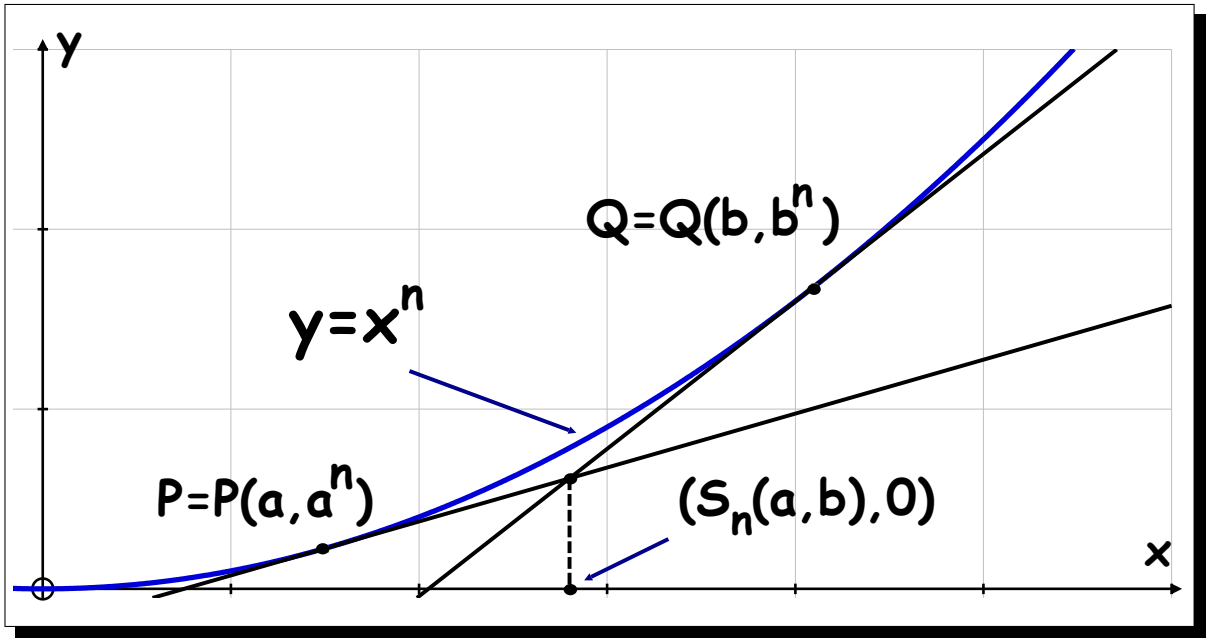
$$F = \frac{a}{x} + \frac{b}{y} + \frac{c}{z}.$$

(Hint: note that $ax+by+cz$ is proportional to the area of $\triangle ABC$. If need be, refer back to Exercise 5 on page 17.[4])

(b) Conclude from part (a) that the inradius r of $\triangle ABC$ (see page 17) is given by $r = 2A/P$, where A and P are the area and perimeter, respectively, of $\triangle ABC$.

The next few exercises will introduce a geometrical notion of the mean of two positive numbers. To do this, fix a positive number $n \neq 1$ (which need not be an integer), and draw the graph of $y = x^n$ for non-negative x. For positive real numbers $a \neq b$, locate the points $P = P(a, a^n)$ and $Q = Q(b, b^n)$ on the graph. Draw the tangents to the graph at these points; the x-coordinate of the point of intersection of these tangents shall be denoted $S_n(a, b)$ and can be regarded as a type of mean of a and b. (If $a = b$, set $S_n(a, b) = a$.) See the figure below:

[4]It turns out that P must be the incenter of $\triangle ABC$.



12. Show that if a, $b > 0$, then

    (a) $S_{-1}(a,b) = \mathrm{HM}(a,b)$;

    (b) $S_{1/2}(a,b) = \mathrm{GM}(a,b)$;

    (c) $S_2(a,b) = \mathrm{AM}(a,b)$.

13. Show that

$$S_n(a,b) = \frac{(n-1)(a^n - b^n)}{n(a^{n-1} - b^{n-1})}, \ (a \neq b).$$

14. Show that if $2 \leq m \leq n$ are integers, and if a, $b > 0$ are real numbers, then $S_m(a,b) \leq S_n(a,b)$. (Hint: this can be carried out in a way similar to the solution of Exercise 11.

    (a) First note that

$$\frac{(m-1)(a^m - b^m)}{m(a^{m-1} - b^{m-1})} \leq \frac{(n-1)(a^n - b^n)}{n(a^{n-1} - b^{n-1})}$$

    if an only if

$$n(m-1)(a^m - b^m)(a^{n-1} - b^{n-1}) \leq m(n-1)(a^n - b^n)(a^{m-1} - b^{m-1}).$$

    (b) Next, define the polynomial

$$P(a,b) = m(n-1)(a^n - b^n)(a^{m-1} - b^{m-1}) - n(m-1)(a^m - b^m)(a^{n-1} - b^{n-1});$$

the objective is to show that $P(a,b) \geq 0$ when a, b, m, n are as given above. Regard the above as a polynomial in a and use Descartes Rule of Signs to conclude that (counting multiplicities) $P(a,b)$ has at most four positive real zeros.

(c) Note that $a = b$ is a zero of $P(a,b)$, i.e., that $P(b,b) = 0$. Next, show that

$$\frac{d}{da}P(a,b)\Big|_{a=b} = \frac{d^2}{da^2}P(a,b)\Big|_{a=b} = \frac{d^3}{da^3}P(a,b)\Big|_{a=b} = 0.$$

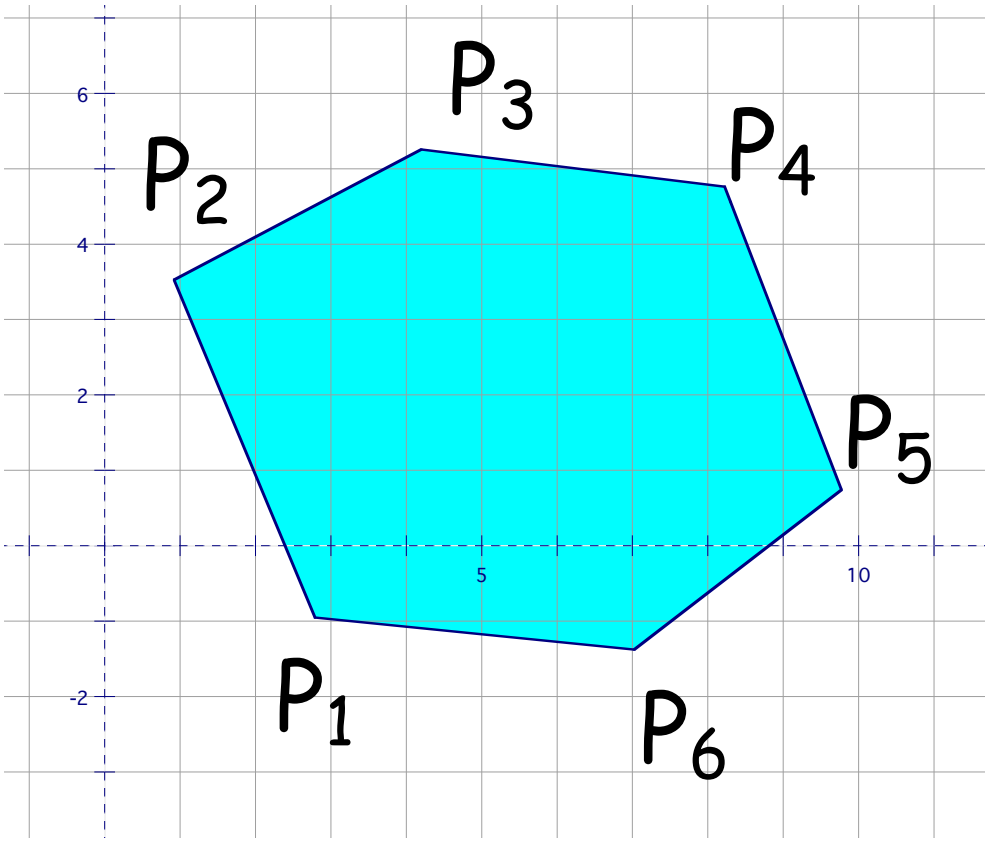
(d) Use the above to conclude that $P(a,b) \geq 0$ with equality if and only if $a = b$ (or $m = n$).

3.3 Jensen's Inequality

If P and Q are points in the coordinate plane, then it's easy to see that the set of points of the form $X = X(t) = (1-t)P + tQ,\ 0 \leq t \leq 1$ is precisely the line segment joining P and Q. We shall call such a point a **convex combination** of P and Q. More generally, if $P_1, P_2, \ldots, P_n$ are points in the plane, and if $t_1, t_2, \ldots, t_n$ are non-negative real numbers satisfying $t_1 + t_2 + \cdots + t_n = 1$, then the point

$$X = X(t) = t_1P_1 + t_2P_2 + \cdots + t_nP_n$$

is a convex combination of $P_1, P_2, \ldots, P_n$. This set is precisely smallest convex polygon in the plane containing the points $P_1, P_2, \ldots, P_n$. Such a polygon is depicted below:



Next, recall from elementary differential calculus that a twice-differentiable function f is concave down on an interval $[a, b]$ if $f''(x) \leq 0$ for all $x \in [a, b]$. Geometrically, this means that if $a \leq c \leq d \leq b$ then the convex combination of the points $P = P(c, f(c))$ and $Q = Q(d, f(d))$ lies on or below the graph of $y = f(x)$. Put more explicitly, this says that when $a \leq c \leq d \leq b$, and when $0 \leq t \leq 1$,

$$f((1-t)c + td) \geq (1-t)f(c) + tf(d).$$

LEMMA 1. *(Jensen's Inequality) Assume that the twice-differentiable function f is concave down on the interval $[a, b]$ and assume that $x_1, x_2, \ldots, x_n \in [a, b]$. If $t_1, t_2, \ldots, t_n$ are non-negative real numbers with $t_1 + t_2 + \cdots + t_n = 1$, then*

$$f(t_1x_1 + t_2x_2 + \cdots + t_nx_n) \geq t_1f(x_1) + t_2f(x_2) + \cdots + t_nf(x_n).$$

PROOF. We shall argue by induction on n with the result already being

true for $n = 2$. We may assume that $0 \leq t_n < 1$; set

$$x_0 = \frac{t_1 x_1 + \cdots + t_{n-1} x_{n-1}}{1 - t_n};$$

note that $x_0 \in [a, b]$. We have

$$\begin{aligned} f(t_1 x_1 + t_2 x_2 + \cdots + t_n x_n) &= f((1 - t_n) x_0 + t_n x_n) \\ &\geq (1 - t_n) f(x_0) + t_n f(x_n) \quad \text{(by induction)} \\ &\geq (1 - t_n) \left(\frac{t_1}{1 - t_n} \cdot f(x_1) + \cdots + \frac{t_{n-1}}{1 - t_n} \cdot f(x_{n-1}) \right) \\ &+ t_n f(x_n) \quad \text{(induction again)} \\ &= t_1 f(x_1) + t_2 f(x_2) + \cdots + t_n f(x_n), \end{aligned}$$

and we're finished.

3.4 The Hölder Inequality

Extending the notion of quadratic mean, we can define, for any real number $p \geq 1$ the "p-mean" of positive real numbers $x_1, \ldots, x_n$:

$$\mathrm{pM}(x_1, x_2, \ldots, x_n) = \sqrt[p]{\frac{x_1^p + x_2^p + \cdots + x_n^p}{n}}.$$

We shall show that if $1 \leq p \leq q$ that for positive real numbers $x_1, \ldots, x_n$ one has

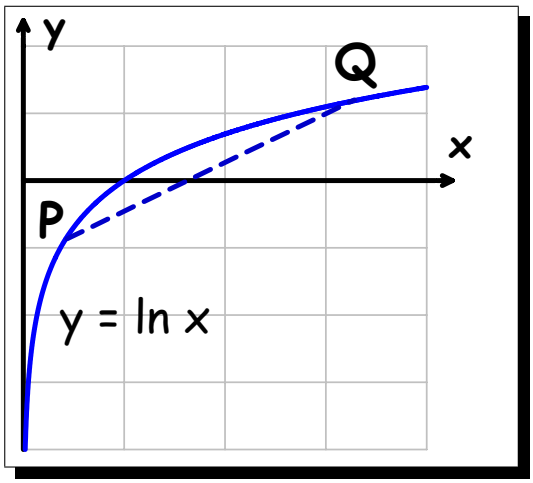
$$\mathrm{pM}(x_1, x_2, \ldots, x_n) \leq \mathrm{qM}(x_1, x_2, \ldots, x_n).$$

The proof is not too difficult—a useful preparatory result is **Young's inequality**, below.

LEMMA 2. (Young's Inequality) *Given real numbers $0 \leq a, b$ and $0 < p, q$ such that $\dfrac{1}{p} + \dfrac{1}{q} = 1$, one has*

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q},$$

with equality if and only if $a^p = b^q$.



PROOF. The proof involves a geometrical fact about the graph of the function $y = \ln x$, namely that for any two points P, Q on the graph, the straight line segment on the graph is completely below the graph.[5] Thus, let $P = P(a^p, \ln(a^p))$ and $Q = Q(b^q, \ln(b^q))$ be two points on the graph of $y = \ln x$. For any value of the parameter t, $0 \leq t \leq 1$, the point $X = X(tb^q + (1-t)a^q, t\ln(b^q) + (1-t)\ln(a^p))$ is a point on the line seqment PQ. Since the graph of $y = \ln x$ lies entirely above the line segment PQ, we conclude that

$$\ln(tb^q + (1-t)a^p) \geq t\ln(b^q) + (1-t)\ln(a^p) = tq\ln b + (1-t)p\ln a.$$

Now set $t = 1/q$ and infer that

$$\ln\left(\frac{b^q}{q} + \frac{a^p}{p}\right) \geq \ln b + \ln a = \ln(ab).$$

Exponentiating both sides yields the desired result, namely that

$$\frac{b^q}{q} + \frac{a^p}{p} \geq ab.$$

THEOREM 2. (Hölder's Inequality) *Given real numbers* $x_1, \ldots, x_n,\ y_1, \ldots, y_n$, *and given non-negative real numbers* p *and* q *such that* $\dfrac{1}{p} + \dfrac{1}{q} = 1$ *then*

$$\sum_{i=1}^{n} |x_i y_i| \leq \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} \left(\sum_{j=1}^{n} |y_j|^q\right)^{1/q}$$

PROOF. Let

$$A = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p},\ B = \left(\sum_{i=1}^{n} |y_i|^q\right)^{1/q}$$

[5] Another way to say this is, of course, that the graph of $y = \ln x$ is **concave down**.

We may assume that both $A,\ B \neq 0$, else the theorem is clearly true. Therefore by using Young's inequality, we see that for each $i = 1, 2, \ldots, n$, that

$$\frac{|x_i|}{A} \cdot \frac{|y_i|}{B} \leq \frac{|x_i|^p}{pA^p} + \frac{|y_i|^q}{qB}.$$

Therefore,

$$\begin{aligned} \frac{1}{AB} \sum_{i=1}^{n} |x_i y_i| &\leq \sum_{i=1}^{n} \left(\frac{|x_i|^p}{pA^p} + \frac{|y_i|^q}{qB} \right) \\ &= \frac{1}{p} + \frac{1}{q} = 1. \end{aligned}$$

This implies that

$$\sum_{i=1}^{n} |x_i y_i| \leq AB = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} \left(\sum_{j=1}^{n} |y_j|^q \right)^{1/q},$$

and we're done.

Note that if we set all $y_i = 1$ then we get

$$\sum_{i=1}^{n} |x_i| \leq n^{1/q} \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} = n^{1-1/p} \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p},$$

and so

$$\frac{1}{n} \sum_{i=1}^{n} |x_i| \leq \left(\sum_{i=1}^{n} |x_i|^p / n \right)^{1/p}$$

for any $p > 1$. This proves that

$$\mathrm{AM}(|x_1|, |x_2|, \ldots, |x_n|) \leq pM(|x_1|, |x_2|, \ldots, |x_n|)$$

whenever $p > 1$.

Finally, assume that $0 < p < q$ and assume that $x_1,\ x_2,\ \ldots,\ x_n$ are non-negative. We shall show that

$$\mathrm{pM}(x_1, x_2, \ldots, x_n) \leq \mathrm{qM}(x_1, x_2, \ldots, x_n).$$

Indeed, from the above, we have by setting $r = \dfrac{q}{p} > 1$ that

$$\begin{aligned}\sum_{i=1}^{n} {x_i}^p/n &= \mathrm{AM}(x_1^p, x_2^p, \ldots, x_n^p) \\ &\leq \mathrm{rM}(x_1^p, x_2^p, \ldots, x_n^p) \\ &= \left(\sum_{i=1}^{n} (x_i^p)^{q/p}/n\right)^{p/q} = \left(\sum_{i=1}^{n} x_i^q/n\right)^{p/q}.\end{aligned}$$

Taking the p-th roots of both sides yields what we were after, viz.,

$$\mathrm{pM}(x_1, x_2, \ldots, x_n) = \left(\sum_{i=1}^{n} {x_i}^p/n\right)^{1/p} \leq \left(\sum_{i=1}^{n} x_i^q/n\right)^{1/q} = \mathrm{qM}(x_1, x_2, \ldots, x_n).$$

EXERCISES.

1. Show how Young's inequality proves that $\mathrm{GM}(x_1, x_2) \leq \mathrm{AM}(x_1, x_2)$, where $x_1, x_2 \geq 0$.

2. Use Jensen's inequality and the fact that the graph of $y = \ln x$ is concave down to obtain a simple proof that

$$\mathrm{AM}(x_1, x_2, \ldots, x_n) \geq \mathrm{GM}(x_1, x_2, \ldots, x_n),$$

   where $x_1, x_2, \ldots, x_n \geq 0$.

3. Use Jensen's inequality to prove that given interior angles A, B, and C of a triangle then

$$\sin A + \sin B + \sin C \leq 3\sqrt{2}/2.$$

   Conclude that for a triangle $\triangle ABC$ inscribed in a circle of radius R, the maximum perimeter occurs for an equilateral triangle. (See Exercise 2 on page 34.)

4. Given $\triangle ABC$ with area K and side lengths a, b, and c, show that

$$ab + ac + bc \geq 4\sqrt{3}K.$$

   Under what circumstances does equality obtain? (Hint: note that $6K = ab\sin C + ac\sin B + bc\sin A$; use Cauchy-Schwarz together with Exercise 3, above.)

3.5 The Discriminant of a Quadratic

The discriminant of a quadratic polynomial, while finding itself in (mostly trivial) discussions in a typical high-school Algebra II course, nonetheless is a highly underused and too narrowly understood concept. This and the next two sections will attempt to provide meaningful applications of the discriminant, as well as put it in its proper algebraic perspective. Before proceeding, let me remind the reader that a possibly surprising application of the discriminant has already occurred in the proof of the Cauchy-Schwarz inequality (page 150).

Given the quadratic polynomial $f(x) = ax^2 + bx + c,\ a, b, c \in \mathbb{R}$, the **discriminant** is defined by the familiar recipe:

$$D \ = \ b^2 - 4ac.$$

This expression is typically introduced as a by-product of the quadratic formula expressing the two roots α, β of the equation $f(x) = 0$ as

$$\alpha,\ \beta \ = \ \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \ = \ \frac{-b \pm \sqrt{D}}{2a}.$$

From the above, the following simple trichotomy emerges.

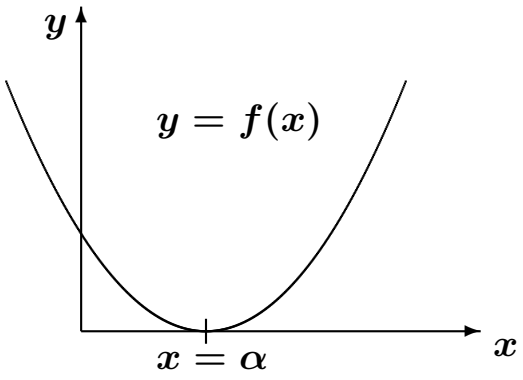
$D > 0 \iff f(x) = 0$ has two distinct real roots;

$D < 0 \iff f(x) = 0$ has two imaginary conjugate roots;

$D = 0 \iff f(x) = 0$ has a double real root.

Note that if $f(x) = ax^2 + bx + c$ with $a > 0$, then the condition $D \leq 0$ implies the unconditional inequality $f(x) \geq 0$.

The $D = 0$ case is the one we shall find to have many applications, especially to constrained extrema problems. Indeed, assuming this to be the case, and denoting α as the double root of $f(x) = 0$, then we have $f(x) = a(x - \alpha)^2$ and that the graph of $y = f(x)$ appears as depicted to the right.

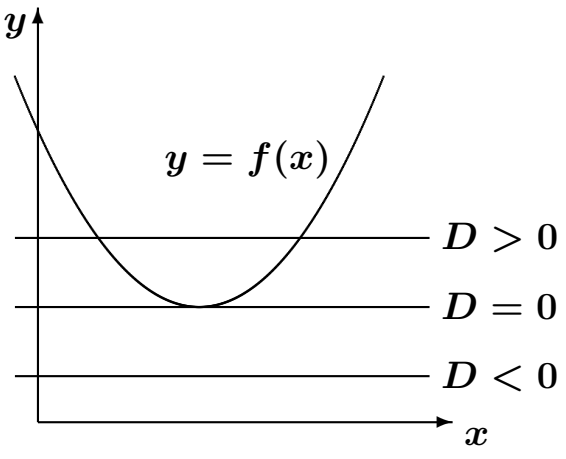


The geometrical implication of the double real root is that the graph of $y = f(x)$ not only has an x-intercept at $x = \alpha$, it is tangent to the x-axis at this point. We shall find this observation extremely useful, as it provides application to a wealth of constrained extrema problems.

The first example is rather pedestrian, but it will serve to introduce the relevant methodology.

EXAMPLE 1. Find the minimum value of the quadratic function $f(x) = 2a^2 - 12x + 23$.

SOLUTION. If we denote this minimum by m, then the graph of $y = m$ will be tangent to the graph of $y = f(x)$ where this minimum occurs. This says that in solving the quadratic equation $f(x) = m$, there must be a double root, i.e., the discriminant of the quadratic $2x^2 - 12x + 23 - m$ must vanish. The geometry of this situation is depicted to the right.



Solving for the discriminant in terms of m, we quickly obtain

$$\begin{aligned} 0 &= b^2 - 4ac \\ &= (-12)^2 - 4 \cdot 2 \cdot (23 - m) \\ &= 144 - 8(23 - m) \\ &= 144 - 184 + 8m = -40 + 8m \end{aligned}$$

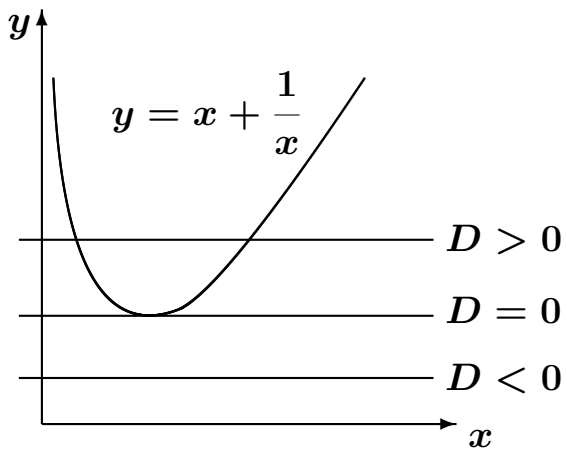
and so one obtains the minimum value of $m = 5$. Of course, this is not the "usual" way students are taught to find the extreme values of a quadratic function: they use the method of "completing the square" (another useful technique).

EXAMPLE 2. Here's an ostensibly harder problem. Find the minimum value of the function $g(x) = x + \dfrac{1}{x},\ x > 0$. Before going further, note that the ideas of Section 3.1 apply very naturally: from

$$0 \leq \left(\sqrt{x} - \frac{1}{\sqrt{x}}\right)^2 = x + \frac{1}{x} - 2$$

we see immediately that $x + \dfrac{1}{x} \geq 2$ with equality precisely when $x = 1$. That is the say, the minimum value of the objective function $x + \dfrac{1}{x}$ is 2.

SOLUTION. Denoting this minimum by m, then the graph of $y = m$ will again be tangent to the graph of $y = g(x)$ where this minimum occurs. Here, the equation is $x + \dfrac{1}{x} = m$, which quickly transforms to the quadratic equation $x^2 - mx + 1 = 0$. For tangency to occur, one must have that the discriminant of $x^2 - mx + 1$ vanishes.

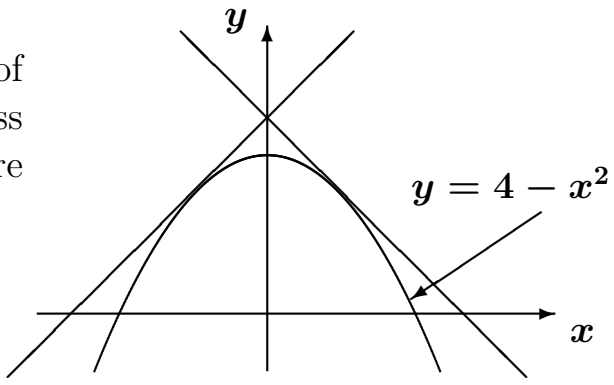


We have

$$\begin{aligned} 0 &= b^2 = 4ac \\ &= m^2 - 4 \end{aligned}$$

which immediately gives $m = \pm 2$. Only the value $m = 2$ is relevant here (as we assumed that $x > 0$) and this is the sought-after minimum value of g. (The extraneous value $m = -2$ can easily be seen to be the maximum value of $x + \dfrac{1}{x},\ x < 0$.)

EXAMPLE 3. Find the equations of the two straight lines which pass through the point $P(0, 5)$ and are both tangent to the graph of $y = 4 - x^2$.



SOLUTION. If we write a line with equation $\ell : y = 5 + mx$, where the slope is to be determined, then we are solving $4 - x^2 = 5 + mx$ so that a double root occurs (i.e., tangency). Clearly, there should result two values of m for this to happen. Again, the discriminant is a very good tool. Write the quadratic equation having the multiple root as $x^2 + mx + 1 = 0$, and so

$$\begin{aligned} 0 &= b^2 - 4ac \\ &= m^2 - 4 \Rightarrow m = \pm 2. \end{aligned}$$

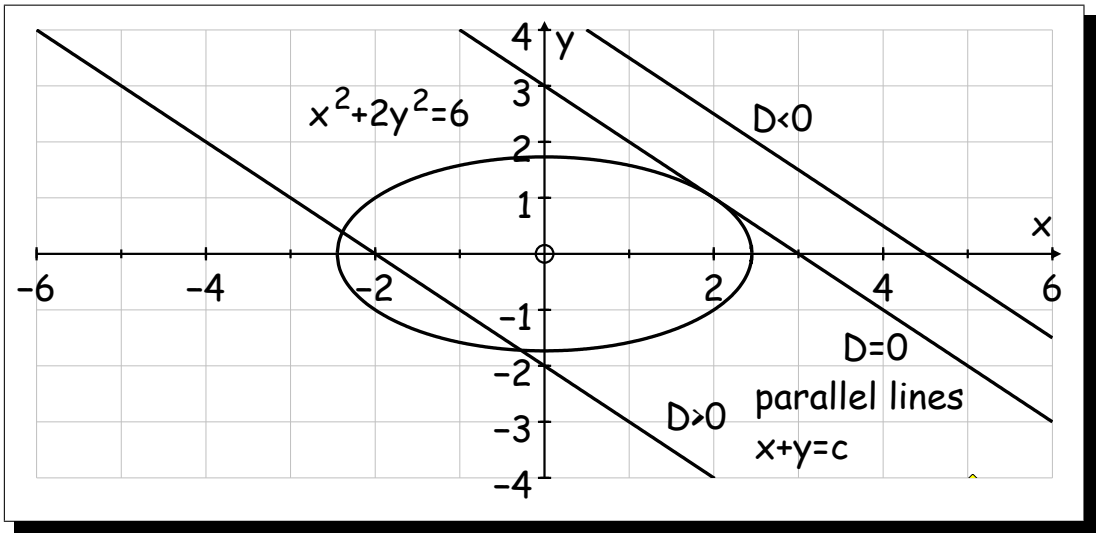
Therefore, the two lines are given by equations

$$y = 5 + 2x \quad \text{and} \quad y = 5 - 2x.$$

(The two points of tangency are at the points with coordinates $(\pm 1, 3)$.)

EXAMPLE 4. Given that $x^2 + 2y^2 = 6$, find the maximum value of $x + y$.

SOLUTION. This problem appears quite a bit different (and more difficult) than the preceding examples, but it's not, and it fits in very well to the present discussion.[6] This problem is very geometrical in nature, as the "constraint equation" $x^2 + 2y^2 = 6$ is an ellipse and the graphs of $x + y = c$ (c = constant) are parallel lines (with slope -1). We seek that value of c which gives the maximum value of $x + y$. See the graphic below:



Clearly the maximum value of $x + y$ will occur where this line is tangent to the ellipse. There will be two points of tangency, one in the third quadrant (where a minimum value of $x+y$ will occur) and one in the first quadrant (where the maximum value of $x + y$ occurs). Next, if we solve $x + y = c$ for y and substitute this into $x^2+2y^{=}6$, then a quadratic equation in x will result. For tangency to occur, one must have that the discriminant is 0. From $y = c-x$, obtain

$$x^2 + 2(c - x)^2 - 6 = 0 \Longrightarrow 3x^2 - 4cx + 2c^2 - 6 = 0.$$

This leads to

$$0 = D = 16c^2 - 12(2c^2 - 6) = -8c^2 + 72 \Longrightarrow c = \pm 3.$$

[6]Problems of this sort are often not considered until such courses as Calculus III, where the method of Lagrange multipliers is applied.

Therefore, the maximum value of $x + y$ is 3 (and the minimum value of $x + y$ is -3).

EXERCISES.

1. Given that $x + y = 1,\ x,\ y > 0$, find the minimum value of $\dfrac{1}{x} + \dfrac{1}{y}$.

2. Given that $\dfrac{1}{x} + \dfrac{1}{y} = 1,\quad x,\ y > 0$, prove that $x + y \geq 4$.

   (Exercises 1 and 2 can be solved very simply by multiplying together $\dfrac{1}{x} + \dfrac{1}{y}$ and $x + y$ and using the result of Example 2.)

3. Find the distance from the origin to the line with equation $x + 3y = 6$.

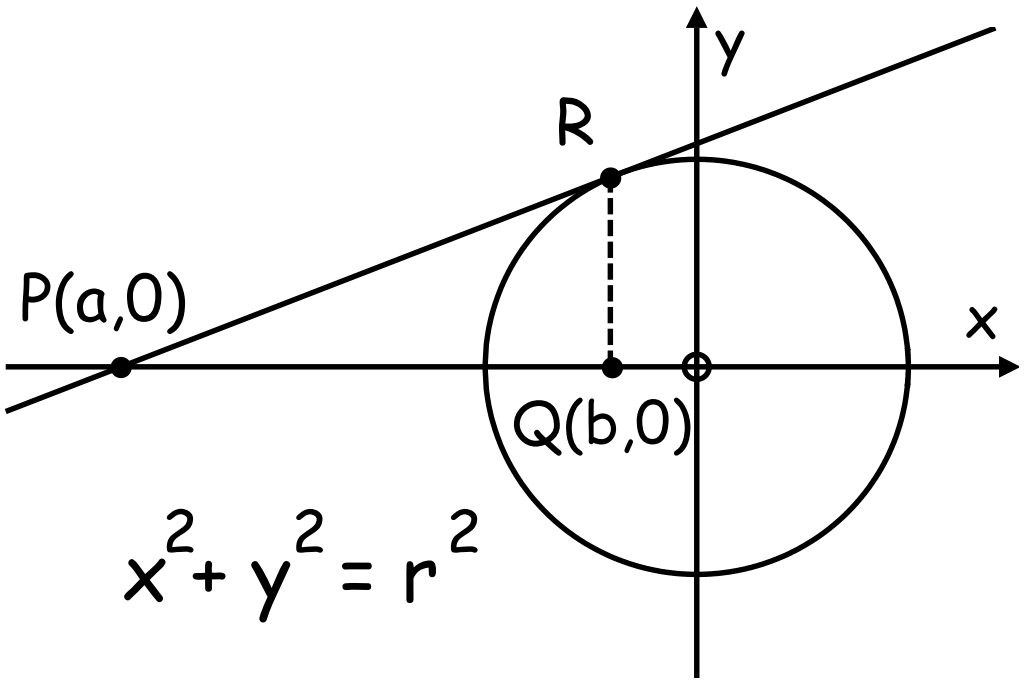
4. Given that $\dfrac{x}{y} + y = 1$ find the minimum value of $x + y,\ x,\ y > 0$.

5. Find the largest value of a so that the parabola with equation $y = a - x^2$ is tangent to the circle with graph $x^2 + y^2 = 4$. Go on to argue that this value of a is the maximum of the function $x^2 + y$ given that $x^2 + y^2 = 4$.

6. Let $f(x) = ax^2 + bx + c$, and so the derivative is $f'(x) = 2ax + b$. Denote by $\mathcal{R}(f)$ the determinant

$$\mathcal{R}(f) \;=\; \det \begin{bmatrix} a & b & c \\ 2a & b & 0 \\ 0 & 2a & b \end{bmatrix}.$$

   Show that $\mathcal{R}(f) = -aD(f)$, where $D = D(f)$ is the discriminant of $f(x)$.



7. Above is depicted the circle whose equation is $x^2 + y^2 = r^2$, as well as the tangent line to this circle at the point R. The point $P = P(a, 0)$ is the intersection of this tangent line with the x-axis and the point $Q = Q(b, 0)$ as the same x-coordinate as the point R.

   (a) Using a discriminant argument show that if m is the slope of the tangent line, then

   $$m^2 = \frac{r^2}{a^2 - r^2}.$$

   Use this to show that $b = r^2/a$.

   (b) Using the Secant-Tangent Theorem (see page 32), give another proof of the fact that $b = r^2/a$. Which is easier?

3.6 The Discriminant of a Cubic

The the quadratic $f(x) = ax^2 + bx + c$ has associated with it the discriminant $D = b^2 - 4ac$, which in turn elucidates the nature of the zeros of $f(x)$. In turn, this information gives very helpful information about tangency which in turn can be applied to constrained extrema

problems. This raises at least a couple of questions. The immediate question raised here would be whether higher-degree polynomials also have discriminants. We'll see that this is the case and will consider the case of cubic polynomials in this section. In the following section we'll introduce the discriminant for arbitrary polynomials. The notion of the determinant of a matrix will play a central role here.

For the quadratic polynomial $f(x) = ax^2+bx+c$ having zeros x_1, x_2, we define the "new" quantity

$$\Delta = a^2 \det \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \end{bmatrix}^2 = a^2(x_2 - x_1)^2.$$

At first blush, it doesn't appear that Δ has anything to do with the discriminant D. However, once we have designated the zeros of f as being x_1 and x_2, then the **Factor Theorem** dictates that

$$f(x) = a(x - x_1)(x - x_2).$$

Since also $f(x) = ax^2 + bx + c$ we conclude by expanding the above that

$$b = -a(x_1 + x_2), \quad \text{and} \quad c = ax_1x_2.$$

Now watch this:

$$\begin{aligned} \Delta &= a^2(x_2 - x_1)^2 \\ &= a^2(x_1^2 + x_2^2 - 2x_1x_2) \\ &= a^2[(x_1 + x_2)^2 - 4x_1x_2] \\ &= [-a(x_1 + x_2)]^2 - 4a(ax_1x_2) \\ &= b^2 - 4ac = D. \end{aligned}$$

In other words, Δ and D **are the same**:

$$\boxed{D = \Delta}.$$

Therefore, D and Δ will satisfy the same trichotomy rule. But let's try to develop the trichotomy rule directly in terms of Δ instead of D.

Case (i): $\Delta > 0$. That is to say, $(x_2 - x_1)^2 > 0$ and so certainly the zeros x_1 and x_2 are distinct. Furthermore, if they were not real, then they would have to be complex conjugates of one another and this would force (think about it!) $(x_2 - x_1)^2 < 0$ (as $x_2 - x_1$ would be purely imaginary). Therefore

$$\Delta > 0 \Longrightarrow \text{ } Q \text{ has two distinct real zeros.}$$

Case (ii): $\Delta = 0$. This is clear in that one immediately has that $x_1 = x_2$. That is to say

$$\Delta = 0 \Longrightarrow \text{ } Q \text{ has a double zero.}$$

Case (iii): $\Delta < 0$. Since $(x_2 - x_1)^2 < 0$ we certainly cannot have both x_1 and x_2 real. Therefore, they're both complex (non-real) as they are complex conjugate. Therefore

$$\Delta < 0 \Longrightarrow \text{ } Q \text{ has two complex (non-real) zeros.}$$

That is to say, D and Δ satisfy the same trichotomy law!

Whereas the definition of D does not suggest a generalization to higher-degree polynomials, the definition of Δ can be easily generalized. We consider first a natural generalization to the cubic polynomial

$$P(x) = ax^3 + bx^2 + cx + d, \quad a,\, b,\, c,\, d \in \mathbb{R},\ a \neq 0.$$

By the **Fundamental Theorem of Algebra**, we know that (counting multiplicities), $P(x)$ has three zeros; we shall denote them by x_1, x_2, and x_3. They may be real or complex, but we do know that one of these zeros must be real.

We set

$$\Delta = a^4 \det \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \end{bmatrix}^2 .$$

With a bit of effort, this determinant can be expanded. It's easier to first compute the determinant of the matrix

$$\begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \end{bmatrix}$$

and then square the result. One has, after a bit of computation, the highly structured answer

$$\det \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \end{bmatrix} = (x_3 - x_2)(x_3 - x_1)(x_2 - x_1),$$

(this is generalized in the next section) which implies that

$$\Delta = a^4(x_3 - x_2)^2(x_3 - x_1)^2(x_2 - x_1)^2.$$

This is all well and good, but two questions immediately arise:

- How does one compute Δ without knowing the zeros of P? Also, and perhaps more importantly,
- what is Δ trying to tell us?

Let's start with the second bullet point and work out the trichotomy law dictated by Δ.

If $P(x)$ has three distinct real zeros, then it's obvious that $\Delta > 0$. If not all of the zeros are real, then $P(x)$ has one real zero (say x_1) and a complex-conjugate pair of non-real zeros (x_2 and x_3). In this case $(x_2 - x_1)$, $(x_3 - x_1)$ would be a complex conjugate pair, forcing $0 < (x_2 - x_1)(x_3 - x_1) \in \mathbb{R}$ and so certainly that $0 < (x_2 - x_1)^2(x_3 - x_1)^2 \in \mathbb{R}$. Furthermore, $(x_3 - x_2)$ is purely imaginary and so $(x_3 - x_2)^2 < 0$, all forcing $\Delta < 0$. Therefore, we see immediately that

$$\Delta > 0 \Longrightarrow P(x) \text{ has three distinct real zeros}$$

and that

$\Delta < 0 \Longrightarrow P(x)$ has one real zero and two non-real complex zeros.

This is all rounded out by the obvious statement that

$$\Delta = 0 \Longrightarrow P(x) \text{ has a multiple zero and all zeros are real.}$$

Of course, none of the above is worth much unless we have a method of computing Δ. The trick is to proceed as in the quadratic case and compute Δ in terms of the coefficients of $P(x)$. We start with the observation that

$$P(x) = a(x - x_1)(x - x_2)(x - x_3),$$

all of which implies that (by expanding)

$$b = -a(x_1 + x_2 + x_3),\ c = a(x_1x_2 + x_1x_3 + x_2x_3),\ d = -ax_1x_2x_3.$$

We set

$$\sigma_1 = x_1 + x_2 + x_3,\ \sigma_2 = x_1x_2 + x_1x_3 + x_2x_3,\ \sigma_3 = x_1x_2x_3,$$

and call them the **elementary symmetric polynomials** (in x_1, x_2, x_3). On the other hand, by expanding out Δ, one has that (after quite a bit of very hard work!)

$$\begin{aligned}
\Delta &= a^4(x_3 - x_2)^2(x_3 - x_1)^2(x_2 - x_1)^2 \\
&= a^4(-4\sigma_1^3\sigma_3 + \sigma_1^2\sigma_2^2 + 18\sigma_1\sigma_2\sigma_3 - 4\sigma_2^3 - 27\sigma_3^2) \\
&= -4b^3d + b^2c^2 + 18abcd - 4ac^3 - 27a^2d^2
\end{aligned}$$

giving a surprisingly complicated homogeneous polynomial in the coefficient a, b, c, and d. (See Exercise 6 below for a more more direct method for computing Δ.)

We'll close this section with a representative example. Keep in mind that just as in the case of the quadratic, when the discriminant of a cubic is 0, then the graph of this cubic is tangent to the x-axis at the multiple zero.

EXAMPLE. Compute the minimum value of the function $f(x) = \dfrac{1}{x^2} + x,\ x > 0.$

SOLUTION. The minimum value will occur where the line $y = c$ is tangent to the graph of $y = f(x)$. We may write the equation $f(x) = c$ in the form of a cubic polynomial in x:

$$x^3 - cx^2 + 1 = 0.$$

As the tangent indicates a multiple zero, we must have $\Delta = 0$. As $a = 1,\ b = -c,\ c = 0,\ d = 1$, we get the equation $4c^3 - 27 = 0$, which implies that the minimum value is given by $c = \dfrac{3}{\sqrt[3]{4}}$ (which can be verified by standard calculus techniques).

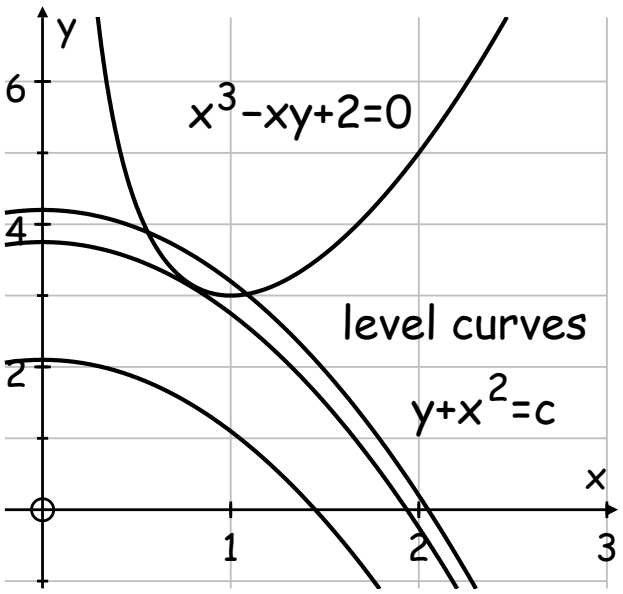
Now try these:

EXERCISES.

1. Compute the minimum of the function $h(x) = \dfrac{1}{x} + x^2,\ x > 0.$

2. Compute the minimum of $2y - x$ given that $x^3 - x^2y + 1 = 0.$

3. Compute the maximum value of $y + x^2$ given that $x^3 - xy + 2 = 0.$



4. Compute the maximum value of xy, given that $x^2 + y = 4$.

5. The polynomial $S(x_1, x_2, x_3) = x_1^3+x_2^3+x_3^3$ is symmetric in x_1, x_2, x_3 and can be expanded in the elementary symmetric polynomials

$$\sigma_1 = x_1 + x_2 + x_3,\ \sigma_2 = x_1x_2 + x_1x_3 + x_2x_3,\ \sigma_3 = x_1x_2x_3.$$

Watch this:

$$\begin{aligned} x_1^3 + x_2^3 + x_3^3 &= (x_1 + x_2 + x_3)^3 \\ &- 3x_1^2(x_2 + x_3) - 3x_2^2(x_1 + x_3) - 3x_3^2(x_1 + x_2) - 6x_1x_2x_3 \\ &= (x_1 + x_2 + x_3)^3 - 3(x_1 + x_2 + x_3)(x_1x_2 + x_1x_3 + x_2x_3) \\ &+ 3x_1x_2x_3 \\ &= \sigma_1^3 - 2\sigma_1\sigma_2 + 3\sigma_3. \end{aligned}$$

Now try to write the symmetric polynomial $x_1^4 + x_2^4 + x_3^4$ as a polynomial in σ_1, σ_2, σ_3.

6. Let $f(x) = ax^3 + bx^2 + cx + d$, and so the derivative is $f'(x) = 3ax^2 + 2bx+$c. Denote by $\mathcal{R}(f)$ the determinant

$$\mathcal{R}(f) = \det \begin{bmatrix} a & b & c & d & 0 \\ 0 & a & b & c & d \\ 3a & 2b & c & 0 & 0 \\ 0 & 3a & 2b & c & 0 \\ 0 & 0 & 3a & 2b & c \end{bmatrix}.$$

Show that $\mathcal{R}(f) = -aD(f)$, where $D = D(f)$ is the discriminant of $f(x)$. (This is the generalization of the result of Exercise 6 to cubic polynomials.)

3.7 The Discriminant (Optional Discussion)

In this section I'll give a couple of equivalent definitions of the discriminant of a polynomial of arbitrary degree. Not all proofs will be given, but an indication of what's involved will be outlined. To this end, let there be given the polynomial

$$f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0,$$

where $a_n \neq 0$ and where all coefficients are real. Denoting by x_1, x_1, $\ldots$, x_n the zeros of $f(x)$ (which may include several complex-conjugate pairs of imaginary zeros), we know that (by the Factor Theorem)

$$f(x) = a_n(x - x_1)(x - x_2)\cdots(x - x_n).$$

In analogy with the above work, we define the **discriminant** of $f(x)$ by setting

$$\Delta = \Delta(f) = a_n^{2n-2}\det \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_n^{n-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{n-1} \\ \vdots & & & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{n-1} \end{bmatrix}^2.$$

The above involves the determinant of the so-called **Vandermonde matrix**,

$$V \;=\; \det \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_n^{n-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{n-1} \\ \vdots & & & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{n-1} \end{bmatrix}$$

which makes frequent appearances throughout mathematics. Its determinant is given in the next theorem.

THEOREM 3. *det* $V = \prod_{i<j}(x_j - x_i)$.

PROOF. We argue by induction on n. Setting $\Delta = \det V$, we start by subtracting row 1 from rows $2, 3, \ldots, n$, which quickly produces

$$\Delta \;=\; \det \begin{bmatrix} x_2 - x_1 & x_2^2 - x_1^2 & \cdots & x_2^{n-1} - x_1^{n-1} \\ x_3 - x_1 & x_3^2 - x_1^2 & \cdots & x_3^{n-1} - x_1^{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ x_n - x_1 & x_n^2 - x_1^2 & \cdots & x_n^{n-1} - x_1^{n-1} \end{bmatrix}.$$

Next, in each row we factor out the common factor of $x_i - x_1$, $i = 2, 4, \ldots, n$, which leads to

$$\Delta \;=\; (x_2 - x_1)(x_3 - x_1)\cdots(x_n - x_1)\times$$
$$\det \begin{bmatrix} 1 & x_2 + x_1 & x_2^2 + x_2x_1 + x_1^2 & \cdots & x_2^{n-2} + x_2^{n-3}x_1 + \cdots + x_1^{n-2} \\ 1 & x_3 + x_1 & x_3^2 + x_2x_1 + x_1^2 & \cdots & x_3^{n-2} + x_3^{n-3}x_1 + \cdots + x_1^{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n + x_1 & x_n^2 + x_nx_1 + x_1^2 & \cdots & x_n^{n-2} + x_n^{n-3}x_1 + \cdots + x_1^{n-2} \end{bmatrix}.$$

Next, if we subtract x_1 times column $n-2$ from column $n-1$, then subtract x_1 times column $n-3$ from column $n-2$, and so on, we'll eventually reach

$$\Delta \;=\; (x_2 - x_1)(x_3 - x_1)\cdots(x_n - x_1) \;\times \det \begin{bmatrix} 1 & x_2 & x_2^2 & \cdots & x_2^{n-2} \\ 1 & x_3 & x_3^2 & \cdots & x_3^{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{n-2} \end{bmatrix}$$

$$= \;(x_2 - x_1)(x_3 - x_1)\cdots(x_n - x_1) \times \prod_{j>i\geq 2}(x_j - x_i) = \prod_{j>i}(x_j - x_i),$$

From the above, we see that

$$\Delta = \Delta(f) = a_n^{2n-2} \prod_{1 \le i < j \le n} (x_j - x_i)^2.$$

The difficulty with the above expression is that its computation appears to require the zeros of $f(x)$. However, this is a symmetric polynomial in the "variables" $x_1, x_2, \ldots, x_n$ and hence[7] can be written as a polynomial in the **elementary symmetric polynomials**

$$\begin{aligned}
\sigma_1 &= x_1 + x_2 + \cdots + \sigma_n \\
\sigma_2 &= x_1x_2 + x_1x_3 + \cdots = \sum_{i<j} x_ix_j \\
\sigma_3 &= \sum_{i<j<k} x_ix_jx_k \\
&\vdots \\
\sigma_n &= x_1x_2 \cdots x_n
\end{aligned}$$

This was carried out for the quadratic polynomial $f(x) = ax^2+bx+c$ on page 168; the result for the cubic polynomial $f(x) = ax^3 + bx^2 + cx + d$ was given on page 171. Carrying this out for higher degree polynomials is quite difficult in practice (see, e.g., Exercise 5, above). The next two subsections will lead to a more direct (but still computationally very complex) method for computing the discriminant of an arbitrary polynomial.

3.7.1 The resultant of $f(x)$ and $g(x)$

Let $f(x) = a_nx^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0$, and $g(x) = b_mx^m + b_{m-1}x^{m-1} + \cdots + b_1x + b_0$ be polynomials having real coefficients of degrees n and m, respectively. Define the $(n+m) \times (n+m)$ **Sylvester matrix** relative to $f(x)$ and $g(x)$, $\mathcal{S}(f, g)$ by setting

[7] This follows from the so-called **Fundamental Theorem on Symmetric Polynomials**.

$$\mathcal{S}(f,g) = \begin{bmatrix} a_n & a_{n-1} & a_{n-2} & \cdots & 0 & 0 & 0 \\ 0 & a_n & a_{n-1} & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & a_1 & a_0 & 0 \\ 0 & 0 & 0 & \cdots & a_2 & a_1 & a_0 \\ b_m & b_{m-1} & b_{m-2} & \cdots & 0 & 0 & 0 \\ 0 & b_m & b_{m-1} & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & b_1 & b_0 & 0 \\ 0 & 0 & 0 & \cdots & b_2 & b_1 & b_0 \end{bmatrix}.$$

The **resultant** $\mathcal{R}(f,g)$ of $f(x)$ and $g(x)$ is the determinant of the corresponding Sylvester matrix:

$$\mathcal{R}(f,g) = \det \mathcal{S}(f,g).$$

For example, if $f(x) = a_2x^2+a_1x+a_0$ and $g(x) = b_3x^3+b_2x^2+b_1x+b_0$, then

$$\mathcal{S}(f,g) = \det \begin{bmatrix} a_2 & a_1 & a_0 & 0 & 0 \\ 0 & a_2 & a_1 & a_0 & 0 \\ 0 & 0 & a_2 & a_1 & a_0 \\ b_3 & b_2 & b_1 & b_0 & 0 \\ 0 & b_3 & b_2 & b_1 & b_0 \end{bmatrix}.$$

Note that the resultant of two polynomials clearly remains unchanged upon field extension.

We aim to list a few simple—albeit technical—results about the resultant. The first is reasonably straightforward and is proved by just keeping track of the sign changes introduced by swapping rows in a determinant.

LEMMA 2.

$$\mathcal{R}(f,g) = (-1)^{mn}\mathcal{R}(g,f)$$

where $\deg f = n$ *and* $\deg g = m$.

Next, write

$$f(x) = a_n(x^n + a'_{n-1}x^{n-1} + \cdots + a'_1 x + a'_0), \;\; a'_i = a_i/a_n, \; i = 0, 1, \ldots, n-1;$$

similarly, write

$$g(x) = b_m(x^m + b'_{m-1}x^{m-1} + \cdots + b'_1 x + b'_0), \;\; b'_j = b_j/b_m, \; j = 0, 1, \ldots, n-1;$$

It follows easily that $\mathcal{R}(f, g) = a_n^m b_m^n \mathcal{R}(f/a_n, g/b_m)$, which reduces computations to resultants of **monic** polynomials.

THEOREM 4.. *Let $f(x)$, $g(x)$, and $h(x)$ be polynomials with real coefficients. Then*

$$\mathcal{R}(fg, h) = \mathcal{R}(f, h)\mathcal{R}(g, h).$$

PROOF. From the above, it suffices to assume that all polynomials are monic (have leading coefficient 1). Furthermore, by the **Fundamental Theorem of Algebra** $f(x)$ splits into linear factors, and so it is sufficient to prove that the above result is true when $f(x) = x + a$ is linear. Here, we let

$$g(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0, \text{ and}$$
$$h(x) = x^m + b_{m-1}x^{m-2} + \cdots + b_1 x + b_0.$$

First of all, one obviously has that

$$\mathcal{R}(x + a, h) = \det \begin{bmatrix} \mathcal{S}(x + a, h(x)) & \boldsymbol{Z} \\ \mathbf{0}_{n,m+1} & \boldsymbol{I}_n \end{bmatrix}$$

where $\boldsymbol{Z}$ is the $(m + 1) \times n$ matrix

$$\boldsymbol{Z} = \begin{bmatrix} 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 \\ -a_{n-1} & \cdots & -a_1 & -a_0 \end{bmatrix}.$$

Next, it is equally clear that

$$\mathcal{R}(g,h) = \det \begin{bmatrix} 1 & a_{n-1} & \cdots & a_1 & a_0 & 0 & \cdots & 0 \\ 0 & & & & & & & \\ 0 & & & & & & & \\ \vdots & & & \mathcal{S}(g,h) & & & & \\ \vdots & & & & & & & \\ 0 & & & & & & & \end{bmatrix}.$$

Finally, one checks that

$$\begin{bmatrix} \mathcal{S}(x+a,h(x)) & \boldsymbol{Z} \\ \mathbf{0}_{n,m+1} & \boldsymbol{I}_n \end{bmatrix} \times \begin{bmatrix} 1 & a_{n-1} & \cdots & a_1 & a_0 & 0 & \cdots & 0 \\ 0 & & & & & & & \\ 0 & & & & & & & \\ \vdots & & & \mathcal{S}(g,h) & & & & \\ \vdots & & & & & & & \\ 0 & & & & & & & \end{bmatrix}$$

$$= \mathcal{S}((x+a)g(x), h(x)),$$

and we're done.

Since $\mathcal{R}(x-a, x-b) = a-b$, we immediately obtain the following:

COROLLARY 1. *Let $f(x)$, $g(x)$ be polynomials with real coefficients, and have leading coefficients a_n and b_m, respectively. Assume $f(x)$ has zeros $\alpha_1, \ldots, \alpha_n$, and that $g(x)$ has zeros $\beta_1, \ldots, \beta_m$. Then*

$$\mathcal{R}(f,g) = a_n^m b_m^n \prod_{i=1}^{n} \prod_{j=1}^{m} (\alpha_i - \beta_j).$$

COROLLARY 2. *$\mathcal{R}(f,g) = 0$ if and only if $f(x)$ and $g(x)$ have a common zero.*

The following corollary will be quite important in the next section.

COROLLARY 3. *Let $f(x)$, $g(x)$ be polynomials with real coefficients, and have leading coefficients a_n and b_m, respectively. Assume $f(x)$ has*

zeros $\alpha_1, \ldots, \alpha_n$. *Then*

$$\mathcal{R}(f,g) = a_n^m \prod_{i=1}^{n} g(\alpha_i).$$

PROOF. Let $g(x)$ have zeros $\beta_1, \ldots, \beta_m$. Since

$$g(x) = b_m(x-\beta_1)\cdots(x-\beta_m),$$

we see that $g(\alpha_i) = b_m(\alpha_i-\beta_1)\cdots(\alpha_i-\beta_m)$. From this, the result follows instantly.

3.7.2 The discriminant as a resultant

As already given above, the discriminant of the polynomial $f(x) = a_n x^n +$ lower-degree terms and having zeros $x_1, x_2, \ldots, x_n$ is given by

$$\Delta(f) = a_n^{2n-2}\det\begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_n^{n-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{n-1} \\ \vdots & & & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{n-1} \end{bmatrix}^2 = a_n^{2n-2}\prod_{i<j}(x_j - x_i)^2.$$

We relate the above to the resultant as follows. Let $f(x) = a_n x^n$ + lower-degree terms, where $a_n \neq 0$, and let $\alpha_1, \ldots, \alpha_n$ be the zeros of $f(x)$. Then, using Corollary 3 above, and letting $f' = f'(x)$ be the derivative of f, we have that

$$\mathcal{R}(f,f') = a_n^{n-1}\prod_{i=1}^{n} f'(\alpha_i).$$

Next, we have $f(x) = a_n(x-\alpha)\cdots(x-\alpha)$; applying the product rule for differentiation leads quickly to

$$f'(x) = a_n\sum_{i=1}^{n}(x-\alpha_1)\cdots\widehat{(x-\alpha_i)}\cdots(x-\alpha_n),$$

where the convention is that the factor under the $\widehat{\ }$ is omitted. From the above, we see immediately that

$$f'(\alpha_i) = a_n \prod_{j \neq i} (\alpha_i - \alpha_j),$$

and so

$$\mathcal{R}(f, f') = a_n^{n-1} \prod_{i=1}^{n} f'(\alpha_i) = a_n^{2n-1} \prod_{i=1}^{n} \prod_{j \neq i} (\alpha_i - \alpha_j) = a_n^{2n-1} \prod_{j \neq i} (\alpha_i - \alpha_j).$$

Finally, one checks that

$$\prod_{j \neq i} (\alpha_i - \alpha_j) = (-1)^{n(n-1)/2} \prod_{1 \leq i < j \leq n} (\alpha_j - \alpha_i)^2,$$

which gives the following relationship between the result of $f(x)$ and $f'(x)$ and the discriminant:

THEOREM 5. *Given the polynomial $f(x)$ with real coefficients, one has that*

$$\mathcal{R}(f, f') = (-1)^{n(n-1)/2} a_n \Delta(f).$$

If we return to the case of the quadratic $f(x) = ax^2 + bx + c$, then

$$\begin{aligned} \mathcal{R}(f, f') &= \mathcal{R}(ax^2 + bx + c, 2ax + b) \\ &= \det \begin{bmatrix} a & b & c \\ 2a & b & 0 \\ 0 & 2a & b \end{bmatrix} \\ &= -(ab^2 - 4a^2c) = -a(b^2 - 4ac). \end{aligned}$$

Since for $n = 2$ we have $(-1)^{n(n-1)/2} = -1$, we see that, indeed, $\mathcal{R}(f, f') = (-1)^{n(n-1)/2} a \Delta(f)$ in this familiar case.

Note, finally, that as a result of the representation of $\Delta(f)$ in terms of $\mathcal{R}(f, f')$ we see that $\Delta(f)$ is a homogeneous polynomial of degree $2n - 2$ in the coefficients a_0, a_1, ..., a_n.

3.7.3 A special class of trinomials

We shall start this discussion with a specific example. Let $f(x) = a_3x^3 + a_2x^2 + a_1x + a_0$ and let $g(x) = b_2x^2 + b_1x + b_0$ and form the Sylvester matrix

$$\mathcal{S}(f,g) = \begin{bmatrix} a_3 & a_2 & a_1 & a_0 & 0 \\ 0 & a_3 & a_2 & a_1 & a_0 \\ b_2 & b_1 & b_0 & 0 & 0 \\ 0 & b_2 & b_1 & b_0 & 0 \\ 0 & 0 & b_2 & b_1 & b_0 \end{bmatrix}.$$

Next assume that in the above, we actually have $a_3 = a_2 = 0$, and that $b_2 \neq 0$. Then the determinant of the above is given by

$$\det \mathcal{S}(f,g) = \det \begin{bmatrix} b_2 & 0 & 0 & 0 & 0 \\ 0 & b_2 & 0 & 0 & 0 \\ 0 & 0 & a_1 & a_0 & 0 \\ 0 & 0 & 0 & a_1 & a_0 \\ 0 & 0 & b_2 & b_1 & b_0 \end{bmatrix} = b_2^2 \mathcal{R}(a_1x + a_0, b_2x^2 + b_1x + b_0).$$

In general, assume that

$$f(x) = a_nx^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0$$

and that

$$g(x) = b_mx^m + b_{m-1}x^{m-1} + \cdots + b_1x + b_0, \quad b_m \neq 0.$$

If we have $a_k \neq 0$ and that $a_{k+1} = a_{k+2} = \cdots = a_n = 0$, then one patterns an argument based on the above to arrive at the conclusion:

LEMMA 3. *With hypotheses as above,*

$$det\, \mathcal{S}(f,g) = (-1)^{m(n-k)} b_m^{n-k} \mathcal{R}(f,g).$$

Assume now that we are given polynomials $f(x) = a_nx^n +$ lower, and $g(x) = b_mx^m +$ lower. The Sylvester matrix $\mathcal{S}(f,g)$ has in rows 1 through m the coefficients of $f(x)$ and in rows $m+1$ through $m+n$

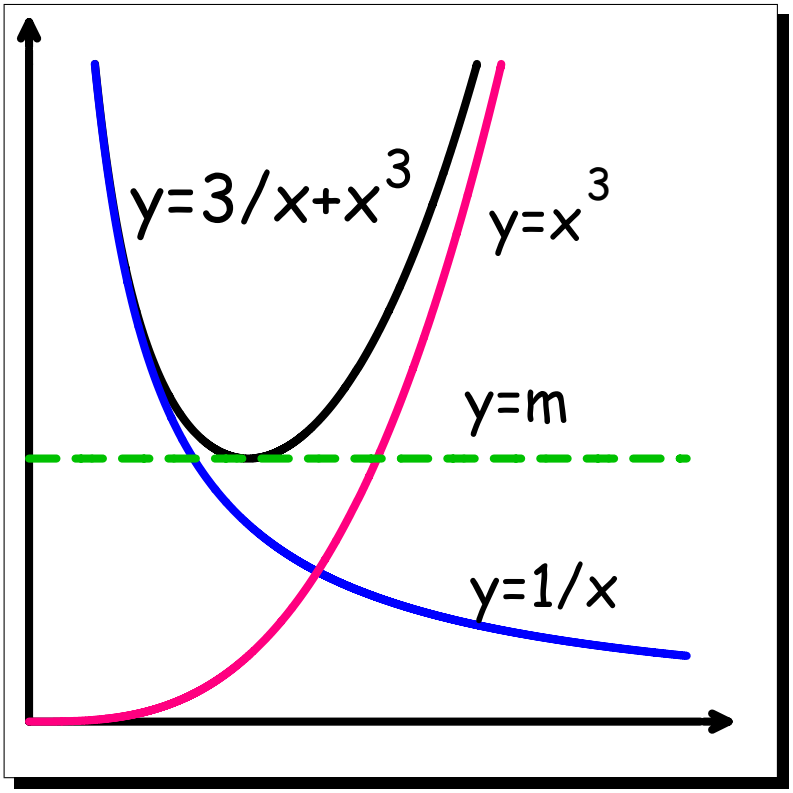
it has the coefficients of $g(x)$. Note that adding a multiple a of row $m+n$ to the first m rows of $\mathcal{S}(f,g)$ will produce the Sylvester matrix $\mathcal{S}(f+ag,g)$, whose determinant is unchanged. If $m<n$, then adding a times row $m+n-1$ to each of the first m rows of $\mathcal{S}(f,g)$ will produce the Sylvester matrix $\mathcal{S}(f+axg,g)$ with the same determinant: $\det\mathcal{S}(f+axg,g)\det\mathcal{S}(f,g)$. More generally, we see that as long as $k\le n-m$, then for any constant a, $\det\mathcal{S}(f+ax^kg,g)=\det\mathcal{S}(f,g)$. This easily implies the following very useful fact:

THEOREM 6. *Given the polynomials $f(x)$, $g(x)$ with real coefficients of degrees $n\ge m$ (respectively), then for any polynomial $h(x)$ of degree $\le n-m$, $det\,\mathcal{S}(f+gh,g)=det\,\mathcal{S}(f,g)$.*

Now consider the monic trinomial of the form $f(x)=x^n+ax+b$ where a, b are real numbers. Applying Theorem 3, we see that

$$\begin{aligned}
\Delta(f) &= (-1)^{n(n-1)/2}\mathcal{R}(f,f')\\
&= (-1)^{n(n-1)/2}\det\mathcal{S}(x^n+ax+b,nx^{n-1}+a)\\
&= (-1)^{n(n-1)/2}\det\mathcal{S}((a-a/n)x+b,nx^{n-1}+a)\quad\text{(Theorem 3)}\\
&= (-1)^{n(n-1)/2}(-1)^{(n-1)^2}n^{n-1}\mathcal{R}((a-a/n)x+b,nx^{n-1}+a)\\
&= (-1)^{n(n-1)/2}(-1)^{(n-1)^2}n^{n-1}(a-a/n)^{n-1}(n(-1)^{n-1}(b/(a-a/n))^{n-1}+a)\\
&\quad\text{(Corollary 3)}\\
&= (-1)^{n(n-1)/2}(-1)^{n-1}((-1)^{n-1}n^nb^{n-1}+a^n(n-1)^{n-1})\\
&= (-1)^{n(n-1)/2}(n^nb^{n-1}+(-1)^{n-1}(n-1)^{n-1}a^n).
\end{aligned}$$

EXAMPLE. Let $f(x)=\dfrac{3}{x}+x^3$, $x>0$ and find the minimum value of $f(x)$. While such a problem is typically the province of differential calculus, we can use a discriminant argument to obtain the solution.



We let m be the minimum value of $f(x)$ and note that the graph of $y = m$ must be tangent to the graph of $y = \dfrac{3}{x} + x^3$. This is equivalent to saying that the solution of $3 + x^4 = mx$ is a multiple root, forcing the discriminant of the quartic polynomial $q(x) = x^4 - mx + 3$ to be zero. That is to say, we need to find that value of m making the discriminant equal to 0. From the above, we have that

$$0 = \Delta(q) = 4^4 3^3 - 3^3 m^4 = 0 \Rightarrow m = 4.$$

In other words, the minimum value of $q(x)$ on the interval $(0, \infty)$ is 4. (The reader should check that the same result is obtained via differential calculus.)

Chapter 4

Abstract Algebra

While an oversimplification, **abstract algebra** grew out of an attempt to solve and otherwise understand polynomial equations (or systems of polynomial equations). A relative high point can be found in the early nineteenth century with E. Galois' proof that polynomial equations of degree at least 5 need not be solvable by the "usual" processes of addition, subtraction, multiplication, division, and extraction of roots as applied to the polynomial's coefficients. What's remarkable is not so much the result itself but rather the methods employed. This marked the beginning of a new enterprise, now called **group theory** which soon took on a life of itself, quite apart from playing a role in polynomial equations.

The language and level of abstraction in group theory quickly began to spread, leading to the somewhat larger discipline of **abstract algebra**. We'll attempt to give the serious student a meaningful introduction in this chapter.

4.1 Basics of Set Theory

In this section we shall consider some elementary concepts related to **sets** and their **elements**, assuming that at a certain level, the students have encountered the notions. In particular we wish to review (not necessarily in this order)

- Element **containment** ($\in$)
- **Containment** relationships between sets ($\subseteq$, $\supseteq$, $\subset$, (same as $\subsetneq$), $\supset$, (same as $\supsetneq$))

- Operations on subsets of a given set: **intersection** ($\cap$), **union**, ($\cup$), **difference** ($-$), and **symmetric difference** ($+$) of two subsets of a given set
- Set-theoretic constructions: **power set** (2^S), and **Cartesian product** ($S \times T$)
- **Mappings** (i.e., functions) between sets
- **Relations** and **equivalence relations** on sets

Looks scary, doesn't it? Don't worry, it's all very natural....

Before we launch into these topics, let's get really crazy for a moment. What we're going to talk about is **naive** set theory. As opposed to what, you might ask? Well, here's the point. When talking about sets, we typically use the language,

"the set of all ..."

Don't we often talk like this? Haven't you heard me say, "consider the set of all integers," or "the set of all real numbers"? Maybe I've even asked you to think about the "set of all differentiable functions defined on the whole real line." Surely none of this can possibly cause any difficulties! But what if we decide to consider something really huge, like the "set of all sets"? Despite the fact that this set is really big, it shouldn't be a problem, should it? The only immediately peculiar aspect of this set—let's call it $\mathcal{B}$ (for "big")—is that not only $\mathcal{B} \subseteq \mathcal{B}$ (which is true for all sets), but also that $\mathcal{B} \in \mathcal{B}$. Since the set $\{1\} \notin \{1\}$, we see that for a given set A, it may or may not happen that $A \in A$. This leads us to consider, as did Bertrand Russell, the set of all sets which don't contain themselves as an element; in symbols we would write this as

$$R \;=\; \{S \mid S \notin S\}.$$

This set R seems strange, but is it really a problem? Well, let's take a closer look, asking the question, is $R \in R$? By looking at the definition,

we see that $R \in R$ if any only if $R \notin R$! This is impossible! **This is a paradox**, often called Russell's paradox (or Russell's Antinomy).

Conclusion: Naive set theory leads to paradoxes! So what do we do? There are basically two choices: we could be much more careful and do **axiomatic set theory**, a highly formalized approach to set theory (I don't care for the theory, myself!) but one that is free of such paradoxes. A more sensible approach for us is simply to continue to engage in naive set theory, trying to avoid sets that seem unreasonably large and hope for the best!

4.1.1 Elementary relationships

When dealing with sets naively, we shall assume that the statement "x in an element of the set A" makes sense and shall symbolically denote this statment by writing $x \in A$. Thus, if $\mathbb{Z}$ denotes the set of integers, we can write such statements as $3 \in \mathbb{Z}$, $-11 \in \mathbb{Z}$, and so on. Likewise, π is not an integer so we'll express this by writing $\pi \notin \mathbb{Z}$.

In the vast majority of our considerations we shall be considering sets in a given "context," i.e., as **subsets** of a given set. Thus, when I speak of the set of integers, I am usually referring to a particular subset of the real numbers. The point here is that while we might not really know what a real number is (and therefore we don't really "understand" the set of real numbers), we probably have a better understanding of the particular subset consisting of integers (whole numbers). Anyway, if we denote by $\mathcal{R}$ the set of all real numbers and write $\mathbb{Z}$ for the subset of of integers, then we can say that

$$\mathbb{Z} \;=\; \{x \in \mathcal{R} \,|\, x \text{ is a whole number}\}.$$

Since $\mathbb{Z}$ is a subset of $\mathcal{R}$ we have the familiar notation $\mathbb{Z} \subseteq \mathcal{R}$; if we wish to emphasize that they're different sets (or that $\mathbb{Z}$ is **properly** contained in $\mathcal{R}$), we write $\mathbb{Z} \subset \mathcal{R}$ (some authors[1] write $\mathbb{Z} \subsetneq \mathcal{R}$). Likewise, if we let $\mathbb{C}$ be the set of all complex numbers, and consider also the set $\mathbb{Q}$ of all rational numbers, then we obviously have

[1] like me, but the former seems more customary in the high-school context.

$$\mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathcal{R} \subseteq \mathbb{C}.$$

As a more geometrical sort of example, let us consider the set $\mathcal{R}^3$ of all points in Cartesian 3-dimensional space. There are certain naturally defined subsets of $\mathcal{R}^3$, the **lines** and the **planes**. Thus, if Π is a plane in $\mathcal{R}^3$, and if L is a line contained in Π, then of course we may write either $L \subset \Pi \subset \mathcal{R}^3$ or $L \subseteq \Pi \subseteq \mathcal{R}^3$. Note, of course, that $\mathcal{R}^3$ has far more subsets that just the subsets of lines and planes!

One more example might be instructive here. First of all, if A is a finite set, we shall denote by $|A|$ the number of elements in A. We often call $|A|$ the **cardinality** or **order** of the set A. Now consider the finite set $S = \{1, 2, 3, \ldots, 8\}$ (and so $|S| = 8$) and ask how many subsets (including S and the empty set $\emptyset$) are contained in S. As you might remember, there are 2^8 such subsets, and this can be shown in at least two ways. The most direct way of seeing this is to form subsets of S by the following process:

1	2	3	4	5	6	7	8
yes or no	yes or no	yes or no	yes or no	yes or no	yes or no	yes or no	yes or no

where in the above table, a subset if formed by a sequence of *yes*'s or *no*'s according as to whether or not the corresponding element is in the subset. Therefore, the subset $\{3, 6, 7, 8\}$ would correspond to the sequence

$$(\text{no, no, yes, no, no, yes, yes, yes}).$$

This makes it already clear that since for each element there are two choices ("yes" or "no"), then there must be

$$2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2 = 2^8$$

possibilities in all.

Another way to count the subsets of the above set is to do this:

$$
\begin{aligned}
\text{Number of subsets} & \\
&= \text{number of subsets of size } 0 \\
&+ \text{number of subsets of size } 1 \\
&+ \text{number of subsets of size } 2 \\
&+ \text{number of subsets of size } 3 \\
&+ \text{number of subsets of size } 4 \\
&+ \text{number of subsets of size } 5 \\
&+ \text{number of subsets of size } 6 \\
&+ \text{number of subsets of size } 7 \\
&+ \text{number of subsets of size } 8 \\
&= \tbinom{8}{0} + \tbinom{8}{1} + \tbinom{8}{2} + \tbinom{8}{3} + \tbinom{8}{4} + \tbinom{8}{5} + \tbinom{8}{6} + \tbinom{8}{7} + \tbinom{8}{8} \\
&= \sum_{k=0}^{8} \binom{8}{k} = (1+1)^8 = 2^8,
\end{aligned}
$$

where we have applied the Binomial Theorem.

In general, if A is any set, we denote by 2^A the set of all subsets of A, often called the **power set** of A. (Many authors denote this set by $\mathcal{P}(A)$.) We'll have more to say about the power set later. At any rate, we showed above that if $S = \{1, 2, 3, \ldots, 8\}$, then $|2^S| = 2^8$. The obvious generalization is this:

Theorem. *Let A be a finite set with cardinality n. The 2^A has cardinality 2^n. Symbolically,*

$$\left|2^A\right| = 2^{|A|}.$$

Exercises

1. Let p be a prime number and define the following subset of the rational numbers $\mathbb{Q}$:

$$\mathbb{Q}_{(p)} = \left\{ \frac{r}{s} \in \mathbb{Q} \,\middle|\, \text{ the fraction } \frac{r}{s} \text{ is in lowest terms, and } p \text{ doesn't evenly divide } s \right\}.$$

Determine which of the following real numbers are in $\mathbb{Q}_{(2)}$:

$$\pi, \quad \frac{2}{3}, \quad \frac{10}{2}, \quad \cos(\pi/4), \quad 12, \quad \frac{3}{4}, \quad 12\pi, \quad \frac{\pi}{3}.$$

2. True or false: $\mathbb{Z} \subseteq \mathbb{Q}_{(p)}$ for any prime number p.

3. Consider the set $S = \{1, 2, 3, \ldots, 10\}$. Define the sets

$$\begin{aligned} A &= \{\text{subsets } T \subseteq S \mid |T| = 2\} \\ B &= \{\text{subsets } T \subseteq S \mid |T| = 2, \text{ and if } x, y \in T \text{ then } |x - y| \geq 2\} \end{aligned}$$

Compute $|A|$ and $|B|$.

4. Given the real number x, denote by $[x]$ the largest integer n not exceeding x. Therefore, we have, for example, that $[4.3] = 4$, $[\pi] = 3$, $[e] = 2$, $[-\pi] = -4$, and $\left[\frac{10}{3}\right] = 3$. Define the set A of integers by setting

$$A = \left\{ \left[\frac{1^2}{100}\right], \left[\frac{2^2}{100}\right], \left[\frac{3^2}{100}\right], \ldots, \left[\frac{99^2}{100}\right], \left[\frac{100^2}{100}\right] \right\}$$

and compute $|A|$.

4.1.2 Elementary operations on subsets of a given set

Let A and B be subsets of some bigger set U (sometimes called the **universal set**; note that U shall just determine a context for the ensuing constructions). We have the familiar **union** and **intersection**, respectively, of these subsets:

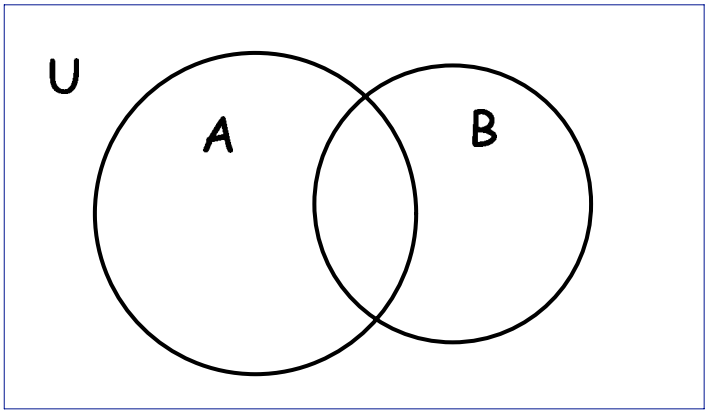
$$\begin{aligned} A \cup B &= \{u \in U \mid u \in A \text{ or } u \in B\}, \text{ and} \\ A \cap B &= \{u \in U \mid u \in A \text{ and } u \in B\}. \end{aligned}$$

I'm sure that you're reasonably comfortable with these notions. Two other important construction are the **difference** and **complement**, respectively:

$$A - B \;=\; \{u \in U \mid u \in A \text{ but } u \notin B\}, \text{ and}$$

$$A' \;=\; \{u \in U \mid u \notin A\} \;=\; U - A.$$

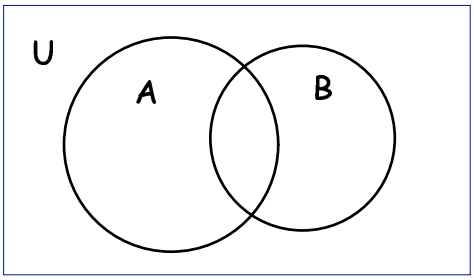
Relationships and operations regarding subsets are often symbolically represented through the familiar **Venn diagram**. For example, in the Venn diagram below, the student should have no difficulty in coloring in any one of the subsets $A \cup B$, $A \cap B$, $A - B$, $B - A$, A' (or any others that might come to mind!)

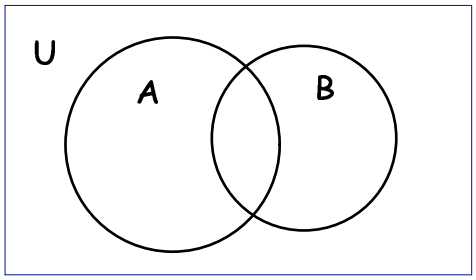


Venn diagrams can be useful in identifying properties of the above operations. One very typical example of such relationships and their Venn diagram proofs are the **De Morgan Laws**: *for subsets A and B of a universal set U, one has*

$$\boxed{(A \cup B)' = A' \cap B' \;\text{ and }\; (A \cap B)' = A' \cup B'}.$$

You can convince yourself of these facts by coloring in the Venn diagrams:





Actually, though, the De Morgan Laws are hardly surprising. If A represents "it will rain on Monday," and B represents "it will rain on Tuesday," then "it will not rain on Monday or Tuesday" is represented by $(A \cup B)'$, which is obviously the same as "it won't rain on Monday and it won't rain on Tuesday," represented mathematically by $A' \cap B'$.

A more formal proof might run along the following lines. In proving that for two sets $S = T$, it is often convenient to prove that $S \subseteq T$ and that $T \subseteq S$.

Theorem. *For subsets A and B of a given set U, $(A \cup B)' = A' \cap B'$.*

Proof. Let $x \in (A \cup B)'$. Then x is not in $A \cup B$, which means that x is not in A and that x is not in B, i.e., $x \in A' \cap B'$. This proves that $(A \cup B)' \subseteq A' \cap B'$. Conversely, if $x \in A' \cap B'$, then x is not in A and that x is not in B, and so x is not in $A \cup B$. But this says that $x \in (A \cup B)'$, proving that $(A \cup B)' \subseteq A' \cap B'$. It follows, therefore, that $(A \cup B)' = A' \cap B'$.

There are two other important results related to unions and intersections, both of which are somewhat less obvious than the De Morgan laws. Let's summarize these results as a theorem:

Theorem. *Let A, B, and C be subsets of some universal set U. Then we have two "distributive laws:"*

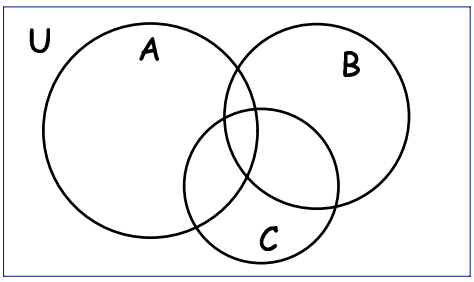
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C), \quad \textit{and} \quad A \cup (B \cap C) = (A \cup B) \cap (A \cup C).$$

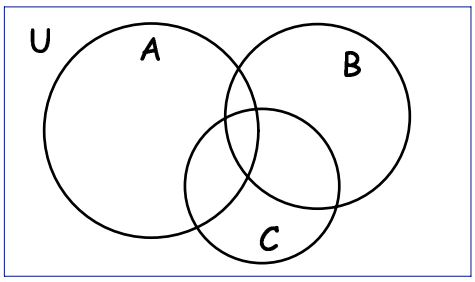
Proof. As you might expect the above can be easily demonstrated through Venn diagrams (see Exercise 1 below). Here, I'll give a formal proof of the first result (viz., that "intersection distributes over union"). Let $x \in A \cap (B \cup C)$ and so $x \in A$ and $x \in B \cup C$. From this we see that either $x \in A$ and $x \in B$ or that $x \in A$ and $x \in C$, which means, of course, that $x \in (A \cap B) \cup (A \cap C)$, proving that $A \cap (B \cup C) \subseteq (A \cap B) \cup (A \cap C)$. Conversely, if $x \in (A \cap B) \cup (A \cap C)$, then $x \in A \cap B$ or $x \in A \cap C$. In either case $x \in A$, but also $x \in B \cup C$, which means that $x \in A \cap (B \cup C)$, proving that $A \cap (B \cup C) \subseteq (A \cap B) \cup (A \cap C)$. It follows that $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$. The motivated student will have no difficulty in likewise providing a formal proof of the second distributive law.

EXERCISES

1. Give Venn diagram proofs of the distributive laws:

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C), \quad \text{and} \quad A \cup (B \cap C) = (A \cup B) \cap (A \cup C).$$





2. Show that if A, $B \subseteq U$, then $A - B = A \cap B'$.

3. Use a Venn diagram argument to show that if A, B, $C \subseteq U$, then

$$A - (B \cup C) = (A - B) \cap (A - C) \quad \text{and} \quad A - (B \cap C) = (A - B) \cup (A - C).$$

4. Show that if A, $B \subseteq U$, and if A and B are **finite** subsets, then $|A \cup B| = |A| + |B| - |A \cap B|$.

5. Show that if A, B, and $C \subseteq U$, and if A, B, and C are finite subsets, then

$$|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|.$$

6. Try to generalize Exercise 5 above.[2]

7. (Compare with Exercise 3 of Subsection 4.1.1) Consider the set $S = \{1, 2, 3, \ldots, 10\}$, and define the sets

$$T = \{\text{ordered pairs } (X, Y) \text{ of subsets } X, Y \subseteq S, \text{ with } |X|, |Y| = 2 \text{ and } X \cap Y = \emptyset\}$$
$$T' = \{\text{subsets } \{X, Y\} \subseteq 2^S \mid |X|, |Y| = 2 \text{ and } X \cap Y = \emptyset\}$$

Compute $|T|$ and $|T'|$.

8. In this problem the universal set is the real line $\mathcal{R}$. Find $A \cup B$, $A \cap B$, $A - B$, $B - A$, and $(A \cup B)'$, where $A =]-10, 5]$ and $B = [-4, \pi]$.

9. In this problem the universal set is the Cartesian plane $\mathcal{R}^2 = \{(x, y) \mid x, y \in \mathcal{R}\}$. Define the subsets

$$A = \{(x, y) \mid x^2 + y^2 < 1\} \quad \text{and} \quad B = \{(x, y) \mid y \geq x^2\}.$$

Sketch the following sets as subsets of $\mathcal{R}^2$: $A \cup B$, $A \cap B$, $A - B$, $B - A$, and $(A \cup B)'$.

10. Let A, $B \subseteq U$ and define the **symmetric difference** of A and B by setting

$$A + B = (A \cup B) - (A \cap B).$$

Using Venn diagram arguments, show the distributive laws

$$A + B = (A - B) \cup (B - A)$$

[2] This is the classical principle of **Inclusion-Exclusion**.

$$A \cap (B + C) \;=\; (A \cap B) + (A \cap C), \text{ where } A,\, B,\, C \subseteq U$$
$$A + (B \cap C) \;=\; (A + B) \cap (A + C), \text{ where } A,\, B,\, C \subseteq U.$$

11. Let p be a fixed prime and let $\mathbb{Q}_{(p)}$ be the set defined in Exercise 1 of Subsection 4.1.1. Interpret and prove the statement that

$$\bigcap_{\text{all primes } p} \mathbb{Q}_{(p)} \;=\; \mathbb{Z}.$$

12. Interpret and prove the statements

(i) $\displaystyle\bigcap_{n=1}^{\infty} \left([0, \tfrac{1}{n}]\right) \;=\; \{0\}$

(ii) $\displaystyle\bigcap_{n=1}^{\infty} \left(]0, \tfrac{1}{n}]\right) \;=\; \emptyset$

4.1.3 Elementary constructions—new sets from old

We have already encountered an elementary construction on a given set: that of the **power set**. That is, if S is a set, then 2^S is the set of all subsets of the set S. Furthermore, we saw in the theorem on page 189 that if S is a finite set containing n elements, then the power set 2^S contains 2^n elements (which motivates the notation in the first place!). Next, let A and B be sets. We form the **Cartesian product** $A \times B$ to be the set of all **ordered pairs** of elements (a, b) formed by elements of A and B, respectively. More formally,

$$A \times B \;=\; \{(a, b) \mid a \in A \text{ and } b \in B\}.$$

From the above, we see that we can regard the Cartesian plane $\mathcal{R}^2$ as the Cartesian product of the real line $\mathcal{R}$ with itself: $\mathcal{R}^2 = \mathcal{R} \times \mathcal{R}$. Similarly, Cartesian 3-space $\mathcal{R}^3$ is just $\mathcal{R} \times \mathcal{R} \times \mathcal{R}$.

Here are a couple of constructions to think about. Perhaps you can see how a right circular cylinder of height h and radius r can be regarded as $S \times [0, h]$, where S is a circle of radius h. Next, can you

see how the product $S \times S$ of two circles could be identified with the *torus* (the surface of a doughnut)?[3]

Finally, it should be obvious that if A and B are finite sets $|A \times B| = |A| \cdot |B|$.

EXERCISES

1. Let n be a positive integer and let $S = \{1, 2, \ldots, n\}$. Define the subset $T \subseteq S \times S$ by $T = \{(a, b) \in S \times S \mid |a - b| = 1\}$. Compute $|T|$ as a function of n.

2. Let n be a positive integer and let S be as above. Define the subset $Z \subseteq S \times S \times S$ by $Z = \{(a, b, c) \in S \times S \times S \mid a, b, c \text{ are all distinct}\}$. Compute $|Z|$ as a function of n.

3. Let X and Y be sets, and let $C, D \subseteq Y$. Prove that $X \times (C \cup D) = (X \times C) \cup (X \times D)$.

4. Let X and Y be sets, let $A, B \subseteq X$ and let $C, D \subseteq Y$. Is it always true that

$$(A \cup B) \times (C \cup D) \;=\; (A \times C) \cup (B \times D)?$$

5. Let T and T' be the sets defined in Exercise 7 of Subsection 4.1.2. Which of the following statements are true:

$$T \in S \times S, \quad T \subseteq S \times S, \quad T \in 2^S, \quad T \subseteq 2^S$$
$$T' \in S \times S, \quad T' \subseteq S \times S, \quad T' \in 2^S, \quad T' \subseteq 2^S$$

[3]Here's a parametrization of the torus which you might find interesting. Let R and r be positive real numbers with $r < R$. The following parametric equations describe a torus of outer radius $r + R$ and inner radius $R - r$:

$$x = (R + r\cos\phi)\cos\theta$$
$$y = (r + r\cos\phi)\sin\theta$$
$$z = r\sin\phi.$$

4.1.4 Mappings between sets

Let A and B be sets. A **mapping** from A to B is simply a function from A to B; we often express this by writing $f : A \to B$ or $A \xrightarrow{f} B$. Let's give some examples (some very familiar):

- $f : \mathcal{R} \to \mathcal{R}$ is given by $f(x) = x^2 - x + 1,\ x \in \mathcal{R}$
- $f : \mathcal{R} \to \mathbb{C}$ is given by $f(x) = (x - 1) + ix^2,\ x \in \mathcal{R}$
- Let $\mathbb{Z}^+ \subseteq \mathbb{Z}$ be the set of **positive** integers and define $g : \mathbb{Z}^2 \to \mathcal{R}$ by $g(m) = \cos(2\pi/n),\ n \in \mathbb{Z}^+$
- $h : \mathcal{R} \times \mathcal{R} \to \mathcal{R}$ is given by $f(x, y) = x - y,\ x,\ y \in \mathcal{R}$.
- $\gamma : \mathcal{R} \times \mathcal{R} \to \mathcal{R}$ is given by $\gamma(x, y) = x^2 + y^2$
- $q : \mathbb{Z} \to \mathbb{Z}$ is given by $q(n) = \frac{1}{2}(n^2 + n),\ n \in \mathbb{Z}$
- $\mu : \mathbb{Z}^+ \to \{-1, 0, 1\}$ is given by

$$\mu(n) = \begin{cases} 1 & \text{if } n \text{ is the product of an even number of distinct primes} \\ -1 & \text{if } n \text{ is the product of an odd number of distinct primes} \\ 0 & \text{if } n \text{ is not the product of distinct primes} \end{cases}$$

  Thus, for example, $\mu(1) = 0$. Also, $\mu(6) = 1$, as $6 = 2 \cdot 3$, the product of two distinct primes. Likewise, $\mu(5) = \mu(30) = -1$, and $\mu(18) = 0$.

- $h : \mathcal{R} \times \mathcal{R} \to \mathbb{C}$ is given by $h(x, y) = x + iy,\ x,\ y \in \mathcal{R}$
- $\sigma : \{1, 2, 3, 4, 5, 6\} \to \{1, 2, 3, 4, 5, 6\}$ is represented by

$$\sigma : \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ 2 & 5 & 3 & 4 & 1 & 6 \end{pmatrix}$$

If $f : A \to B$ is a mapping we call A the **domain** of f and call B the **codomain** of f. The **range** of f is the subset $\{f(a) \mid a \in A\} \subseteq B$.

Some definitions. Let A and B be sets and let $f : A \to B$. We say that

- f is **one-to-one** (or is **injective**) if whenever $x, y \in A,\ x \neq y$ then $f(x) \neq f(y)$. (This is equivalent with saying that $f(x) = f(y) \Rightarrow x = y$, where $x, y \in A$.)
- f is **onto** (or is **surjective**) if for any $z \in B$ there is an element $x \in A$ such that $f(x) = z$. Put differently, f is onto if the range of r is all of B.
- f is **bijective** if f is both one-to-one and onto.

The following definition is extremely useful. Let A and B be sets and let $f : A \to B$ be a mapping. Let $b \in B$; the **fibre** of f over b, written $f^{-1}(b)$ is the set

$$f^{-1}(b) = \{a \in A \mid f(a) = b\} \subseteq A.$$

Please do not confuse fibres with anything having to do with the inverse function f^{-1}, as this **might not exist**! Note that if $b \in B$ then the fibre over b might be the empty set. However, if we know that $f : A \to B$ is onto, then the fibre over each element of B is nonempty. If, in fact, for each $b \in B$ the fibre $f^{-1}(b)$ over b consists of a single element, then we are guaranteed that f is a bijection.

Finally, a mapping $f : A \to A$ from a set into itself is called a **permutation** if it is a bijection. It should be clear that if $|A| = n$, then there are $n!$ bijections on A.

EXERCISES

1. Let $f : \mathcal{R} \to \mathcal{R}$ be a **quadratic** function. Therefore, there are real constants a, b, $c \in \mathcal{R}$ with $a \neq 0$ such that $f(x) = ax^2 + bx + c$ for all $x \in \mathcal{R}$. Prove that f cannot be either injective or surjective.

2. Suppose that $f : \mathcal{R} \to \mathcal{R}$ is a cubic function such that $f'(x) \neq 0$ for all $x \in \mathcal{R}$. Give an intuitive argument (I'm not asking for a formal proof) that f must be bijective.

3. Define $f : \mathcal{R} \times \mathcal{R} \to \mathcal{R}$ by setting $f(x, y) = x - y$. Show that f is onto but is not one-to-one.

4. Define $f : \mathcal{R} \times \mathcal{R} \to \mathcal{R}$ by setting $f(x, y) = x^2 + y$. Show that f is onto but is not one-to-one.

5. Let $f : \mathbb{C} \to \mathbb{C}$ be a **quadratic** function. Therefore, there are complex constants a, b, $c \in \mathbb{C}$ with $a \neq 0$ such that $f(x) = ax^2 + bx + c$ for all $x \in \mathbb{C}$. Prove that f is onto but not one-to-one. (Compare with Exercise 1, above.)

6. For the mapping given in Exercise 3, above, show that the fibre over each point in $\mathcal{R}$ is a line in $\mathcal{R} \times \mathcal{R}$.

7. What are the fibres of the mapping in Exercise 4?

8. (A guided exercise) Let A be a set and let 2^A be its power set. Let's show that **there cannot exist any surjective function $f : A \to 2^A$**. A good way to proceed is to **argue by contradiction**, which means that we'll assume that, in fact, a surjective function exists and then reach a contradiction! So let's assume that $f : A \to 2^A$ is surjective. Note first that for any element $a \in A$, it may or may not happen that $a \in f(a)$ (this is important!). Now consider the following strange subset of A:

$$A_0 \;=\; \{a \in A \mid a \notin f(a)\} \in 2^A.$$

Is $A_0 = f(a_0)$ for some element $a_0 \in A$? Think about it! This contradiction has the same flavor as Russell's paradox!

4.1.5 Relations and equivalence relations

Let S be a set. A **relation** R on S is simply a subset of $S \times S$. Nothing more, nothing less. If $(x, y) \in R \subseteq S \times S$, then we typically write xRy and say that ***x* is related to *y***. A few examples might clarify this.

(i) Let R be the relation "$<$" on the set $\mathcal{R}$ of real numbers. Therefore, $R = \{(x, y) \in \mathcal{R} \times \mathcal{R} \mid x < y\}$.

(ii) Fix a positive integer m and recall that if $a \in \mathbb{Z}$ then $m|a$ means that a is a multiple of m. Now let R be the relation on the set $\mathbb{Z}$ of integers defined by

$$aRb \Leftrightarrow m|(a - b).$$

Note that we already met this relation in Section 2.1.3.

This relation is, as we have seen, customarily denoted "mod m" and read "congruence modulo m." Thus if $m = 7$, then we can say that $1 \equiv 15 \pmod 7$" where we read this as "1 is congruent to 15 modulo 7."

Note, in particular, that if $m = 7$ then the integers which are congruent modulo 7 to -1 are precisely those of the form $-1 + 7k$, $k = 0, \pm 1, \pm 2, \ldots$.

(iii) Let $S = \{1, 2, 3, 4, 5, 6\}$. We may express a relation R on S by specifying a matrix P containing 0s and 1s and where the rows and columns are labeled by the elements of S in the order 1, 2, 3, 4, 5, and 6 and where a "1" in row i and column j designates that iRj. More specifically, let's consider the relation defined as by the matrix:

$$\begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

In this example we see that sRs for all $s \in S$. You should have no trouble in writing down all the correct relational expressions xRy.

(iv) Here's one of my favorite examples. Let T_5 be the set of all 2-element subsets of $\{1, 2, 3, 4, 5\}$ and define the relation R on T_5 by stipulating that $A_1 R A_2 \Leftrightarrow A_1 \cap A_2 = \emptyset$. Can you compute $|R|$ in this example (see Exercise 3, below)? Can you reformulate this in terms of an appropriate graph, having T_5 as its set of vertices?

(v) Define the relation R on the real line $\mathcal{R}$ by stipulating that $xRy \Leftrightarrow x - y \in \mathbb{Z}$. What are the elements related to π?

Let R be a relation on a set S. We say that R is an **equivalence relation** if the following three properties hold:

R is **reflexive**: sRs for any $s \in S$;

R is **symmetric**: $s_1 R s_2 \Leftrightarrow s_2 R s_1$, $s_1, s_2 \in S$;

R is **transitive**: $s_1 R s_2$ and $s_2 R s_3 \Rightarrow s_1 R s_3$, $s_1, s_2, s_3 \in S$.

Of the five examples given above, the relations (ii), (iii), and (v) are equivalence relations. The relation given in (i) is neither reflexive (since $x < x$ is **false** for all real numbers) nor is it symmetric ($1 < 2$ but $2 \not< 1$). This relation is, however transitive (easy to check!). The analysis of the remaining cases are left to the exercises.

Example (iii) is a bit different from the others, which warrant a few extra words. Two (almost) obvious facts are the following: Since the matrix has all 1s down the diagonal, this already proves the reflexivity of the relation. Next, the matrix is symmetric which proves that the

relation is symmetric. How about the transitivity? This involves a more work, but a bit of thought reveals the following. If P denotes the above matrix, and if P^2 has nonzero entries in exactly the same places as P, then the relation is also transitive.

Let S be a set and let R be an equivalence relation on S. For any element $s \in S$ we denote by $[s]$ the set

$$[s] \;=\; \{s' \in S \mid sRs'\} \subseteq S,$$

and call this set the **equivalence class** in S containing $s \in S$. Note that if $s_1 R s_2$ then $[s_1] = [s_2]$ because s_1 and s_2 are equivalent to exactly the same elements of S.

Proposition. *Let R be an equivalence relation on the set S and let $[s]$ and $[s']$ be two equivalence classes in S. Then either $[s] = [s']$, in which case sRs' or $[s] \cap [s'] = \emptyset$, where then $s \not R s'$.*

Proof. Assume that $[s] \cap [s'] \neq \emptyset$, say that there is some element $t \in [s] \cap [s']$. Therefore sRt and $s'Rt$ which implies by symmetry that sRt and tRs'. Using transitivity, we see that sRs' which means that s, s' are equivalent to exactly the same elements of S. From this it follows that $[s] = [s']$. The only other possibility is that $[s] \cap [s'] = \emptyset$ in which case one obviously has $s \not R s'$.

As a result of the above proposition we see that an equivalence relation R on a set S partitions the set into disjoint equivalence classes. In light of this, let's take a look at a few examples.

(i) Consider the equivalence relation "$\equiv \pmod 7$" on the set $\mathbb{Z}$ of integers. We have the following decomposition of $\mathbb{Z}$ into exactly 7 equivalence classes:

$$[0] = \{\ldots, -14, -7, 0, 7, 14, \ldots\}$$
$$[1] = \{\ldots, -13, -6, 1, 8, 15, \ldots\}$$
$$[2] = \{\ldots, -12, -5, 2, 9, 16, \ldots\}$$
$$[3] = \{\ldots, -11, -4, 3, 10, 14, \ldots\}$$

$$[4] = \{\ldots, -10, -3, 4, 11, 14, \ldots\}$$
$$[5] = \{\ldots, -9, -2, 5, 7, 12, \ldots\}$$
$$[6] = \{\ldots, -8, -1, 6, 7, 13, \ldots\}$$

(ii) Let R be the relation on $\mathcal{R}$ given by $xRy \Longrightarrow x - y \in \mathbb{Q}$. This is easily shown to be an equivalence relation, as follows. First xRx as $x - x = 0 \in \mathbb{Q}$. Next, if xRy, then $x - y \in \mathbb{Q}$ and so $y - x = -(x - y) \in \mathbb{Q}$, i.e., yRx. Finally, assume that xRy and that yRz. Then $x-y$, $y-z \in \mathbb{Q}$ and so $x-z = (x-y)+(y-z) \in \mathbb{Q}$ and so xRz. Note that the equivalence class containing the real number x is $\{x + r \mid r \in \mathbb{Q}\}$.

(iii) Define the function $f : \mathcal{R}^2 \to \mathcal{R}$ by setting $f(x, y) = x-y$. Define an equivalence relation on $\mathcal{R}^2$ by stipulating that $(x_1, y_1)R(x_2, y_2) \Leftrightarrow f(x_1, y_1) = f(x_2, y_2)$. Note that this is the same as saying that $x_1 - y_1 = x_2 - y_2$. Thus, the equivalence classes are nothing more than the fibres of the mapping f. We can visualize these equivalence classes by noting that the above condition can be expressed as $\dfrac{y_2 - y_1}{x_2 - x_1} = 1$, which says that the equivalence classes are precisely the various lines of slope 1 in the Cartesian plane $\mathcal{R}^2$.

One final definition is appropriate here. Let S be a set and let R be an equivalence relation on S. The **quotient set** of S by R is the set of equivalence classes in S. In symbols, this is

$$S/R \;=\; \{[a] \mid a \in S\}.$$

We shall conclude this subsection with a particularly important quotient set. Let $n \in \mathbb{Z}^+$ and let R be the relation "$\equiv \pmod n$." One usually writes $\mathbb{Z}_n$ for the corresponding quotient set. That is,

$$\mathbb{Z}_n \;=\; \{[m] \mid m \in \mathbb{Z}\} \;=\; \{[0], [1], [2], \ldots, [n-1]\}.^4$$

[4] It's important that the IB examination authors do not use the brackets in writing the elements of $\mathbb{Z}_n$; they simply write $\mathbb{Z}_n = \{0, 1, 2, \ldots, n-1\}$. While logically incorrect, this really shouldn't cause too much confusion.

EXERCISES

1. Let $S = \{1, 2, 3, 4\}$. How many relations are there on S?

2. Let $m \in Z^+$ and show that "$\equiv \pmod{m}$" is an equivalence relation on $\mathbb{Z}$. How many distinct equivalence classes mod m are there in $\mathbb{Z}$?

3. Let T_5 be the set of all 2-element subsets of $\{1, 2, 3, 4, 5\}$ and say define the relation R on T_5 by stipulating that $A_1 R A_2 \Leftrightarrow A_1 \cap A_2 = \emptyset$. Compute $|R|$. Which of the three properties of an equivalence relation does R satisfy?

4. Let $f : S \to T$ be a mapping and define a relation on S by stipulating that $sRs' \Leftrightarrow f(s) = f(s')$. (Note that this is says that $sRs' \Leftrightarrow s$ and s' are in the same fibre of f.) Show that R is an equivalence relation.

5. Define the following relation on the Cartesian 3-space $\mathcal{R}^3$: $PRQ \Leftrightarrow$ P and Q are the same distance from the origin. Prove that R is an equivalence relation on $\mathcal{R}^3$ and determine the equivalence classes in $\mathcal{R}^3$.

6. Suppose that we try to define a function $f : \mathbb{Z}_4 \to \mathbb{Z}$ by setting $f([n]) = n - 2$. What's wrong with this definition?

7. Suppose that we try to define a function $g : \mathbb{Z}_4 \to \{\pm 1, \pm i\}$ by setting $g([n]) = i^n$. Does this function suffer the same difficulty as that in Exercise 6?

8. Suppose that we try to define function $\tau : \mathbb{Z} \to \mathbb{Z}_4$ by setting $\tau(n) = [n - 2]$. Does this function suffer the same difficulty as that in Exercise 6? What's going on here?

9. Let R be the relation on the real line given by $xRy \Leftrightarrow x - y \in \mathbb{Z}$, and denote by the $\mathcal{R}/R$ the corresponding quotient set.[5] Suppose that we try to define $p : \mathcal{R}/R \to \mathbb{C}$ by setting $p([x]) = \cos 2\pi x + i \sin 2\pi x$. Does this definition make sense?

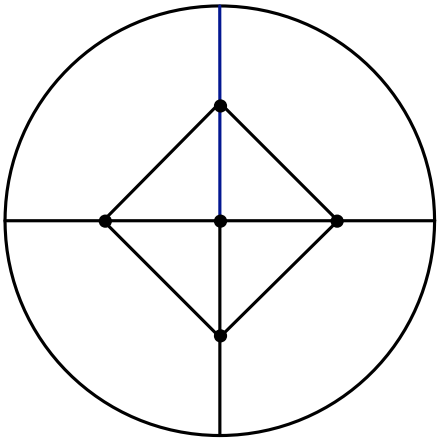
[5]Most authors denote this quotient set by $\mathcal{R}/\mathbb{Z}$.

10. We saw on page 141 that the complete graph K_5 cannot be planar, i.e., cannot be drawn in the plane. Let's see if we can draw it elsewhere. Start by letting

$$C \;=\; \{(x, y) \in \mathcal{R}^2 \,|\, x^2 + y^2 \leq 1\};$$

therefore, C is just the "disk" in the plane with radius 1. We define an equivalence relation R on C by specifying that a point (x, y) on the boundary of C (so $x^2 + y^2 = 1$) is equivalent with its "antipodal" point $(-x, -y)$. Points on the interior of C are equivalent only with themselves. We call the quotient set C/R the **real projective plane**, often written $\mathcal{R}\mathrm{P}^2$. (Recall that C/R is just the set of equivalence classes.)

Explain how the drawing to the right can be interpreted as a drawing of K_5 in the real projective plane. Also, compute the Euler characteristic $v-e+f$ for this drawing.



11. Here's another construction of $\mathcal{R}\mathrm{P}^2$, the real projective plane; see Exercise 10, above. Namely, take the unit sphere $S^2 \subseteq \mathcal{R}^3$, defined by $S^2 = \{(x, y, z) \in \{\mathcal{R}^3 \mid x^2 + y^2 + z^2 = 1\}$. We define a "geometry" on S^2 by defining **points** to be the usual points on the sphere and defining **lines** to be "great circles," i.e., circles on the sphere which form the shortest path between any two distinct points on such a circle. Notice, therefore, that the equator on the earth (approximately a sphere) is a great circle; so are the latitude lines. With this definition of lines and points we see that Euclid's parallel postulate[6] is violated as distinct parallel lines simply don't exist: any pair of distinct lines must meet in exactly two points.

[6]viz., that through any line ℓ_1 and any point P not on ℓ_1 there is a unique line ℓ_2 through the point P and not intersecting ℓ_1.

Form an equivalence relation R on S^2 by declaring any point on the sphere to be equivalent with its "antipode." (Thus on the earth, the north and south poles would be equivalent.) The quotient set S^2/R is often called the **real projective plane** and denoted $\mathcal{R}\mathrm{P}^2$.

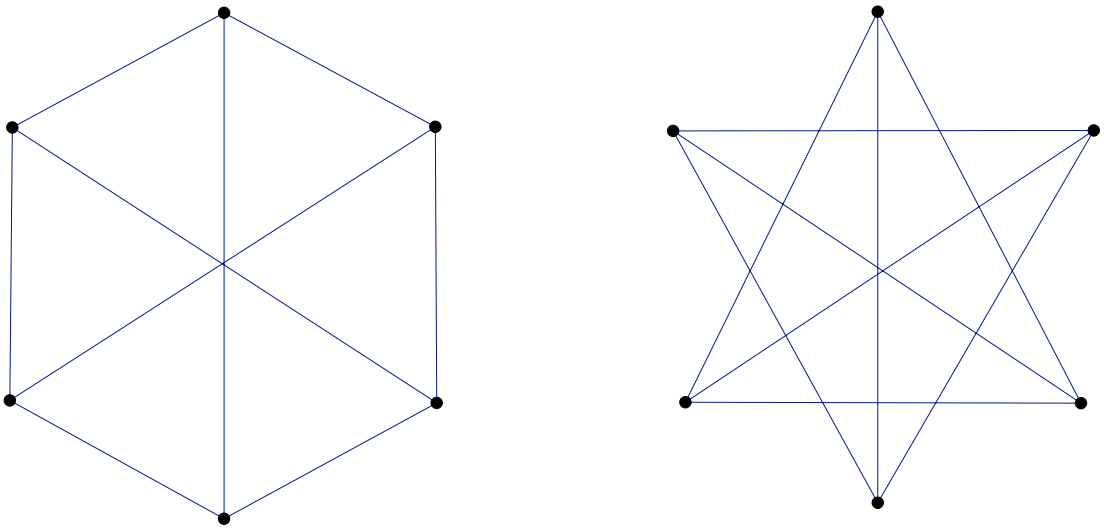
(a) Give at least a heuristic argument that the constructions of $\mathcal{R}\mathrm{P}^2$ given in this and Exercise 10 are equivalent.

(b) Show that on $\mathcal{R}\mathbf{P}^2$ that any pair of distinct points determine a unique line and that any pair of distinct lines intersect in a unique point.[7]

4.2 Basics of Group Theory

4.2.1 Motivation—graph automorphisms

We shall start this discussion with one of my favorite questions, namely which of the following graphs is more "symmetrical?"

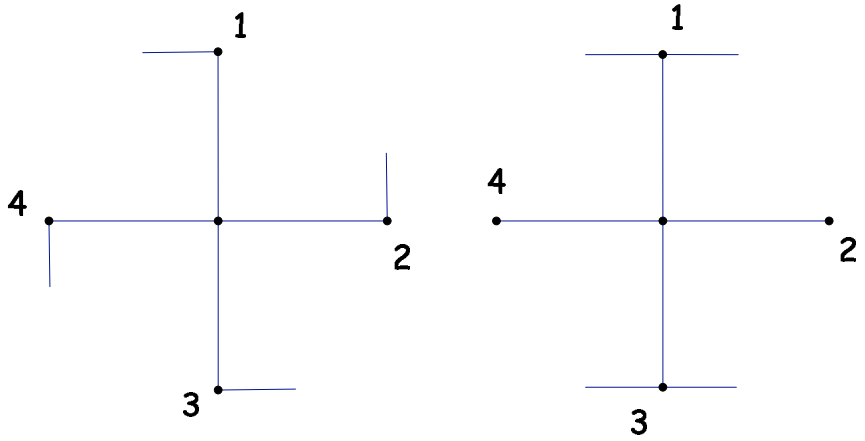


While this question might not quite make sense at the outset, it is my intention to have the reader rely mostly on intuition. Incidently,

[7]This says that the real projective plane has a "point-line duality" not enjoyed by the usual Euclidean plane.

I have asked this question many times and to many people—some mathematicians—and often, if not usually, I get the wrong intuitive response! Without pursuing the details any further, suffice it to say for now that group theory is the "algebraitization" of symmetry. Put less obtusely, groups give us a way of "quantifying" symmetry: the larger the group (which is something we can often compute!) the greater the symmetry. This is hardly a novel view of group theory. Indeed the prominent mathematician of the late 19-th and early 20-th century Felix Klein regarded all of geometry as nothing more than the study of properties invariant under groups.

Apart from quantifying symmetry, groups can give us a more explicit way to separate types of symmetry. As we'll see shortly, the two geometrical figures below both have four-fold symmetry (that is, they have groups of order 4), but the nature of the symmetry is different (the groups are not isomorphic).



Anyway, let's return briefly to the question raised above, namely that of the relative symmetry of the two diagrams above. Given a graph G we now consider the set of all permutations σ of the set of vertices of G such that

vertices a and b form an edge of $G \Leftrightarrow \sigma(a)$ and $\sigma(b)$ form an edge of G.

A permutation satisfying the above is called an **automorphism** of the graph G, and the set of all such automorphisms is often denoted

$\mathrm{Aut}(G)$. The most important facts related to graph automorphisms is the following:

Proposition. *The composition of two graph automorphisms is a graph automorphism. Also the inverse of a graph automorphism is a graph automorphism.*

Proof. This is quite simple. Let σ and τ be two graph automorphisms, and let v and w be vertices of the graph. Then $\sigma \circ \tau(v)$ and $\sigma \circ \tau(w)$ form an edge $\Leftrightarrow$ $\tau(v)$ and $\tau(w)$ form an edge $\Leftrightarrow$ v and w form an edge. Next, let σ be a graph automorphism. Since σ is a permutation, it is bijective and so the inverse function σ^{-1} exists. Thus we need to show that vertices v and w form an edge $\Leftrightarrow$ $\sigma^{-1}(v)$ and $\sigma^{-1}(w)$ forms an edge. If v and w form an edge, this says that $\sigma(\sigma^{-1}(v))$, $\sigma(\sigma^{-1}(w))$ form an edge. But since σ is an automorphism, we see that $\sigma^{-1}(v)$ and $\sigma^{-1}(w)$ form and edge. In other words,

$$v \text{ and } w \text{ form an edge } \Rightarrow \sigma^{-1}(v) \text{ and } \sigma^{-1}(w) \text{ form an edge.}$$

Conversely, assume that $\sigma^{-1}(v)$ and $\sigma^{-1}(w)$ form an edge. Then since σ is an automorphism, we may apply σ to conclude that the vertices $\sigma(\sigma^{-1}(v))$, $\sigma(\sigma^{-1}(w))$ form an edge. But this says that v and w form an edge, i.e.,

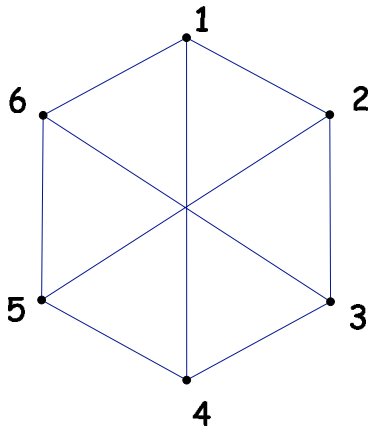
$$\sigma^{-1}(v) \text{ and } \sigma^{-1}(w) \text{ form an edge} \Rightarrow v \text{ and } w \text{ form an edge,}$$

which proves that σ^{-1} is a graph automorphism.

The above fact will turn out to be hugely important!

EXERCISES

1. Consider the graph shown below.



(i) Give an automorphism σ which takes vertex 1 to vertex 2 by completing the following:

$$\sigma : \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ 2 & & & & & \end{pmatrix}$$

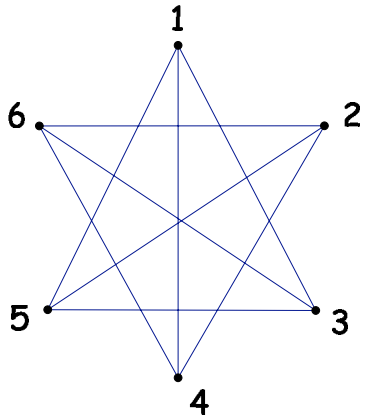
Compute the inverse of this automorphism.

(ii) Give an automorphism τ which maps vertex 1 to 3 and fixes vertex 5 (that is $\tau(5) = 5$).

$$\tau : \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ 3 & & & & 5 & \end{pmatrix}$$

Compute the inverse of this automorphism.

2. Consider the graph shown below.



(i) Give an automorphism σ which takes vertex 1 to vertex 2 by completing the following:

$$\sigma : \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ 2 & & & & & \end{pmatrix}$$

Compute the inverse of this automorphism.

(ii) Give an automorphism τ which maps vertex 1 to 3 and fixes vertex 5 (that is $\tau(5) = 5$).

$$\tau : \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ 3 & & & & 5 & \end{pmatrix}$$

Compute the inverse of this automorphism.

4.2.2 Abstract algebra—the concept of a binary operation

Intuitively, a **binary operation** on a set S is a rule for "multiplying" elements of S together to form new elements.

Definition of Binary Operation on a Set. *A* **binary operation** *on a non-empty set S is a mapping $* : S \times S \to S$.*

No more, no less! We usually write $s * s'$ in place of the more formal $*(s, s')$.

We have a wealth of examples available; we'll review just a few of them here.

- The familiar operations $+$ and $\cdot$ are binary operations on our favorite number systems: $\mathbb{Z}$, $\mathbb{Q}$, $\mathcal{R}$, $\mathbb{C}$.
- Note that if S is the set of irrational numbers then neither $+$ nor $\cdot$ defines a binary operation on S. (Why not?)
- Note that subtraction $-$ defines a binary operation on $\mathcal{R}$.
- Let $\mathrm{Mat}_n(\mathcal{R})$ denote the $n \times n$ matrices with real coefficients. Then both $+$ (matrix addition) and $\cdot$ (matrix multiplication) define binary operations on $\mathrm{Mat}_n(\mathcal{R})$.
- Let S be any set and let $\mathrm{F}(S) = \{\text{functions } : S \to S\}$. Then function composition $\circ$ defines a binary operation on $\mathrm{F}(S)$. (This is a particularly important example.)
- Let $\mathrm{Vect}_3(\mathcal{R})$ denote the vectors in 3-space. Then the **vector cross product** $\times$ is a binary operation on $\mathrm{Vect}_3(\mathcal{R})$. Note that the scalar product $\cdot$ does not define a binary operation on $\mathrm{Vect}_3(\mathcal{R})$.
- Let A be a set and let 2^A be its power set. The operations $\cap$, $\cup$, and $+$ (symmetric difference) are all important binary operations on 2^A.
- Let S be a set and let $\mathrm{Sym}(S)$ be the set of all permutations on S. Then function composition $\circ$ defines a binary operation on $\mathrm{Sym}(S)$. We really should prove this. Thus let σ, $\tau : S \to S$ be permutations; thus they are one-to-one and onto. We need to show that $\sigma \circ \tau : S \to S$ is also one-to-one and onto.

  $\underline{\sigma \circ \tau \text{ is one-to-one:}}$ Assume that s, $s' \in S$ and that $\sigma \circ \tau(s) = \sigma \circ \tau(s')$. Since σ is one-to-one, we conclude that $\tau(s) = \tau(s')$.

Since τ is one-to-one, we conclude that $s = s'$, which proves that $\sigma \circ \tau$ is also one-to-one.

$\sigma \circ \tau$ is onto: We need to prove that for any $s \in S$ there exists some $s' \in S$ such that $\sigma \circ \tau(s') = s$. However, since σ is onto, there must exist some element $s'' \in S$ such that $\sigma(s'') = s$. But since τ is onto there exists some element $s' \in S$ such that $\tau(s') = s''$. Therefore, it follows that $\sigma \circ \tau(s') = \sigma(\tau(s')) = \sigma(s'') = s$, proving that $\sigma \circ \tau$ is onto.

Before looking further for examples, I'd like to amplify the issue of "closure," as it will given many additional examples of binary operations.

Definition of Closure. Let S be a set, let $*$ be a binary operation on S, and let $\emptyset \neq T \subseteq S$. We say that T is **closed** under the binary operation $*$ if $t * t' \in T$ whenever $t, t' \in T$. In this case it is then follows that $*$ also defines a binary operation on T. Where the above IB remark is misleading is that we don't speak of a binary operation as being closed, we speak of a subset being closed under the given binary operation!

More examples ...

- Let $\mathcal{R}$ be the real numbers. Then $\mathbb{Z}$ and $\mathbb{Q}$ are both closed under both addition and multiplication.
- Note that the negative real numbers are not closed under multiplication.
- Let $\mathbb{Z}[\sqrt{5}] = \{a + b\sqrt{5} \mid a, b \in \mathbb{Z}\}$. Then $\mathbb{Z}[\sqrt{5}]$ is easily checked to be closed under both addition $+$ and multiplicaton $\cdot$ of complex numbers. (Addition is easy. For multiplication, note that if $a, b, c, d \in \mathbb{Z}$, then

$$(a + b\sqrt{5}) \cdot (c + d\sqrt{5}) \;=\; (ac + 5bd) + (ad + bc)\sqrt{5}.$$

Note that the above example depends heavily on the fact that $\mathbb{Z}$ is closed under both addition and multiplication.)

- Let $\mathrm{GL}_n(\mathcal{R}) \subseteq \mathrm{Mat}_n(\mathcal{R})$ denote the matrices of determinant $\neq 0$, and let $\mathrm{GL}_n^+(\mathcal{R}) \subseteq \mathrm{Mat}_n(\mathcal{R})$ denote the matrices of positive determinant. Then both of these sets are closed under multiplication; neither of these sets are closed under addition.
- The subset $\{\mathbf{0}, \pm\boldsymbol{i}, \pm\boldsymbol{j}, \pm\boldsymbol{k}\} \subseteq \mathrm{Vect}_3(\mathcal{R})$ is closed under vector cross product $\times$. The subset $\{\mathbf{0}, \boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}\} \subseteq \mathrm{Vect}_3(\mathcal{R})$ is not. (Why not?)
- The subset $\{-1, 0, 1\} \subseteq \mathbb{Z}$ is closed under multiplication but not under addition.
- Let X be a set and let $\mathrm{Sym}(X)$ denote the set of permutations. Fix an element $x \in X$ and let $\mathrm{Sym}_x(X) \subseteq \mathrm{Sym}(S)$ be the subset of all permutations which fix the element s. That is to say,

$$\mathrm{Sym}_x(X) \;=\; \{\sigma \in \mathrm{Sym}(X) \mid \sigma(x) = x\}.$$

Then $\mathrm{Sym}_x(X)$ is closed under function composition $\circ$ (Exercise 5).

We have two more extremely important binary operations, namely addition and subtraction on $\mathbb{Z}_n$, the integers modulo n. These operations are defined by setting

$$[a] + [b] \;=\; [a+b], \text{ and } \; [a] \cdot [b] \;=\; [a \cdot b], \quad a, b \in \mathbb{Z}.^8$$

We shall sometimes drop the $[\cdot]$ notation; as long as the context is clear, this shouldn't cause any confusion.

[8] A somewhat subtle issue related to this "definition" is whether it makes sense. The problem is that the same equivalence class can have many names: for example if we are considering congruence modulo 5, we have $[3] = [8] = [-2]$, and so on. Likewise $[4] = [-1] = [14]$. Note that $[3]+[4] = [7] = [2]$. Since $[3] = [-2]$ and since $[4] = [14]$, adding $[-2]$ and $[14]$ should give the same result. But they do: $[-2]+[14] = [12] = [2]$. Here how a proof that this definition of addition really makes sense (i.e., that it is **well defined**) would run. Let $[a] = [a']$ and $[b] = [b']$. Then $a' = a + 5k$ for some integer k and $b' = b + 5l$ for some integer l. Therefore $[a'] + [b'] = [a + 5k] + [b + 5l] = [a + b + 5k + 5l] = [a + b] = [a] + [b]$. Similar comments show that multiplication likewise makes sense. Finally this generalizes immediately to $\mathbb{Z}_n$, for any positive integer n.

We display addition and multiplication on the integers modulo 5 in the following obvious tables:

+	0	1	2	3	4
0	0	1	2	3	4
1	1	2	3	4	0
2	2	3	4	0	1
3	3	4	0	1	2
4	4	0	1	2	3

×	0	1	2	3	4
0	0	0	0	0	0
1	0	1	2	3	4
2	0	2	4	1	3
3	0	3	1	4	2
4	0	4	3	2	1

EXERCISES

1. Denote by $2\mathbb{Z} \subseteq \mathbb{Z}$ the even integers. Is $2\mathbb{Z}$ closed under addition? Under multiplication?

2. Is the set of **odd** integers closed under either addition or multiplication?

3. On the set $\mathbb{Z}$ of integers define the binary operation $*$ by setting $x * y = x + 2y \in \mathbb{Z}$. Is the set of even integers closed under $*$? Is the set of odd integers closed under $*$?

4. Let $U_2(\mathcal{R}) \subseteq \mathrm{Mat}_2(\mathcal{R})$ be defined by setting

$$U_2(\mathcal{R}) \;=\; \left\{ \left(\begin{array}{cc} 1 & x \\ 0 & 1 \end{array} \right) \;\middle|\; x \in \mathcal{R} \right\}.$$

   Is $U_2(\mathcal{R})$ closed under matrix addition? Under matrix multiplication?

5. Let X be a set and let $\mathrm{Sym}(X)$ be the set of permutations of X. Fix an element $x \in X$ and show that $\mathrm{Sym}_x(X)$ is closed under function composition "$\circ$."

6. Let A be a set and let $\mathcal{A} \subseteq 2^A$ be the subset of the power set consisting of all finite subsets of **even** cardinality. Show that if $|A| \geq 3$, then $\mathcal{A}$ is not closed under either $\cap$ or $\cup$ but it **is** closed under $+$. (Why do we need to assume that $|A| \geq 3$?)

7. Are the non-zero elements $\{1, 2, 3, 4, 5\}$ in $\mathbb{Z}_6$ closed under multiplication?

8. Are the non-zero elements of $\mathbb{Z}_p$, where p is a prime number, closed under multiplication?

9. For any positive integer n, set $\mathbb{N}_n = \{1, 2, \ldots, n\}$, and let $\mathcal{P}(\mathbb{N}_n)$ be the power set of $\mathbb{N}_n$ (see page 189). Show that for any integer N with $0 \leq N \leq 2^n$ there exists subsets $\mathcal{A}$, $\mathcal{B} \subseteq \mathcal{P}(\mathbb{N}_n)$ such that

   (a) $|\mathcal{A}| = |\mathcal{B}| = \mathcal{N}$,

   (b) $\mathcal{A}$ is closed under $\cap$, and

   (c) $\mathcal{B}$ is closed under $\cup$.

   (Hint: Use induction, together with the De Morgan laws.)

4.2.3 Properties of binary operations

Ordinary addition and multiplication enjoy very desirable properties, most notably, associativity and commutativity. Matrix multiplication is also associative (though proving this takes a little work), but not commutative. The vector cross product of vectors in 3-space is neither associative nor is it commutative. (The cross product is "anticommutative" in the sense that for vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, $\boldsymbol{u} \times \boldsymbol{v} = -\boldsymbol{v} \times \boldsymbol{u}$. The nonassociativity is called for in Exercise 2, below.) This motivates the following general definition: let S be a set with a binary operation $*$. We recall that

$*$ is **associative** if $s_1 * (s_2 * s_3) = (s_1 * s_2) * s_3$, for all s_1, s_2, $s_3 \in S$;

$*$ is **commutative** if $s_1 * s_2 = s_2 * s_1$, for all s_1, $s_2 \in S$.

Next, we say that e is an **identity** with respect to the binary operation $*$ if

$$e * s = s * e = s \quad \text{for all} \quad s \in S.$$

If the binary operation has an identity e, then this identity is **unique**. Indeed, if e' were another identity, then we would have

$$e \underbrace{=}_{\substack{\text{because } e' \\ \text{is an identity}}} e * e' \underbrace{=}_{\substack{\text{because } e \\ \text{is an identity}}} e'.$$

(Cute, huh?)

Finally, assume that the binary operation $*$ is associative and has an identity element e. The element $s' \in S$ is said to be an **inverse** of $s \in S$ relative to $*$ if $s' * s = s * s' = e$. Note that if s has an inverse, then **this inverse is unique**. Indeed, suppose that s' and s'' are both inverses of s. Watch this:

$$s' = s' * e = \underbrace{s' * (s * s'') = (s' * s) * s''}_{\text{note how associativity is used}} = e * s'' = s''.$$

EXERCISES

1. In each case below, a binary operation $*$ is given on the set $\mathbb{Z}$ of integers. Determine whether the operation is associative, commutative, and whether an identity exists.

   (a) $x * y = x + xy$

   (b) $x * y = x$

   (c) $x * y = 4x + 5y$

   (d) $x * y = x + xy + y$

   (e) $x * y = x^2 - y^2$

2. Give an example of three vectors $\boldsymbol{u}$, $\boldsymbol{v}$, and $\boldsymbol{w}$ in 3-space such that $\boldsymbol{u} \times (\boldsymbol{v} \times \boldsymbol{w}) \neq (\boldsymbol{u} \times \boldsymbol{v}) \times \boldsymbol{w}$.

3. Let A be a set and let 2^A be it power set. Determine whether the operations $\cap$, $\cup$, and $+$ are associative, commutative, and whether an identity element exists for the operation.

4. Let A be a nonempty set. For any non-empty set $B \subseteq A$ find the inverse of B with respect to symmetic difference $+$.

5. Let $\mathrm{Mat}_n(\mathcal{R})$ be the $n \times n$ matrices with real coefficients and define the binary operation $*$ by setting

$$\boldsymbol{A} * \boldsymbol{B} = \boldsymbol{A}\boldsymbol{B} - \boldsymbol{B}\boldsymbol{A},$$

   where $\boldsymbol{A}$, $\boldsymbol{B} \in \mathrm{Mat}_n(\mathcal{R})$. Is $*$ associative? Commutative? Is there an identity?

6. Let S be a set and let $\mathrm{F}(S)$ be the set of all functions $f : S \to S$. Is composition "$\circ$" associative? Commutative? Is there an identity?

7. Let S be a set and consider the set $\mathrm{F}(S, \mathcal{R})$ of all real-valued functions $f : S \to \mathcal{R}$. Define addition $+$ and multiplication "$\cdot$" on $\mathrm{F}(S, \mathcal{R})$ by the rules

$$(f+g)(s) \;=\; f(s)+g(s), \quad (f \cdot g)(s) \;=\; f(s) \cdot g(s), \quad s \in \mathcal{R}.$$

   Are these operations associative? Commutative? What about identities? What about inverses?

4.2.4 The concept of a group

Let $(G, *)$ be a set together with a binary operation. We say that $(G, *)$ is a **group** if the following three properties hold:

$*$ is associative: that is $g_1 * (g_2 * g_3) = (g_1 * g_2) * g_3$ for all $g_1,\ g_2,\ g_3 \in G$;

***G* has an identity:** that is, there exists an element $e \in G$ such that $e * g = g * e = e$, for all $g \in G$;

Existence of inverses: that is, for every $g \in G$, there exists an element $g' \in G$ with the property that $g' * g = g * g' = e$.

We have already noted on page 216 that the identity element and inverses are unique. This says that in denoting the inverse of an element $g \in G$ we may use, for example, the notation g^{-1} to denote this inverse, knowing that we are unambiguously referring to a unique element. However, inverses (and identities) aren't always denoted in this way. If we use the symbol $+$ for our binary operation, it's more customary to write "0" for the identity and to write $-a$ for the inverse of the element a. Finally it's worth mentioning that in certain contexts, the binary operation is simply denoted by "juxtaposition," writing, for example xy in place of $x * y$. This happens, for instance, in denoting multiplication of complex numbers, polynomials, matrices, and is even used to denote the binary operation in an abstract group when no confusion is likely to result.

We shall now survey some very important examples of groups.

1. (The **symmetric group**) Let X be a set and let $(\mathrm{Sym}(X), \circ)$ be the set of all bijections on X, with function composition as the binary operation. At the risk of being redundant, we shall carefully show that $(\mathrm{Sym}(X), \circ)$ is a group.

   $\circ$ is associative: let σ_1, σ_2, $\sigma_3 \in \mathrm{Sym}(X)$, and let $x \in X$. Then to show that $\sigma_1 \circ (\sigma_2 \circ \sigma_3) = (\sigma_1 \circ \sigma_2) \circ \sigma_3$ we need to show that they are the same permutations on X, i.e., we must show that for all $x \in X$, $\sigma_1 \circ (\sigma_2 \circ \sigma_3)(x) = (\sigma_1 \circ \sigma_2) \circ \sigma_3(x)$. But

   $$\sigma_1 \circ (\sigma_2 \circ \sigma_3)(x) = \sigma_1((\sigma_2 \circ \sigma_3)(x)) = \sigma_1(\sigma_2(\sigma_3(x))),$$

   whereas

   $$(\sigma_1 \circ \sigma_2) \circ \sigma_3(x) = (\sigma_1 \circ \sigma_2)(\sigma_3(x)) = \sigma_1(\sigma_2(\sigma_3(x))),$$

   which is the same thing! Thus we have proved that $\circ$ is associative.

   Existence of identity: Let $e : X \to X$ be the function $e(x) = x$, for all $x \in X$. Then clearly e is a permutation, i.e., $e \in \mathrm{Sym}(X)$.

Furthermore, for all $\sigma \in \mathrm{Sym}(X)$, and for all $x \in X$, we have $e \circ \sigma(x) = e(\sigma(x)) = \sigma(x)$, and $\sigma \circ e(x) = \sigma(e(x)) = \sigma(x)$, which proves that $e \circ \sigma = \sigma \circ e = \sigma$.

Existence of inverses: Let $\sigma \in \mathrm{Sym}(X)$ and let $\sigma^{-1} : X \to X$ denote its inverse function. Therefore $\sigma^{-1}(x) = y$ means precisely that $\sigma(y) = x$ from which it follows that σ^{-1} is a permutation (i.e., $\sigma^{-1} \in \mathrm{Sym}(X)$) and $(\sigma^{-1} \circ \sigma)(x) = x = (\sigma \circ \sigma^{-1})(x)$ for all $x \in X$, which says that $\sigma^{-1} \circ \sigma = e = \sigma \circ \sigma^{-1}$.

I firmly believe that the vast majority of practicing group theorests consider the symmetric groups the most important of all groups!

2. (The **General Linear Group**) Let $\mathcal{R}$ be the real number, let n be a positive integer, and let $(\mathrm{GL}_n(\mathcal{R}), \cdot)$ be the set of all $n \times n$ matrices with coefficients in $\mathcal{R}$ and having non-zero determinant, and where $\cdot$ denotes matrix multiplication. (However, we have already noted above that we'll often use juxtaposition to denote matrix multiplication.) Since matrix multiplication is associative, since the identity matrix has determinant 1 ($\neq 0$), and since the inverse of any matrix of non-zero determinant exists (and also has non-zero determinant) we conclude that $(\mathrm{GL}_n(\mathcal{R}), \cdot)$ is a group. We remark here that we could substitute the coefficients $\mathcal{R}$ with other systems of coefficients, such as $\mathbb{C}$ or $\mathbb{Q}$. (We'll look at another important example in Exercise 4 on page 223.)

3. $(\mathbb{C}, +)$, $(\mathcal{R}, +)$, $(\mathbb{Q}, +)$, $(\mathbb{Z}, +)$ are all groups.

4. Let $\mathbb{C}^* = \mathbb{C} - \{0\}$ (similarly can denote $\mathcal{R}^*$, $\mathbb{Q}^*, \mathbb{Z}^*$, etc.); then $(\mathbb{C}^*, \cdot)$ is a group. Likewise, so are $(\mathcal{R}^*, \cdot)$ and $(\mathbb{Q}^*, \cdot)$ but not $(\mathbb{Z}^*, \cdot)$.

5. Let $\mathbb{Z}_n$ denote the integers modulo n. Then $(\mathbb{Z}_n, +)$ is a group with identity $[0]$ (again, we'll often just denote 0); the inverse of $[x]$ is just $[-x]$.

6. Let $(G, \circ)$ be the set of automorphisms of some graph. If X is the set of vertices of this graph, then $G \subseteq \mathrm{Sym}(X)$; by the proposition

on page 208, we see that G is closed under $\circ$. Since we already know that $\circ$ is associative on $\mathrm{Sym}(X)$ it certainly continues to be associative for G. Next, the identity permutation of the vertices of the graph is clearly a graph automorphism. Finally, the same proposition on page 208 shows that each element $g \in G$ has as inverse, proving that G is a group.

7. There is one other group that is well worth mentioning, and is a multiplicative version of $(\mathbb{Z}_n, +)$. We start by writing $\mathbb{Z}_n^* = \{1, 2, 3, \ldots, n-1\}$ (note, again, that we have dispensed with writing the brackets ($[\cdot]$)). We would like to consider whether this is a group relative to multiplication. Consider, for example, the special case $n = 10$. Note that despite the fact that $2, 5 \in \mathbb{Z}_{10}^*$ we have $2 \cdot 5 = 0 \notin \mathbb{Z}_{10}^*$. In other words $\mathbb{Z}_{10}^*$ is not closed under multiplication and, hence, certainly cannot compose a group.

   The problem here is pretty simple. If the integer n is **not a prime number**, say, $n = n_1 n_2$, where $1 < n_1, n_2 < n$ then it's clear that while $n_1, n_2 \in \mathbb{Z}_n^*$ we have $n_1 n_2 = 0 \notin \mathbb{Z}_n^*$. This says already that $(\mathbb{Z}_n^*, \cdot)$ is not a group. Thus, in order for $(\mathbb{Z}_n^*, \cdot)$ to have any chance at all of being a group, we must have that $n = p$, some prime number. Next, we shall show that if p is prime, then $\mathbb{Z}_p^*$ is closed under multiplication. This is easy, if $a, b \in \mathbb{Z}_p{}^*$, then neither a nor b is divisible by p. But then ab is not divisible by p which means that $ab \neq 0$ and so, in fact, $ab \in \mathbb{Z}_p^*$, proving that $\mathbb{Z}_p^*$ is closed under multiplication.

   Next, note that since multiplication is associative in $\mathbb{Z}_p$, and since $\mathbb{Z}_p^* \subseteq \mathbb{Z}_p$ we have that multiplication is associative in $\mathbb{Z}_p^*$. Clearly $1 \in \mathbb{Z}_p^*$ and is the multiplicative identity. It remains only to show that every element of $\mathbb{Z}_p^*$ has a multiplicative inverse. There are a number of ways to do this; perhaps the following argument is the most elementary. Fix an element $a \in \mathbb{Z}_p^*$ and consider the elements

$$1 \cdot a,\ 2 \cdot a,\ 3 \cdot a,\ \ldots, (p-1) \cdot a \in \mathbb{Z}_p^*.$$

If any two of these elements are the same, say $a'a = a''a$, for distinct elements a' and a'', then $(a' - a'')a = 0$. But this would say that $p \,|(a' - a'')a$; since p is prime, and since $p \not| a$, this implies that $p \,|(a' - a'')$. But $1 \leq a'$, $a'' < p$ and so this is impossible unless $a' = a''$, contradicting our assumption that they were distinct in the first place! Finally, since we now know that the elements in the above list are all distinct, there are exactly $p-1$ such elements, which proves already that

$$\{1 \cdot a,\ 2 \cdot a,\ 3 \cdot a,\ \ldots, (p-1) \cdot a\} \ = \ \mathbb{Z}_p^*.$$

In particular, it follows that $1 \in \{1 \cdot a,\ 2 \cdot a,\ 3 \cdot a,\ \ldots, (p-1) \cdot a\}$, and so $a'a = 1$ for some $a' \in \mathbb{Z}_p^*$, proving that $a' = a^{-1}$. In short, we have proved that $(\mathbb{Z}_p^*, \cdot)$ is a group.

The **multiplication table** for a (finite) group $(G, *)$ is just a table listing all possible products.[9] We give the multiplication table for the group $(\mathbb{Z}_7^*, \cdot)$ below:

$\cdot$	1	2	3	4	5	6
1	1	2	3	4	5	6
2	2	4	6	1	3	5
3	3	6	2	5	1	4
4	4	1	5	2	6	3
5	5	3	1	6	4	2
6	6	5	4	3	2	1

On the basis of the above table, we find, for instance that $4^{-1} = 2$ and that $3^{-1} = 5$. Even more importantly, is this: if we let $x = 3$, we get

$$x^0 = 1,\ x^1 = 3,\ x^2 = 2,\ x^3 = 6,\ x^4 = 4,\ x^5 = 5,\ x^6 = 1,$$

which says that **every element of $\mathbb{Z}_7^*$ can be expressed as some power of the single element 3**. This is both important and more

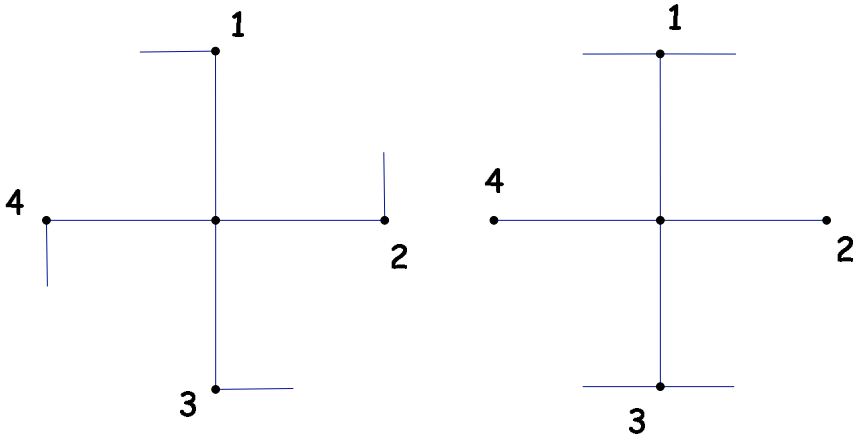
[9]The multiplication table for a group is often called the **Cayley table** after the English mathematician Arthur Cayley (1821–1895).

subtle than it looks, and shall be the topic of the next subsection.

A group $(G, *)$ is called **Abelian**[10] if the operation $*$ is commutative. Granted, it would make good sense to call such groups "commutative," but we enjoy naming concepts after influencial mathematicians. In the above list of groups you should be able to separate the Abelian groups from the non-Abelian ones.

EXERCISES

1. Consider the two graphs given at the beginning of this section; here they are again:



   Write down the elements of the corresponding automorphism groups, and then give the corresponding Cayley tables.

2. In the group $(\mathbb{Z}_{17}, \cdot)$, find 2^{-1} and 5^{-1}. Find any elements x such that $x^2 = 1$.

3. Let $X = \{1, 2, 3\}$ and consider the group $\mathrm{Sym}(X)$ of permutations on X. Define the following two permutations:

$$\sigma = \begin{pmatrix} 1 & 2 & 3 \\ \downarrow & \downarrow & \downarrow \\ 2 & 3 & 1 \end{pmatrix} \qquad \tau = \begin{pmatrix} 1 & 2 & 3 \\ \downarrow & \downarrow & \downarrow \\ 1 & 3 & 2 \end{pmatrix}$$

[10] after the Norwegian mathematics Niels Henrik Abel (1802–1829)

(i) Show that the six elements e, σ, σ^2, τ, $\sigma\tau$, $\sigma^2\tau$ comprise all of the elements of this group.

(ii) Show that $\sigma^3 = \tau^2 = e$ and that $\tau\sigma = \sigma^2\tau$.

(iii) From the above, complete the multiplication table:

$\circ$	e	σ	σ^2	τ	$\sigma\tau$	$\sigma^2\tau$
e	e	σ	σ^2	τ	$\sigma\tau$	$\sigma^2\tau$
σ	σ	σ^2				
σ^2	σ^2				τ	
τ	τ					σ
$\sigma\tau$	$\sigma\tau$				e	
$\sigma^2\tau$	$\sigma^2\tau$					

4. Let G be the set of all 2×2 matrices with coefficients in $\mathbb{Z}_2$ with determinant $\neq 0$. Assuming that multiplication is associative, show that G is a group of order 6. Next, set

$$\boldsymbol{A} = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}, \quad \boldsymbol{B} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}.$$

Show that $\boldsymbol{A}^3 = \boldsymbol{B}^2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ (the identity of G), and that $\boldsymbol{BAB} = \boldsymbol{A}^{-1}$.

5. Let $(G, *)$ be a group such that for every $g \in G$, $g^2 = e$. Prove that G must be an Abelian group.

6. Let $\mathbb{Z}_3$ be the integers modulo 3 and consider the set U of matrices with entries in $\mathbb{Z}_3$, defined by setting

$$U = \left\{ \begin{pmatrix} 1 & a & c \\ 0 & 1 & b \\ 0 & 0 & 1 \end{pmatrix} \middle| \; a, b, c \in \mathbb{Z}_3 \right\}.$$

(a) Show that U is a group relative to ordinary matrix multiplication.

(b) Show that $|U| = 27$.

(c) Show that for every element $x \in U$, $x^3 = e$, where e is the identity of U.

(d) Show that U is **not** abelian.

7. Let $(G, *)$ be a group such that $|G| \leq 4$. Prove that G must be abelian.

8. Let $(G, *)$ be a group such that for all a, $b \in G$ we have $(ab)^2 = a^2b^2$. Prove that G must be Abelian. Find elements a, $b \in \mathrm{Sym}(X)$ in Exercise 3 above such that $(ab)^2 \neq a^2b^2$.

9. Let A be a set. Show that $(2^A, +)$ is an Abelian group, but that if $|A| \geq 2$ then $(2^A, \cap)$ and $(2^A, \cup)$ are **not** groups at all.

10. Here's another proof of the fact that if p is prime, then every element $a \in \mathbb{Z}_p^*$ has an inverse. Given that a, p are relatively prime, then by the Euclidean trick (page 58) there exist integers s and t with $sa + tp = 1$. Now what?

4.2.5 Cyclic groups

At the end of the previous subsection we observed that the multiplicative group $(\mathbb{Z}_7^*, \cdot)$ has every element representable as a power of the element 3. This is a very special property, which we formalize as follows.

Definition of Cyclic Group. Let $(G, *)$ be a group. If there exists an element $g \in G$ such that every element is a power (possibly negative) of x, then $(G, *)$ is called a **cyclic group**, and the element x is called a **generator** of G. Note that a cyclic group is necessarily Abelian. To see this, assume that the group G is cyclic with generator x and that g, $g' \in G$. Then $g = x^m$ and $g' = x^n$ for suitable powers m, n, and so

$$gg' \;=\; x^m * x^n \;=\; x^{m+n} \;=\; x^{n+m} \;=\; x^n x^m \;=\; g'g,$$

proving that G is abelian.

Let's look at a few examples.

1. The infinite additive group $(\mathbb{Z}, +)$ is cyclic, with generator 1. Note, however, in this context, we wouldn't write 1^n for powers of 1 as this notation is suggestive of multiplication and $1^n = 1$. Rather, in this additive setting we write

$$n1 = \underbrace{1 + 1 + 1 + \cdots + 1}_{n \text{ terms}}.$$

   As any integer can be written as a positive or negative multiple ("multiple" is the additive version of "power"), we conclude that $(\mathbb{Z}, +)$ is an (infinite) cyclic group.

2. If n is a positive integer, then the additive group $(\mathbb{Z}_n, +)$ is cyclic. Notice here that we don't really need any negative multiples of 1 to obtain all of $\mathbb{Z}_n$. One easy way to see this is that $-1 = (n-1)1$ and so if $[a] \in \mathbb{Z}_n$, then $-[a] = a(n-1)1$.

3. If p is prime, then the multiplicative group $(\mathbb{Z}_p^*, \cdot)$ is cyclic. While not a deep fact, this is not easy to show using only what we've learned up to this point.[11] As examples, note that $\mathbb{Z}_5^*$ is cyclic, with generator 2, as $2^1 = 2$, $2^2 = 4$, $2^3 = 8 = 3$, and $2^4 = 16 = 1$. Next, $\mathbb{Z}_7^*$ is cyclic, with generator 3, as

$$3^1 = 3,\ 3^2 = 9 = 2,\ 3^3 = 6,\ 3^4 = 4,\ 3^5 = 5,\ 3^6 = 1.$$

   Note, however, that while 2 is a generator of $\mathbb{Z}_5^*$, it is **not** a generator of $\mathbb{Z}_7^*$.

   Related to the above is the following famous **unsolved conjecture**: that the congruence class of the integer "2" a generator of

[11] Most proofs proceed along the following lines. One argues that if $\mathbb{Z}_p^*$ is not cyclic, then there will have to exist a proper divisor k of $p-1$ such that every element x of $\mathbb{Z}_p^*$ satisfies $x^k = 1$. However, this can be interpreted as a polynomial equation of degree k which has $(p-1) > k$ solutions. Since $(\mathbb{Z}_p, +, \cdot)$ can be shown to be a "field," one obtains a contradiction.

$\mathbb{Z}_p^*$ for infinitely many primes p. This is often called the **Artin Conjecture**, and the answer is "yes" if one knows that the so-called **Generalized Riemann Hypothesis** is true! Try checking this out for the first few primes, noting (as above) that 2 is **not** a generator for $\mathbb{Z}_7^*$.

4. Let n be a positive integer and consider the set of complex numbers

$$C_n = \{e^{2\pi ki/n} = \cos 2\pi k/n + i \sin 2\pi k/n \mid k = 0,\ 1,\ 2,\ \ldots,\ n-1\} \subseteq \mathbb{C}.$$

If we set $\zeta = e^{2\pi/n}$, then $e^{2\pi ki/n} = \zeta^k$. Since also $\zeta^n = 1$ and $\zeta^{-1} = \zeta^{n-1}$ we see that not only is C_n closed under multiplication, it is in fact, a cyclic group.

We hasten to warn the reader that in a cyclic group the generator is almost never unique. Indeed, the inverse of any generator is certainly also a generator, but there can be even more. For example, it is easy to check that every non-identity element of the additive cyclic $(\mathbb{Z}_5, +)$ is a generator. This follows by noting that 1 is a generator and that

$$1 = 3 \cdot 2 \ = \ 2 \cdot 3 \ = \ 4 \cdot 4.$$

On the other hand, we showed above that 3 is a generator of the cyclic group $\mathbb{Z}_7^*$, and since $3^{-1} = 5$ (because $3 \cdot 5 = 1$), we see that 5 is also a generator. However, these can be shown to be the only generators of $\mathbb{Z}_7^*$. In general, if G is a cyclic group of order n, then the number of generators of G is $\phi(n)$, where, as usual, ϕ is the Euler ϕ-function; see Exercise 6, below.

In fact, we'll see in the next section that if $(G, *)$ is a group of prime order p, then not only is G cyclic, every non-identity element of G is a generator.

We shall conclude this section with a useful definition. Let $(G, *)$ be a group, and let $g \in G$. The **order** of g is the least positive integer n such that $g^n = e$. We denote this integer by $o(g)$. If no such integer exists, we say that g has **infinite order**, and write $o(g) = \infty$. Therefore, for example, in the group $\mathbb{Z}_7^*$, we have

$$o(1) = 1,\ o(2) = 3,\ o(3) = 6,\ o(4) = 3,\ o(5) = 6,\ o(6) = 2.$$

Note that if the element g has order n, then

$$\{g^k \mid k \in \mathbb{Z}\} \ = \ \{e,\, g,\, g^2,\, \cdots, g^{n-1}\}$$

and all of the elements of $\{e,\, g,\, g^2,\, \cdots, g^{n-1}\}$ **are distinct**. To see this, note that when we divide any integer k by n we may produce a quotient q and a remainder r, where $0 \leq r \leq n-1$. In other words we may express $k = qn + r$, which implies that $g^k = g^{qn+r} = g^{qn}g^r = (g^n)^q g^r = eg^r = g^r$. Therefore we already conclude that $\{g^k \mid k \in \mathbb{Z}\} = \{e,\, g,\, g^2,\, \cdots, g^{n-1}\}$. Next, if $e,\, g,\, g^2,\, \cdots, g^{n-1}$ aren't all distinct, then there must exist integers $k < m,\ 0 \leq k < m \leq n-1$ such that $g^k = g^m$. But then $e = g^m g^{-k} = g^{m-k}$. But clearly $0 < m - k \leq n - 1$ which contradicts the definition of the order of g. This proves our assertion.

Note that in general, $o(g) = 1$ precisely when $g = e$, the identity element of G. Also, if G is a finite group with n elements, and if G has an element g of order n, then G is cyclic and g is a generator of G (Exercise 4).

Exercises

1. The two groups you computed in Exercise 1 of Subsection 4.2.4 both have order 4: one is cyclic and one is not. Which one is cyclic? What are the generators of this group?

2. Let G be a group and let g be an element of finite order n. Show that if $g^m = e$ then m must be a multiple of n, i.e., $n|m$.

3. Assume that G is a group and that $g \in G$ is an element of finite order n. Assume that k is a positive integer which is **relatively prime** to n (see page 60). Show that the element g^k also has order n.

4. Let G be a finite group of order n, and assume that G has a element g of order n. Show that G is a cyclic group and that g is a generator.

5. Let G be a finite cyclic group of order n and assume that x is a generator of G. Show that $|G| = o(x)$, i.e., the **order of the group** G is the same as the **order of the element** x.

6. Let G be a cyclic group of order n. Show that the number of generators of G is $\phi(n)$. (Hint: let $x \in G$ be a fixed generator; therefore, any element of G is of the form x^k for some integer k, $0 \leq k \leq n-1$. Show that x^k is also a generator if and only if k and n are relatively prime.)

4.2.6 Subgroups

Most important groups actually appear as "subgroups" of larger groups; we shall try to get a glimpse of how such a relationship can be exploited.

Definition. Let $(G, *)$ be a group and let $H \subseteq G$ be a subset of G. We say that H is a **subgroup** of G if

(i) H is closed under the operation $*$, and

(ii) $(H, *)$ is also a group.

Interestingly enough, the condition (i) above (closure) is almost enough to guarantee that a subset $H \subseteq G$ is actually a subgroup. There are two very useful and simple criteria each of which guarantee that a given subset is actually a subgroup.

Proposition. *Let $(G, *)$ be a group and let $H \subseteq G$ be a non-empty subset.*

(a) If for any pair of elements $h, h' \in H$, $h^{-1}h' \in H$, then H is a subgroup of G.

(b) If $|H| < \infty$ and H is closed under $*$, then H is a subgroup of G.

Proof. Notice first that we don't have to check the associativity of $*$, as this is already inherited from the "parent" group G. Now assume condition (a). Since H is non-empty, we may choose an element $h \in H$. By condition (a), we know that $e = h^{-1}h \in H$, and so H contains the identity element of G (which is therefore also the identity element of H). Next, given $h \in H$ we appeal again to condition (a) to obtain (since $e \in H$) $h^{-1} = h^{-1} * e \in H$. It follows that H is a subgroup of G.

Next, assume condition (b), and let $h \in H$. Since H is closed under $*$, we conclude that all of the products h, h^2, h^3, ... are all in H. Since H is a finite set, it is impossible for all of these elements to be distinct, meaning that there must be powers $m < n$ with $h^m = h^n$. This implies that $e = h^{n-m} \in H$, forcing H to contain the identity of G. Furthermore, the same equation above shows that $e = h^{n-m-1} * h$, where $n-m-1 \geq 0$. Therefore, $h^{n-m-1} \in H$ and $e = h^{n-m-1} * h$ implies that $h^{-1} = h^{n-m-1} \in H$. Therefore, we have shown that H contains both the identity and the inverses of all of its elements, forcing H again to be a subgroup of G.

I can't overestate how useful the above result is!

One very easy way to obtain a subgroup of a given group $(G, *)$ is start with an element $x \in G$ and form the set $H = \{x^k \mid k \in \mathbb{Z}\}$; that is, H contains all the positive and negative powers of x. Clearly H satisfies condition (a) of the above proposition since x^m, $x^n \in H \Rightarrow x^{-m}x^n = x^{m-n} \in H$. Therefore H is a subgroup of G; as H is cyclic, we say that H is the **cyclic subgroup of G generated by** x. This cyclic subgroup generated by x is often denoted $\langle x \rangle$.

Exercises

1. Let G be a group, and let H_1 and H_2 be subgroups. Prove that the intersection $H_1 \cap H_2$ is also a subgroup of G.

2. Let G be a group, and let H_1 and H_2 be subgroups. Prove that unless $H_1 \subseteq H_2$ or $H_2 \subseteq H_1$, then $H_1 \cup H_2$ is **not** a subgroup of G.

3. Show that the even integers $2\mathbb{Z}$ is a subgroup of the additive group of the integers $(\mathbb{Z}, +)$. In fact, show that if n is any positive integer, then the set $n\mathbb{Z}$ of multiples of n is a subgroup of $(\mathbb{Z}, +)$.

4. Show that any subgroup H of the additive group $(\mathbb{Z}, +)$ of the integers must be cyclic.

5. Show that any subgroup $H \neq \{0\}$ of the additive group $(\mathbb{C}, +)$ of complex numbers must be infinite.

6. Consider the group $G = \mathrm{GL}_2(\mathcal{R})$ of 2×2 matrices of non-zero determinant. Find an element (i.e., a matrix) $\boldsymbol{A}$ of finite order and an element $\boldsymbol{B}$ of infinite order. Conclude that G has both finite and infinite subgroups.

7. Let $X = \{1, 2, 3, 4\}$ and set $G = \mathrm{Sym}(X)$, the group of permutations of X. Find all of the elements in G having order 2. Find all of the elements of G having order 3. Find all of the elements of G having order 4.

8. Let $(G, *)$ be a cyclic group and let $H \subseteq G$, $H \neq \{e\}$ be a subgroup. Show that H is also cyclic. (This is not entirely trivial! Here's a hint as to how to proceed. Let G have generator x and let n be the **smallest** positive integer such that $x^n \in H$. Show that, in fact, x^n is a generator of H.)

9. Consider the set $\mathcal{R}^+$ of positive real numbers and note that $(\mathcal{R}^+, \cdot)$ is a group, where "$\cdot$" denotes ordinary multiplication. Show that $\mathcal{R}^+$ has elements of finite order as well as elements of infinite order and hence has both finite and infinite subgroups.

10. Consider a graph with set X of vertices, and let G be the automorphism group of this graph. Now fix a vertex $x \in X$ and set $G_x = \{\sigma \in G \mid \sigma(x) = x\}$. Prove that G_x is a subgroup of G, often called the **stabilizer** in G of the vertex x.

11. Find the orders of each of the elements in the cyclic group $(\mathbb{Z}_{12}, +)$.

12. Let p be a prime number and let $\mathbb{Z}_p$ be the integers modulo three and consider the group $\mathrm{GL}_2(\mathbb{Z}_p)$ of matrices having entries in $\mathbb{Z}_3$ and all having nonzero determinant.

(a) Show that $\mathrm{GL}_2(\mathbb{Z}_p)$ is a group.

(b) Show that there are $p(p+1)(p-1)^2$ elements in this group.[12]

(c) Let B be the set of *upper triangular* matrices inside $\mathrm{GL}_2(\mathbb{Z}_p)$. Therefore,

$$B = \left\{ \begin{pmatrix} a & b \\ 0 & c \end{pmatrix} \;\middle|\; ac \neq 0 \right\} \subseteq \mathrm{GL}_2(\mathbb{Z}_p).$$

Show that B is a subgroup of $\mathrm{GL}_2(\mathbb{Z}_p)$, and show that $|B| = p(p-1)^2$.

(d) Define $U \subseteq B$ to consist of matrices with 1s on the diagonal. Show that U is a subgroup of B and consists of p elements.

4.2.7 Lagrange's theorem

In this subsection we shall show a potentially surprising fact, namely that if H is a subgroup of the finite group G, then the order $|H|$ evenly divides the order $|G|$ of G. This severly restricts the nature of subgroups of G.

The fundamental idea rests on an equivalence relation in the given group, relative to a subgroup. This relationship is very similar to the congruence relation (mod n) on the additive group $\mathbb{Z}$ of integers. Thus, let $(G, *)$ be a group and let $H \subseteq G$ be a subgroup. Define a relation on G, denoted (mod H) defined by stipulating that

$$g \equiv g' \,(\text{ mod } H) \;\Leftrightarrow g^{-1}g' \in H.$$

This is easy to show is an equivalence relation:

[12]This takes a little work. However, notice that a matrix of the form $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ will have nonzero determinant precisely when not both a and b are 0 and when the "vector" (c, d) is not a multiple of the "vector" (a, b). This implies that there are $p^2 - 1$ possibilities for the first row of the matrix and $p^2 - p$ possibilities for the second row. Now put this together!

reflexivity: $g \equiv g(\bmod H)$ since $g^{-1}g = e \in H$.

symmetry: If $g \equiv g'(\bmod H)$ then $g^{-1}g' \in H$, and so $g'^{-1}g = (g^{-1}g')^{-1} \in H$. Therefore, also $g' \equiv g(\bmod H)$

transitivity: If $g \equiv g'(\bmod H)$ and $g' \equiv g''(\bmod H)$, then $g^{-1}g'$, $g'^{-1}g'' \in H$. But then $g^{-1}g'' = g^{-1}g'g'^{-1}g'' \in H$, proving that also $g \equiv g''(\bmod H)$.

Pretty easy, eh?

As a result we see that G is partitioned into mutually disjoint equivalence classes. Next we shall actually determine what these equivalence classes look like. Thus let $g \in G$ and let $[g]$ be the equivalence class (relative to the above equivalence relation) containing g.

We Claim: $[g] = gH = \{gh \mid h \in H\}$.

Proof of Claim: Note first that an element of gH looks like gh, for some $h \in H$. Since $g^{-1}(gh) = h \in H$ we see that $g \equiv gh(\bmod H)$, i.e., $gh \in [g]$. This proves that $gH \subseteq [g]$. Conversely, assume that $g \equiv g'(\bmod H)$, i.e., that $g \equiv g'(\bmod H)$, which says that $g^{-1}g' \in H$. But then $g^{-1}g' = h$ for some $g' = gh \in gH$. This proves that $[g] \subseteq gH$ and so $[g] = gH$.

Next we would like to show that H is a finite subgroup of G then the elements of each equivalence class gH, $g \in G$ have the same number of elements. In fact, we shall show that $|gH| = |H|$, for each $g \in G$. To prove this we shall define a mapping $f : H \to gH$ and show that it is a bijection. Namely, we define $f(h) = gh$, $h \in H$.

f **is one-to-one**: If h, $h' \in H$ and if $f(h) = f(h')$, then $gh = gh'$. We now multiply each side by g^{-1} and get $h = g^{-1}gh = g^{-1}gh' = h'$. Thus f is one-to-one.

f **is onto**: If $gh \in gH$, then $gh = f(h)$ and so f is onto.

It follows, therefore, that $|gH| = |H|$ for each element $g \in G$. If G is also a finite group, this says that G is partitioned into sets, each of

which has cardinality $|H|$. If G is partitioned into k such sets, then obviously $|G| = k|H|$, which proves that $|H|$ is a divisor of the group order G.

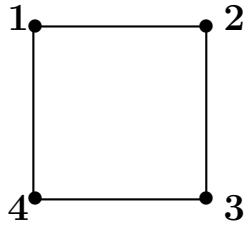
We summarize the above in the following theorem.

Lagrange's Theorem. *Let G be a finite group and let H be a subgroup of G. Then $|H|\big||G|$.*

If G is a finite group and $g \in G$, then we have seen that $o(g)$ is the order of the subgroup $\langle g \rangle$ that it generates. Therefore,

Corollary. *If G is a finite group and $g \in G$, then $o(g)\big||G|$.*

Example. Consider the graph given to the right, with four vertices, and let G be the automorphism group of this graph. Notice that if $X = \{1, 2, 3, 4\}$, then G is a subgroup of $\mathrm{Sym}(X)$, the group of permutations of of the four vertices. Therefore, we infer immediately that $|G|$ is a divisor of $4! = 24$.



Note that two very obvious automorphisms of this graph are the permutations

$$\sigma : \begin{pmatrix} 1 & 2 & 3 & 4 \\ \downarrow & \downarrow & \downarrow & \downarrow \\ 2 & 3 & 4 & 1 \end{pmatrix}, \quad \tau : \begin{pmatrix} 1 & 2 & 3 & 4 \\ \downarrow & \downarrow & \downarrow & \downarrow \\ 2 & 1 & 4 & 3 \end{pmatrix}$$

Next, note that σ has order 4 and τ has order 2. Finally, note that $\tau\sigma\tau = \sigma^3$ $(= \sigma^{-1})$. Let $C = \langle \sigma \rangle$ and set $D = \langle \tau \rangle$ be the cyclic subgroups generated by σ and τ. Note that since $\tau \notin C$, we conclude that $|G| > 4 = |C|$; since by Langrange's Theorem we must have $|C|\big||G|$, we must have that $|G|$ is a multiple of 4 and is strictly larger than 4. Therefore $|G| \geq 8$. Also since G is a subgroup of $\mathrm{Sym}(X)$, we

see that $|G|\big|24$. But there are plenty of permutations of the vertices 1, 2, 3, and 4 which are not graph automorphisms (find one!), and so $|G| < 24$.

On the other hand, note that the powers e, σ, σ^2, σ^3 give four automorphisms of the graph, and the elements τ, $\sigma\tau$, $\sigma^2\tau$, $\sigma^3\tau$ give four more. Furthermore, since $\tau\sigma\tau = \sigma^3$ we can show that the set $\{e, \sigma, \sigma^2, \sigma^3, \tau, \sigma\tau, \sigma^2\tau, \sigma^3\tau\}$ is closed under multiplication and hence is a subgroup of G. Therefore $8\big||G|$ and so it follows that $|G| = 8$ and the above set is all of G:

$$G = \{e, \sigma, \sigma^2, \sigma^3, \tau, \sigma\tau, \sigma^2\tau, \sigma^3\tau\}.$$

Below is the multiplication table for G (you can fill in the missing elements). Notice that G has quite a few subgroups—you should be able to find them all (Exercise 3).

$\circ$	e	σ	σ^2	σ^3	τ	$\sigma\tau$	$\sigma^2\tau$	$\sigma^3\tau$
e	e	σ	σ^2	σ^3	τ	$\sigma\tau$	$\sigma^2\tau$	$\sigma^3\tau$
σ	σ	σ^2	σ^3	e				
σ^2	σ^2	σ^3	e	σ				
σ^3	σ^3	e	σ	σ^3				
τ	τ							
$\sigma\tau$	$\sigma\tau$							
$\sigma^2\tau$								
$\sigma^3\tau$								

EXERCISES

1. Use the corollary on page 233 to give another proof of Fermat's Little Theorem; see page 86.

2. Suppose that G is a finite group of prime order p. Prove that G must be cyclic.

3. Refer to the multiplication table above for the group G of symmetries of the square and list all of the subgroups.

4. Let G be a group, let H be a subgroup, and recall the equivalence relation (mod H) defined by

$$g \equiv g' \text{ (mod } H) \Leftrightarrow g^{-1}g' \in H.$$

The equivalence classes in G relative to this equivalence relation are called the (left) **cosets** of H in G. Are the cosets also subgroups of G? Why or why not?

5. Let G be the group of Exercise 3 and let K be the cyclic subgroup generated by $\sigma\tau$. Compute the left cosets of K in G.

6. Let G be the group of Exercise 3 and let L be the subgroup $\{e, \tau, \sigma^2, \sigma^2\tau\}$. Compute the left cosets of L in G.

7. Here we shall give yet another proof of the infinitude of primes. Define, for each prime p the corresponding **Mersenne number** by setting $M_p = 2^p - 1$ (these are often primes themselves). Assume by contradiction that there are only finitely many primes and let p be the **largest** prime. Let q be a prime divisor of $M_p = 2^p - 1$. Then we have, in the multiplicative group $\mathbb{Z}_q^*$ of nonzero integers modulo q, that $2^p \equiv 1 (\text{mod } q)$. This says, by exercise 2 on page 227 that p is the order of 2 in the group $\mathbb{Z}_q^*$. Apply Lagrange's theorem to obtain $p \,|\, (q-1)$, proving in particular that q is a larger prime than p, a contradiction.

4.2.8 Homomorphisms and isomorphisms

What is the difference between the additive group $(\mathbb{Z}_6, +)$ and the multiplicative group $(\mathbb{Z}_7^*, \cdot)$? After all, they are both cyclic: $(\mathbb{Z}_6, +)$ has generator 1 (actially, [1]), and $(\mathbb{Z}_7^*, \cdot)$ has generater 3 ([3]). So wouldn't it be more sensible to regard these two groups as algebraically the same, the only differences being purely cosmetic? Indeed, doesn't any cyclic group of order 6 look like $\{e, x, x^2, x^3, x^4, x^5, x^6\}$?

Here's a much less obvious example. Consider the two infinite groups $(\mathcal{R}, +)$ and $(\mathcal{R}^+, \cdot)$. At first blush these would seem quite different. Yet, if we consider the mapping $f : \mathcal{R} \to \mathcal{R}^+$ given by $f(x) = e^x$ (the exponential function) then f is not only a bijection (having inverse ln) but this mapping actually matches up the two binary operations:

$$f(x + y) \;=\; e^{x+y} \;=\; e^x \cdot e^y \;=\; f(x) \cdot f(y).$$

Notice that the inverse mapping $g(x) = \ln x$, does the same, but in the reverse order:

$$g(x \cdot y) \;=\; \ln(x \cdot y) \;=\; \ln x + \ln y \;=\; g(x) + g(y).$$

The point of the above is that through the mappings f and its inverse g we see that group structure of $(\mathcal{R}^+, \cdot)$ is faithfully represented by the group structure of $(\mathcal{R}^+, \cdot)$, i.e., the two groups are "isomorphic." We shall formalize this concept below.

Definition of Homomorphism: Let $(G, *)$ and $(H, \star)$ be groups, and let $f : G \to H$ be a mapping. We say that f is a **homomorphism** if for all g, $g' \in G$ we have $f(g * g') = f(g) \star f(g')$. In other words, in finding the image of the product of elements g, $g' \in G$, it doesn't matter whether you first compute the product $g * g'$ in G and then apply f or to first apply f to g and g' and then compute the product $f(g) \star f(g')$ in H.

Of course, we now see that the exponential mapping from $(\mathcal{R}, +)$ to $(\mathcal{R}^+, \cdot)$ is a homomorphism.

Here's another example. Recall the group $\mathrm{GL}_2(\mathcal{R})$ of 2×2 matrices having real coefficients and non-zero determinants. Since we know that $\det(\boldsymbol{A} \cdot \boldsymbol{B}) = \det(\boldsymbol{A}) \cdot \det(\boldsymbol{B})$ we see that $\det : \mathrm{GL}_2(\mathcal{R}) \to \mathcal{R}^*$ is a homomorphism, where $\mathcal{R}^*$ denotes the multiplicative group of non-zero real numbers.

Definition of Isomorphism: If $f : G \to H$ is a homomorphism of groups $(G, *)$ and $(H, \star)$, we say that f is an **isomorphism** if f is

bijective. Note that in this case, the inverse mapping $f^{-1} : H \to G$ is also a homomorphism. The argument is as follows: if h, $h' \in H$ then watch this:

$$\begin{aligned} f(f^{-1}(h) * f^{-1}(h')) &= f(f^{-1}(h)) \star f f^{-1}(h')) \quad \begin{pmatrix}\text{since } f \text{ is a homo-}\\ \text{morphism}\end{pmatrix} \\ &= h \star h' \qquad\qquad \begin{pmatrix}\text{since } f \text{ and } f^{-1}\\ \text{are inverse func-}\\ \text{tions}\end{pmatrix} \\ &= f(f^{-1}(h \star h')) \quad \text{(same reason!)} \end{aligned}$$

However, since f is one-to-one, we infer from the above that

$$f^{-1}(h) * f^{-1}(h') = f^{-1}(h \star h'),$$

i.e., that f^{-1} is a homomorphism.

Before going any further, a few comments about homomorphisms are needed here. Namely, let G_1 and G_2 be groups (we don't need to emphasize the operations here), and assume that e_1 and e_2 are the identity elements of G_1 and G_2, respectively Assume that $f : G_1 \to G_2$ is a homomorphism. Then,

$\underline{f(e_1) = e_2}$. This is because $f(e_1)^2 = f(e_1)f(e_1) = f(e_1e_1) = f(e_1)$. Now multiply both sides by $f(e_1)^{-1}$ and get $f(e_1) = e_2$.

$\underline{\text{If } x \in G_1, \text{ then } f(x^{-1}) = f(x)^{-1}}$. Note that by what we just proved, $e_2 = f(e_1) = f(xx^{-1}) = f(x)f(x^{-1})$. Now multiply both sides by $f(x)^{-1}$ and get $f(x)^{-1} = f(x)^{-1}e_2 = f(x)^{-1}(f(x)f(x^{-1})) = (f(x)f(x)^{-1})f(x^{-1}) = e_2f(x^{-1}) = f(x^{-1})$.

Theorem. *Let $(G, *)$ and $(H, \star)$ be cyclic groups of the same order n. Then $(G, *)$ and $(H, \star)$ are isomorphic.*

Proof. Let G have generator x and let H have generator y. Define the mapping $f : G \to H$ by setting $f(x^k) = x^k$, $k = 0, 1, 2, \ldots, n-1$. Note that f is obviously onto. But since $|G| = |H| = n$ it is obvious

that f is also one-to-one. (Is this obvious?) Finally, let x^k, $x^l \in G$; if $k + l \leq n - 1$, then $f(x^k x^l) = f(x^{k+l}) = y^{k+l} = y^k y^l = f(x^k) f(x^l)$. However, if $k + l \geq n$, then we need to divide n into $k + l$ and get a remainder r, where $0 \leq r \leq n - 1$, say $k + l = qn + r$, where q is the quotient and r is the remainder. We therefore have that

$$\begin{aligned}
f(x^k x^l) &= f(x^{k+l}) \\
&= f(x^{qn+r}) \\
&= f(x^{qn} x^r) \\
&= f(e_G x^r) \qquad \text{(since } x^{qn} = e_G\text{, the identity of } G\text{)} \\
&= f(x^r) \\
&= y^r \qquad \text{(by definition of } f\text{)} \\
&= e_H y^r \qquad \text{(where } e_H \text{ is the identity of } H\text{)} \\
&= y^{qn} y^r \qquad \text{(since } y^{qn} = e_H\text{)} \\
&= y^{qn+r} \\
&= y^{k+l} \\
&= y^k y^l.
\end{aligned}$$

EXERCISES

1. Let $f : G_1 \to G_2$ be a homomorphism of groups, and let $H_1 \subseteq G_1$ be a subgroup of G_1. Prove that the image, $f(H_1) \subseteq G_2$ is a subgroup of G_2.

2. (A little harder) Let $f : G_1 \to G_2$ be a homomorphism of groups, and let $H_2 \subseteq G_2$ be a subgroup of G_2. Set $f^{-1}(H_2) = \{g_1 \in G_1 \mid f(g_1) \in H_2\}$. Prove that $f^{-1}(H_2)$ is a subgroup of G_1.

3. Let $\mathrm{GL}_2(\mathcal{R})$ be the group of 2×2 matrices with real coefficients and determinant not 0, and let $\boldsymbol{S}$ be a nonsingular matrix. Define the mapping $f : \mathrm{GL}_2(\mathcal{R}) \to \mathrm{GL}_2(\mathcal{R})$ by setting $f(\boldsymbol{A}) = \boldsymbol{S}\boldsymbol{A}\boldsymbol{S}^{-1}$. Prove that f is an isomorphism of $\mathrm{GL}_2(\mathcal{R})$ onto itself.

4. Again, let $\mathrm{GL}_2(\mathcal{R})$ be the group of 2×2 matrices with real coefficients and determinant not 0. Show that the determinant defines

a homomorphism of $\mathrm{GL}_2(\mathcal{R})$ into the multiplicative group of nonzero real numbers.

5. (Really the same as Exercise 3) Let G be any group and fix an element $x \in G$. Prove that the mapping $f : G \to G$ defined by setting $f(g) = xgx^{-1}$ is an isomorphism of G onto itself.

6. Let A be an **Abelian** group and let $f : A \to B$ be a **surjective** homomorphism, where B is also a group. Prove that B is also Abelian.

7. Let $f : G \to H$ be a homomorphism of groups and set $K = \{g \in G \mid f(g) = e_H\}$, where e_H is the identity of H. Prove that K is a subgroup of G.[13]

8. Let $X = \{1, 2, 3, \ldots, n\}$, where n is a positive integer. Recall that we have the group $(2^X, +)$, where, as usual, 2^X is the power set of X and $+$ is symmetric difference (see page 194). Define $f : 2^X \to \{-1, 1\}$ (where $\{\pm 1\}$ is a group with respect to multiplication) by setting

$$f(A) = \begin{cases} +1 & \text{if } |A| \text{ is even} \\ -1 & \text{if } |A| \text{ is odd.} \end{cases}$$

   Prove that f is a homomorphism.

9. Let G be a group and define $f : G \to G$ by setting $f(g) = g^{-1}$.

   (a) Show that f is a bijection.

   (b) Under what circumstances is f a homomorphism?

10. Prove that the automorphism groups of the graphs on page 207 (each having four vertices) are not isomorphic.

11. Let $\mathbb{R}$ be the additive group of real numbers and assume that $f : \mathbb{R} \to \mathbb{R}$ is a function which satisfies $f(x - y) = f(x) - f(y)$, for $0 < x, y \in \mathbb{R}$. Prove that $f : \mathbb{R} \to \mathbb{R}$ is a homomorphism.

[13] This subgroup of G is usually called the **kernel** of the homomorphism f.

12. Let $\mathbb{R}$ be the additive group of real numbers and assume that $f : \mathbb{R} \to \mathbb{R}$ is a function which satisfies the peculiar property that $f(x^2 - y^2) = xf(x) - yf(y)$ for all $x, y \in \mathbb{R}$.

    (a) Prove that $f : \mathbb{R} \to \mathbb{R}$ is a homomorphism, and that

    (b) there exists a real number $c \in \mathbb{R}$ such that $f(x) = cx,\ x \in \mathbb{R}$.

    The result of Exercise 12 is strictly stronger that that of Exercise 11. Indeed the condition of Exercise 12 shows that the homomorphism is **continuous** and of a special form. We'll return to this briefly in Chapter 5; see Exercise 6 on page 252.

13. Let G be a group and let $\mathbb{C}^*$ denote the multiplicative group of complex numbers. By a (linear) **character** we mean a homomorphism $\chi : G \to \mathbb{C}^*$, i.e., $\chi(g_1 g_2) = \chi(g_1)\chi(g_2)$ for all $g_1, g_2 \in G$.

    (a) Prove that if $\chi : G \to \mathbb{C}^*$ is a character, then $\chi(g^{-1}) = \overline{\chi(g)}$ (complex conjugate) for all $g \in G$.

    Now assume that G is finite and that $\chi : G \to \mathbb{C}^*$ is a character such that for at least one $g \in G$, $\chi(g) \neq 1$. Prove that

    (b) $\sum_{g \in G} \chi(g) = 0$.

    (c) Let $\chi_1, \chi_2 : G \to \mathbb{C}^*$ be distinct characters and prove that $\sum_{g \in G} \chi_1(g)\overline{\chi_2(g)} = 0$.

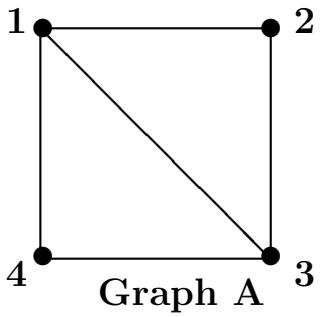
    (d) Fix the positive integer n and show that for any integer $k = 0, 1, 2, \ldots, n-1$ the mapping $\chi_k : \mathbb{Z}_n \to \mathbb{C}^*$ given by $\chi_k(a) = \cos(2\pi ka/n) + i \sin(2\pi ka/n)$ is a character. Show that any character of $\mathbb{Z}_n$ must be of the form $\chi_k,\ 0 \leq k < n$, as above.

4.2.9 Return to the motivation

We return to the two graphs having six vertices each on page 206, and make a simple observation about their automorphism groups. Namely,

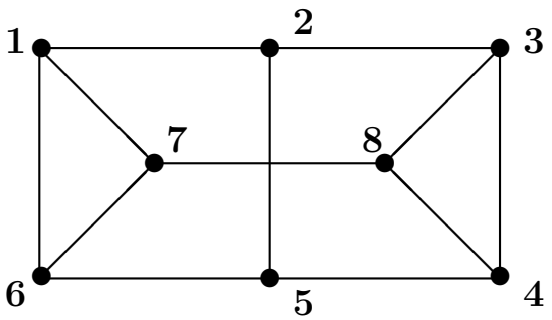
for any two vertices, call them x_1 and x_2, there is always an automorphism σ which carries one to the other. This is certainly a property of the graph—such graphs are called **vertex-transitive** graphs. A simple example of a non vertex-transitive graph is as follows:

Since vertex 1 is contained in three edges, and since vertex 2 is contained in only two, it is obviously impossible to construct an automorphism which will map vertex 1 to vertex 2.



Graph A

The graph to the right doesn't have the same deficiency as the above graph, and yet a moment's thought will reveal that there cannot exist an automorphism which carries vertex 1 to vertex 2.



Graph B

The following result is fundamental to the computation of the order of the automorphism group of a vertex-transitive graph.[14] Its usefulness is that it reduces the computation of the size of the automorphism group of a vertex-transitive graph to the computation of a stabilizer (which is often much easier).

Theorem. *Let G be the automorphism group of a vertex-transitive graph having vertex set X. Fix $x \in X$ and let G_x be the stabilizer in G of x. Then $|G| = |X| \cdot |G_x|$.*

Proof. Let H be the stabilizer in G of the fixed vertex x: $H = \{\sigma \in G \mid \sigma(x) = x\}$. Recall the equivalence relation on G introducted in Subsection 4.2.7, namely

[14] I have used this result many times in my own research!

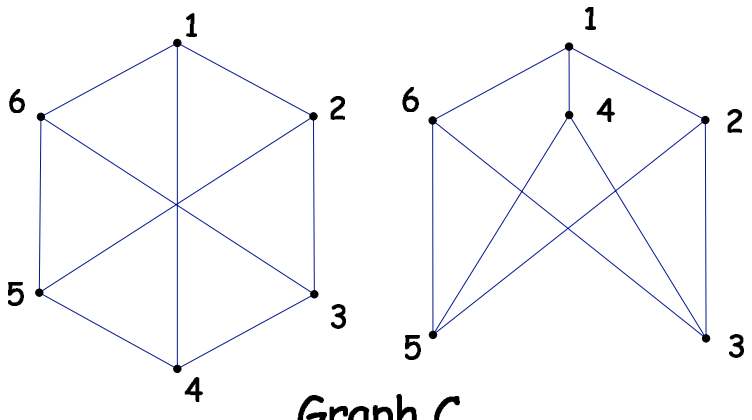
$$\sigma \equiv \sigma' \,(\mathrm{mod}\, H) \iff \sigma^{-1}\sigma' \in H.$$

Recall also from Subsection 4.2.7 that the equivalence classes each have order $|H|$; if we can count the number of equivalence classes in G, we'll be done! Now define a mapping $f : G \to X$ by the rule $f(\sigma) = \sigma(x)$. Since the graph is vertex-transitive, we see that this mapping is surjective. Note that

$$\begin{aligned} f(\sigma) = f(\sigma') &\Leftrightarrow \sigma(x) = \sigma'(x) \\ &\Leftrightarrow \sigma^{-1}\sigma(x) = x \\ &\Leftrightarrow \sigma^{-1}\sigma \in H \\ &\Leftrightarrow \sigma \equiv \sigma' \,(\mathrm{mod}\, H). \end{aligned}$$

In other words, there are exactly as many equivalence classes mod H as there are vertices of X! This proves the theorem.

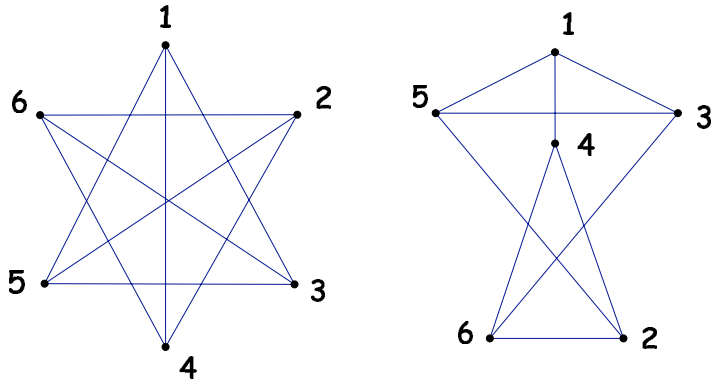
We turn now to the computation of the size of the automorphism groups of the two graphs introduced at the begining of this section. In order to compute the order of a stabilizer, we re-draw the graph from the "point of view" of a particular vertex. Thus, consider the following graphs, where the second emphasizes the role of the vertex 1:



Graph C

If H is the stabilizer of the vertex 1, then surely H must permute the three vertices 2, 4, 6 and must permute the vertices 3 and 5. Furthermore, it is easy to see that **any** permutation of 2, 4, 6 and of 3, 5 will determine an automorphism of the graph which fixes vertex 1. Since there are $6 = 3!$ permutations of 2, 4, and 6, and since there are 2 permutations of 3 and 5, we conclude that there are exactly $6 \times 2 = 12$ automorphisms which fix the vertex 1. Therefore the full automorphism has order $6 \times 12 = 72$.

We turn now to the second graph considered in our introduction; again we draw two versions:



Graph D

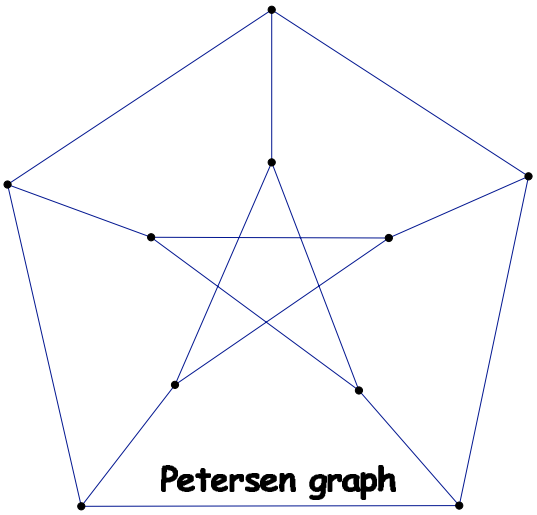
If H is the stabilizer of the vertex 1, then H must permute the three vertices 3, 4, 5 and must permute the vertices 2 and 6. However, in this case, there are some restrictions. Note that H must actually fix the vertex 4 (because there's an edge joining 3 and 5). Thus an automorphism $\tau \in H$ can only either fix the vertices 3 and 5 or can transpose them: $\tau(3) = 5$, $\tau(5) = 3$. However, once we know what τ does to $\{3, 4, 5\}$ we can determine its effect on 2 and 6. If τ fixes 3 and 5, then it's easy to see that τ also fixes 2 and 6 (verify this!). Likewise, if τ transposes 3 and 5, then τ also transposes 2 and 6, meaning that there are only two elements in H, e and the element

$$\tau = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ 1 & 6 & 5 & 4 & 3 & 2 \end{pmatrix}.$$

From the above we conclude that the automorphism group of this graph has $6 \times 2 = 12$ elements, meaning that the first graph is **six times more symmetrical than the second**!

EXERCISES

1. Compute $|G|$, where G is the automorphism group of Graph A, above. Is G abelian?
2. Compute $|G|$, where G is the automorphism group of Graph B, above. Is G abelian?
3. Find an element of order 6 in the stabilizer of vertex 1 of Graph C, above.
4. As a challenge, compute the order of the automorphism group of the Petersen graph.[15]



[15]The answer is 120=5!. Indeed, the automorphism group is isomorphic with the symmetric group S_5. Here's a possible approach. Construct the graph Γ whose vertices are the 2-element subsets of $\{1, 2, 3, 4, 5\}$ and where we stipulate that A and B are adjacent precisely when $A \cap B = \emptyset$. One shows first that this graph is actually isomorphic with the Petersen graph. (Just draw a picture!) Next, if we let S_5 operate on the elements of $\{1, 2, 3, 4, 5\}$ in the natural way, then S_5 actually acts as a group of automorphisms of Γ.

Chapter 5

Series and Differential Equations

The methods and results of this chapter pave the road to the students' more serious study of "mathematical analysis," that branch of mathematics which includes calculus and differential equations. It is assumed that the student has had a backgroud in calculus at least equivalent with that represented either in IB mathemtics HL year 2 or AP Calculus (AB). The key ideas revolving around limits will be reviewed, leading to substantial coverage of series and differential equations.

5.1 Quick Survey of Limits

As quickly becomes obvious to even the causual learner, the study of calculus rests in a fundamental way on the notion of limit. Thus, a reasonable starting point in this somewhat more "advanced" study is to be reminded of the notion of the "limit of a function as x approaches a (either of which might be $\pm\infty$)."

5.1.1 Basic definitions

DEFINITION. *Let f be a function defined in a neighborhood of the real number a (except possibly at $x = a$). We say that the* **limit** *of $f(x)$ is L as x* **approaches** *a, and write*

$$\lim_{x \to a} f(x) \;=\; L,$$

if for any real number $\epsilon > 0$, there is another real number $\delta > 0$ (which in general depends on ϵ) such that whenever $0 < |x - a| < \delta$ then

$|f(x) - L| < \epsilon$.

Notice that in the above definition we stipulate $0 < |x - a| < \delta$ rather than just saying $|x - a| < \delta$ because we really don't care what happens when $x = a$.

In defining limits involving ∞ only slight modifications are necessary.

DEFINITION.

Limits at ∞. Let f be a function defined for all $x > N$. We say that We say that the **limit** *of $f(x)$ is L as x* **approaches** *∞, and write*

$$\lim_{x \to \infty} f(x) \;=\; L,$$

if for any real number $\epsilon > 0$, there is another real number K (which in general depends on ϵ) such that whenever $x > K$ then $|f(x) - L| < \epsilon$.

In an entirely similarly way, we may define $\lim_{x \to -\infty} f(x) \;=\; L$.

Limits of ∞. Let f be a function defined in a neighborhood of the real number a. We say that the **limit** *of $f(x)$ is L as x approaches ∞, and write*

$$\lim_{x \to a} f(x) \;=\; \infty,$$

if for any real number N, there is another real number $\delta > 0$ (which in general depends on N) such that whenever $0 < |x - a| < \delta$ then $|f(x)| > N$.

Similarly, one defines

$$\lim_{x \to a} f(x) \;=\; -\infty, \quad \lim_{x \to \infty} f(x) \;=\; \infty,$$

and so on.

Occasionally, we need to consider **one-sided limits**, defined as follows.

Definition. *Let f be a function defined on an interval of the form $a < x < b$. We say that the* **limit** *of $f(x)$ is L as x* **approaches** *a* **from the right**, *and write*

$$\lim_{x \to a^+} f(x) \;=\; L,$$

if for any real number $\epsilon > 0$, there is another real number $\delta > 0$ (which in general depends on ϵ) such that whenever $0 < x - a < \delta$ then $|f(x) - L| < \epsilon$.

Similarly, one defines $\lim_{x \to a^-} f(x) \;=\; L$.

Limits behave in a very reasonable manner, as indicated in the following theorem.

Theorem. *Let f and g be functions defined in a punctured neighborhood of a.*[1] *Assume that*

$$\lim_{x \to a} f(x) \;=\; L, \quad \lim_{x \to a} g(x) \;=\; M.$$

Then,

$$\lim_{x \to a}(f(x) + g(x)) \;=\; L + M, \quad \textit{and} \quad \lim_{x \to a} f(x)g(x) \;=\; LM.$$

Proof. Assume that $\delta > 0$ has been chosen so as to guarantee that whenever $0 < |x - a| < \delta$, then

$$|f(x) - L| < \frac{\epsilon}{2}, \quad \text{and} \quad |g(x) - M| < \frac{\epsilon}{2}.$$

Then,

$$|f(x) + g(x) - (L + M)| < |f(x) - L| + |g(x) - M| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,$$

proving that $\lim_{x \to a}(f(x) + g(x)) \;=\; L + M$.

[1] A "punctured" neighborhood of the real number a is simply a subset of the form $\{x \in \mathcal{R} \,|\, 0 < |x - a| < d$, for some positive real number d.

Next, assume that $1 > \epsilon > 0$, and let $\delta > 0$ be a real number such that whenever $0 < |x - a| < \delta$,

$$|f(x) - L| < \frac{\epsilon}{3}, \frac{\epsilon}{3|M|}, \quad |g(x) - M| < \frac{\epsilon}{3}, \frac{\epsilon}{3|L|}.$$

(If either of L, $M = 0$, ignore the corresponding fraction $\frac{\epsilon}{3|L|}$, $\frac{\epsilon}{3|M|}$)

$$\begin{aligned} |f(x)g(x) - LM| &= |(f(x) - L)(g(x) - M) + (f(x) - L)M \\ &\quad +(g(x) - M)L| \\ &\leq |(f(x) - L)(g(x) - M)| + |(f(x) - L)M| \\ &\quad +|(g(x) - M)L| \\ &< \frac{\epsilon^2}{9} + \frac{\epsilon}{3} + \frac{\epsilon}{3} \\ &< \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon, \end{aligned}$$

proving that $\lim_{x \to a} f(x)g(x) = LM$.[2]

As indicated above, in computing $\lim_{x \to a} f(x)$ we are not concerned with $f(a)$; in fact a need not even be in the domain of f. However, in defining **continuity** at a point, we require more:

DEFINITION. *Let f be defined in a neighborhood of a. We say that f is* **continuous** *at $x = a$ if*

$$\lim_{x \to a} f(x) = f(a)$$

As a simply corollary of the above theorem we may conclude that polynomial functions are everywhere continuous.

The student will recall that the **derivative** of a function is defined in terms of a limit. We recall this important concept here.

[2]Note that in the above proof we have repeatedly used the so-called "triangle inequality," which states that for real numbers a, $b \in \mathcal{R}$, $|a + b| \leq |a| + |b|$. A moment's thought reveals that this is pretty obvious!

DEFINITION. *Let f be a function defined in a neighborhood of a. If*

$$\lim_{x \to a} \frac{f(x) - f(a)}{x - a} = L,$$

we say that f is **differentiable** *at $x = a$ and write $f'(a) = L$, calling $f'(a)$ the* **derivative of f at** a.

In mathematical analysis we often encounter the notion of a **sequence**, which is nothing more than a function

$$f : \{0, 1, 2, \ldots\} \longrightarrow \mathcal{R}.$$

It is customary to write the individual terms of a sequence $f(0)$, $f(1)$, $f(2), \ldots$ as subscripted quantities, say, as a_0, a_1, $a_2, \ldots$.

Sequences may or may not have limits.

DEFINITION. *Let $(a_n)_{n \geq 0}$ be a sequence. We say that the* **limit** *of the sequence is the real number $L \in \mathcal{R}$ and write $\lim_{n \to \infty} a_n = L$, if given $\epsilon > 0$ there exists a real number N such that whenever $n > N$ then $|a_n - L| < \epsilon$.*

We shall begin a systematic study of sequences (and "series") in the next section.

Finally, we would like to give one more example of a limiting process: that associated with the "Riemann integral." Here we have a function f defined on the closed interval $[a, b]$, and a **partition** P of the interval into n subintervals

$$P : \ a = x_0 < x_1 < x_2 < \cdots < x_n = b.$$

On each subinterval $[x_{i-1}, x_i]$ let

$$M_i = \max_{x_{i-1} < x < x_i} f(x), \quad m_i = \min_{x_{i-1} < x < x_i} f(x).$$

The **upper Riemann sum** relative to the above partition is the sum

$$U(f;P) \;=\; \sum_{i=i}^{n} M_i(x_i - x_{i-1}),$$

and the **lower Riemann sum** relative to the above partition is the sum

$$L(f;P) \;=\; \sum_{i=i}^{n} m_i(x_i - x_{i-1}).$$

Before continuing, we need two more fundamental concepts, the **least upper bound** and **greatest lower bound** of a set of real numbers. Namely, if $A \subseteq \mathcal{R}$, we set

$$\text{LUB}(A) = \min_d\{d \geq a \,|\, a \in A\}, \text{GLB}(A) = \max_d\{d \leq a \,|\, a \in A\}.$$

Finally, we define the sets

$$U(f) = \{U(f;P) \,|\, P \text{ is a partition of } [a,b]\},$$
$$L(f) = \{L(f;P) \,|\, P \text{ is a partition of } [a,b]\}.$$

DEFINITION. *If* $\text{LUB}(L(f))$ *and* $\text{GLB}(U(f))$ *both exist, and if* $\text{LUB}(L(f)) = \text{GLB}(U(f))$, *we say that* f *is* **Riemann integrable over** $[a,b]$ *and call the common value the* **Riemann integral of** f **over the interval** $[a,b]$.

EXAMPLE. Consider the function $f(x) = x^3,\ 0 \leq x \leq 2$, and consider the partition of $[0,2]$ into n equally-spaced subintervals. Thus, if $P:\ 0 = x_0 < x_1 < \cdots x_n = 2$ is this partitiion, then $x_i = \frac{2i}{n},\ i = 0,1,2,\ldots,n$. Since f is increasing over this interval, we see that the maximim of f over each subinterval occurs at the right endpoint and that the minimum occurs at the left endpoint. It follows, therefore, that

$$U(f;P) = \sum_{i=1}^{n}\left(\frac{2i}{n}\right)^3 \cdot \frac{2}{n} = \frac{16}{n^4}\sum_{i=1}^{n} i^3, \quad L(f;P) = \sum_{i=0}^{n-1}\left(\frac{2i}{n}\right)^3 \cdot \frac{2}{n} = \frac{16}{n^4}\sum_{i=0}^{n-1} i^3.$$

Next, one knows that $\sum_{i=1}^{n} i^3 = \frac{1}{4}n^2(n+1)^2$; therefore,

$$U(f;P) = \frac{4n^2(n+1)^2}{n^4}, \qquad L(f;P) = \frac{4n^2(n-1)^2}{n^4}.$$

Finally, we note that for any partition P' of $[0,2]$ $0 < L(f;P') < U(f;P') < 16$, and so it is clear that $\mathrm{GLB}(U(f))$ and $\mathrm{LUB}(L(f))$ both exist and that $\mathrm{GLB}(U(f)) \geq \mathrm{LUB}(L(f))$. Finally, for the partition P above we have

$$L(f;P) \leq \mathrm{LUB}(L(f)) \leq \mathrm{GLB}(U(f)) \leq U(f;P).$$

Therefore, we have

$$4 = \lim_{n\to\infty} L(f;P) \leq \mathrm{LUB}(L(f)) \leq \mathrm{GLB}(U(f)) \leq \lim_{n\to\infty} L(f;P) = 4,$$

and so it follows that

$$\int_0^2 x^3\,dx \;=\; 4.$$

For completeness' sake, we present the following fundamental result without proof.

Theorem. (Fundamental Theorem of Calculus) *Assume that we are given the function f defined on the interval $[a,b]$. If there exists a differentiable function F also defined on $[a,b]$ and such that $F'(x) = f(x)$ for all $x \in [a,b]$, then f is Riemann integral on $[a,b]$ and that*

$$\int_a^b f(x)\,dx \;=\; F(b) - F(a).$$

Exercises

1. Let f and g be functions such that

   (a) f is defined in a punctured neighborhood of a,

(b) $\lim_{x\to a} f(x) = L$,

(c) g is defined in a punctured neighborhood of L, and

(d) $\lim_{x\to L} g(x) = M$.

Show that $\lim_{x\to a} g(f(x)) = M$.

2. Show that if $a \geq 0$, then $\lim_{x\to a} \sqrt{x} = \sqrt{a}$.

3. Show that if $a \in \mathcal{R}$, then $\lim_{x\to a} \sqrt{1+x^2} = \sqrt{1+a^2}$.

4. Prove that the sequence 1, 0, 1, 0, ... does not have a limit.

5. Let $f : D \to \mathcal{R}$ be a real-valued function, where D (the domain) is some subset of $\mathcal{R}$. Prove that f is continuous at $a \in D$ if and only if for every sequence $a_1, a_2, \ldots$, which converges to a, then $\lim_{n\to\infty} f(a_n) = f(a)$.

6. Here we revisit Exercises 11 and 12 on page 239.

   (a) Let $f : \mathcal{R} \to \mathcal{R}$ be a differentiable homomorphism, i.e., f is differentiable and satisfies $f(x+y) = f(x)+f(y)$ for all $x, y \in \mathcal{R}$. Prove that there exists $c \in \mathcal{R}$ such that $f(x) = cx,\ x \in \mathcal{R}$. (This is easy!)

   (b) Let $f : \mathcal{R} \to \mathcal{R}$ be a continuous homomorphism, and prove that the same conclusion of part (a) holds. (Hint: you want to prove that for all $a, x \in \mathcal{R}$, $f(ax) = af(x)$. This guarantees that $f(x) = xf(1)$.)[3]

7. Define $f(x) = \sqrt{x + \sqrt{x + \sqrt{x + \cdots}}},\ x \geq 0$. Clearly $f(x) = 0$. Is f continuous at $x = 0$? (Hint: note that if we set $y = f(x)$, then $y = \sqrt{x+y}$, and so $y^2 = x + y$. Using the quadratic formula you can solve for y in terms of x.)

8. Recall Euler's ϕ-function ϕ, given on page 63. Define the set A of real numbers

$$A = \left\{ \frac{\phi(n)}{n} \mid n \in \mathbb{N} \right\}.$$

[3]It may surprise you to learn that "most" homomorphisms $\mathcal{R} \to \mathcal{R}$ are **not** continuous. However, the discontinuous homomorphisms are essentially impossible to write down. (Their existence involves a bit of advanced mathematics: using a "Hamel basis" for the real numbers, one can easily write down discontinuous homomorphisms $\mathcal{R} \to \mathcal{R}$.)

Show that $\text{LUB}(A) = 1$ and $\text{GLB}(A) = 0$.[4]

9. (The irrationality of π.) This exercise will guide you through a proof of the irrationality of π and follows roughly the proof of Ivan Niven[5]. Assume, by way of contradiction, that we may express $\pi = \dfrac{a}{b}$, where a and b are positive integers. For each integer $n \geq 1$ define the function

$$f_n(x) = \frac{x^n(a - bx)^n}{n!}.$$

Using the assumption that $\pi = \dfrac{a}{b}$, show that

(a) $f_n(0) = 0,\ f_n(\pi) = 0$;

(b) $f_n(x - \pi) = f_n(x)$;

(c) $f_n^{(i)}(0)$ is an integer for all $i \geq 0$;

(d) $f_n^{(i)}(\pi)$ is an integer for all $i \geq 0$ (use (b)).

Next, define the new functions

$$F_n(x) = f(x) - f^{(2)}(x) + f^{(4)}(x) - \cdots + (-1)^n f^{(2n)}(x)$$

[4]Showing that $\text{LUB}(A) = 1$ is pretty easy: define $P = \left\{ \frac{\phi(p)}{p} \mid p \text{ is prime} \right\}$ and show that $\text{LUB}(P) = 1$. Showing that $\text{GLB}(A) = 0$ is a bit trickier. Try this: let $p_1 = 2, p_2 = 3, p_3 = 5, \cdots$, and so p_k is the k-th prime. Note that

$$\frac{\phi(p_1 p_2 \cdots p_r)}{p_1 p_2 \cdots p_r} = \left(1 - \frac{1}{p_1}\right)\left(1 - \frac{1}{p_2}\right) \cdots \left(1 - \frac{1}{p_r}\right).$$

The trick is to show that $\dfrac{\phi(p_1 p_2 \cdots p_r)}{p_1 p_2 \cdots p_r} \to 0$ as $r \to \infty$. Next, one has the "harmonic series" (see page 265) $\displaystyle\sum_{n=1}^{\infty} \frac{1}{n}$, which is shown on page 265 to be infinite. However, from the Fundamental Theorem of Arithmetic (see page 76), one has

$$\begin{aligned} \sum_{n=1}^{\infty} \frac{1}{n} &= \left(1 + \frac{1}{2} + \frac{1}{2^2} + \cdots\right)\left(1 + \frac{1}{3} + \frac{1}{3^2} \cdots\right)\left(1 + \frac{1}{5} + \frac{1}{5^2} \cdots\right) \cdots \\ &= \left(1 - \frac{1}{2}\right)^{-1}\left(1 - \frac{1}{3}\right)^{-1}\left(1 - \frac{1}{5}\right)^{-1} \cdots \end{aligned}$$

From this one concludes that, as stated above, $\dfrac{\phi(p_1 p_2 \cdots p_r)}{p_1 p_2 \cdots p_r} \to 0$ as $r \to \infty$.

[5]Bull. Amer. Math. Soc. Volume 53, Number 6 (1947), 509.

$$= \sum_{i=0}^{n} (-1)^i f^{(2i)}(x), \quad n \geq 1,$$

and show that

(e) $F_n(0)$ and $F_n(\pi)$ are both integers;

(f) $F_n''(x) + F_n(x) = f(x)$ (note that if $i > 2n$, then $f_n^{(i)}(x) = 0$);

(g) $\dfrac{d}{dx}[F_n'(x)\sin x - F_n(x)\cos x] = f(x)\sin x$;

(h) $\displaystyle\int_0^{\pi} f_n(x)\sin x\,dx$ is an integer for all $n \geq 1$.

Finally, show that

(i) $f_n(x)\sin x > 0$ when $0 < x < \pi$;

(j) $f_n(x)\sin x < \dfrac{\pi^n a^n}{n!}$ when $0 < x < \pi$;

Conclude from (j) that

(k) $\displaystyle\int_0^{\pi} f_n(x)\sin x\,dx < \frac{\pi^{n+1}a^n}{n!}$ for all $n \geq 1$.

Why are (h) and (k) incompatible? (Note that $\displaystyle\lim_{n\to\infty} \frac{\pi^{n+1}a^n}{n!} = 0$.)

5.1.2 Improper integrals

There are two type of **improper intergrals** of concern to us.

(I.) Those having at least one infinite limit of integration, such as

$$\int_a^{\infty} f(x)\,dx \quad \text{or} \quad \int_{-\infty}^{\infty} f(x)\,dx.$$

(II.) Those for which the integrand becomes unbounded within the interval over which the integral is computed. Examples of these include

$$\int_0^1 \frac{dx}{x^p}, \quad (p > 0), \quad \int_1^3 \frac{dx}{x-2}.$$

The definitions of these improper integrals are in terms of limits. For example

$$\begin{aligned}\int_0^\infty f(x)\,dx &= \lim_{b\to\infty}\int_a^b f(x)\,dx\\ \int_{-\infty}^\infty f(x)\,dx &= \lim_{a\to-\infty}\int_a^0 f(x)\,dx+\lim_{b\to\infty}\int_0^b f(x)\,dx.\end{aligned}$$

Likewise, for example,

$$\begin{aligned}\int_0^1 \frac{dx}{x^p} &= \lim_{a\to0^+}\int_a^1 \frac{dx}{x^p},\\ \int_1^3 \frac{dx}{x-2} &= \lim_{a\to2^-}\int_1^a \frac{dx}{x-2}+\lim_{b\to2^+}\int_b^3 \frac{dx}{x-2}.\end{aligned}$$

Relative to the above definition, the following is easy.

Theorem. *We have*

$$\int_1^\infty \frac{dx}{x^p} = \begin{cases}\dfrac{1}{p-1} & \text{if } p>1\\ \infty & \text{if } p\le 1.\end{cases}.$$

Proof. We have, if $p\neq 1$, that

$$\begin{aligned}\int_1^\infty \frac{dx}{x^p} &= \lim_{a\to\infty}\frac{x^{1-p}}{1-p}\Big|_1^a = \lim_{a\to\infty}\left(\frac{a^{1-p}}{1-p}-\frac{1}{1-p}\right)\\ &= \begin{cases}\dfrac{1}{p-1} & \text{if } p>1\\ \infty & \text{if } p<1.\end{cases}\end{aligned}$$

If $p=1$, then

$$\int_1^\infty \frac{dx}{x} = \lim_{a\to\infty}\ln x\Big|_1^a = \lim_{a\to\infty}\ln a = \infty.$$

Example. Compute the improper integral $\displaystyle\int_2^\infty \frac{dx}{x\ln^p x}$, where $p>1$.

This is a fairly simple integration: using the substitution ($u = \ln x$) one first computes the indefinite integral

$$\int \frac{dx}{x \ln^p x} = \frac{1}{(1-p)x \ln^p x}.$$

Therefore,

$$\int_2^\infty \frac{dx}{x \ln^p x} = \lim_{a\to\infty} \left(\frac{1}{(1-p)x \ln^p x} \right) \Bigg|_2^a = \frac{1}{2(1-p)\ln^p 2}.$$

EXERCISES

1. For each improper integral below, compute its value (which might be $\pm\infty$) or determine that the integral does not exist.
   (a) $\displaystyle\int_2^\infty \frac{dx}{\sqrt{x-2}}$
   (b) $\displaystyle\int_{-1}^1 \frac{dx}{\sqrt{1-x^2}}$
   (c) $\displaystyle\int_2^\infty \frac{dx}{\sqrt{x^2-4}}$
   (d) $\displaystyle\int_0^1 \ln x \, dx$
2. Let $k \geq 1$ and $p > 1$ and prove that $\displaystyle\int_1^\infty \sin^k\left(\frac{2\pi}{x^p}\right) dx < \infty$. (Hint: note that if $x^p > 4$ then $\displaystyle\sin^k\left(\frac{2\pi}{x^p}\right) \leq \sin\left(\frac{2\pi}{x^p}\right) < \frac{2\pi}{x^p}$.)
3. Compute
$$\lim_{x\to\infty} \int_0^\infty e^{-t}\cos(xt)\, dt.$$
4. Let A, B be constants, $A > 0$. Show that
$$\int_0^\infty e^{-At} \sin Bt \, dt = \frac{B}{A^2 + B^2}.$$
5. Define the function
$$\Pi(x) = \begin{cases} 1 & \text{if } |x| \leq \frac{1}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

Now let f be a continuous function defined for all real numbers and compute

$$\lim_{T\to 0} \frac{1}{T} \int_{-\infty}^{\infty} \Pi\left(\frac{x-a}{T}\right) f(x)\, dx$$

in terms of f and a.

6. (The (real) **Laplace transform**) Let $f = f(x)$ be a function defined for $x \geq 0$. Define a new function $F = F(s)$, called the **Laplace transform** of f by setting $F(s) = \int_0^{\infty} e^{-sx} f(x)\, dx$, where $s \geq 0$. Now let f be the function defined by

$$f(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{if } x > 1. \end{cases}$$

Compute the Laplace transform $F = F(s)$ explicitly as a function of s.

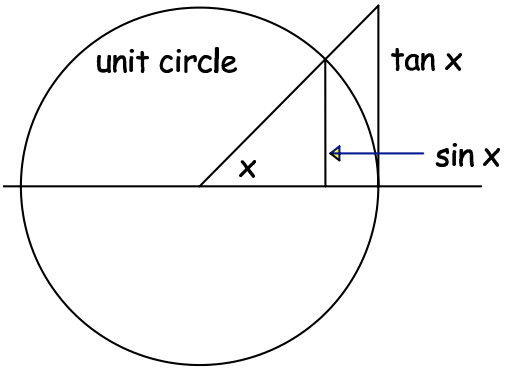
7. Let $f(x) = \sin 2\pi x,\ x \geq 0$. Compute the Laplace transform $F = F(s)$ explicitly as a function of s. (You'll need to do integration by parts twice!)

5.1.3 Indeterminate forms and l'Hôpital's rule

Most interesting limits—such as those defining the derivative—are "indeterminate" in the sense that they are of the form $\lim_{x\to a} \dfrac{f(x)}{g(x)}$ where the numerator and denominator both tend to 0 (or to ∞). Students learn to compute the derivatives of trigonometric functions only after they have been shown that the limit

$$\lim_{x\to 0} \frac{\sin x}{x} = 1.$$

At the same time, you'll no doubt remember that the computation of this limit was geometrical in nature and involved an analysis of the diagram to the right.



The above limit is called a **0/0 indeterminate form** because the limits of both the numerator and denominator are 0.

You've seen many others; here are two more:

$$\lim_{x\to 3} \frac{2x^2-7x+3}{x-3} \quad \text{and} \quad \lim_{x\to 1} \frac{x^5-1}{x-1}.$$

Note that in both cases the limits of the numerator and denominator are both 0. Thus, these limits, too, are 0/0 indeterminate forms.

While the above limits can be computed using purely algebraic methods, there is an alternative—and often quicker—method that can be used when algebra is combined with a little differential calculus.

In general, a **0/0 indeterminate form** is a limit of the form $\lim_{x\to a} \frac{f(x)}{g(x)}$ where both $\lim_{x\to a} f(x) = 0$ and $\lim_{x\to a} g(x) = 0$. Assume, in addition, that f and g are both differentiable and that f' and g' are both continuous at $x = a$ (a very reasonable assumption, indeed!). Then we have

$$\begin{aligned}
\lim_{x\to a} \frac{f(x)}{g(x)} &= \lim_{x\to a} \frac{\left(\frac{f(x)}{x-a}\right)}{\left(\frac{g(x)}{x-a}\right)} \\
&= \frac{\lim_{x\to a} \left(\frac{f(x)}{x-a}\right)}{\lim_{x\to a} \left(\frac{g(x)}{x-a}\right)} \\
&= \frac{f'(a)}{g'(a)} \\
&= \frac{\lim_{x\to a} f'(x)}{\lim_{x\to a} g'(x)}. \qquad \text{(by continuity of the derivatives)}
\end{aligned}$$

This result we summarize as

l'Hôpital's Rule. *Let f and g be functions differentiable on some interval containing $x = a$, and assume that f' and g' are continuous at $x = a$. Then*

$$\boxed{\lim_{x\to a} \frac{f(x)}{g(x)} = \frac{\lim\limits_{x\to a} f'(x)}{\lim\limits_{x\to a} g'(x)}}.$$

As a simple illustration, watch this:

$$\lim_{x\to 3} \frac{2x^2 - 7x + 3}{x - 3} = \frac{\lim\limits_{x\to 3} 4x - 7}{\lim\limits_{x\to 3} 1} = 5,$$

which agrees with the answer obtained algebraically.

In a similar manner, one defines $\boldsymbol{\infty/\infty}$ indeterminate forms; these are treated as above, namely by differentiating numerator and denominator:

l'Hôpital's Rule (∞/∞). *Let f and g be functions differentiable on some interval containing $x = a$, that $\lim\limits_{x\to a} f(x) = \pm\infty = \lim\limits_{x\to a} g(x)$, and assume that f' and g' are continuous at $x = a$. Then*

$$\boxed{\lim_{x\to a} \frac{f(x)}{g(x)} = \frac{\lim\limits_{x\to a} f'(x)}{\lim\limits_{x\to a} g'(x)}}.$$

There are other indeterminate forms as well: $\boldsymbol{0\cdot\infty}$, $\boldsymbol{1^\infty}$, and $\boldsymbol{\infty^0}$. These can be treated as indicated in the examples below.

Example 1. Compute $\lim\limits_{x\to 0^+} x^2 \ln x$. Note that this is a $\boldsymbol{0\cdot\infty}$ indeterminate form. It can easily be converted to an $\frac{\infty}{\infty}$ indeterminate form and handled as above:

$$\lim_{x\to 0^+} x^2 \ln x \;=\; \lim_{x\to 0^+} \frac{\ln x}{(1/x^2)} \overset{l'H}{=} \lim_{x\to 0^+} \frac{1/x}{-2/x^3} = \lim_{x\to 0^+} \frac{-x^2}{2} = 0.$$

Other indeterminate forms can be treated as in the following examples.

EXAMPLE 2. Compute $\lim_{x\to\infty}\left(1-\frac{4}{x}\right)^x$. Here, if we set L equal to this limit (if it exists!), then we have, by continuity of the logarithm, that

$$\begin{aligned}
\ln L &= \ln \lim_{x\to\infty}\left(1-\frac{4}{x}\right)^x \\
&= \lim_{x\to\infty} \ln\left(1-\frac{4}{x}\right)^x \\
&= \lim_{x\to\infty} x\ln\left(1-\frac{4}{x}\right) \\
&= \lim_{x\to\infty} \frac{\ln\left(1-\frac{4}{x}\right)}{1/x} \\
&\overset{l'H}{=} \lim_{x\to\infty} \frac{4/\left(x^2\left(1-\frac{4}{x}\right)\right)}{-1/x^2} \\
&= \lim_{x\to\infty} \frac{-4}{\left(1-\frac{4}{x}\right)} = -4
\end{aligned}$$

This says that $\ln L = -4$ which implies that $L = e^{-4}$.

EXAMPLE 3. This time, try $\lim_{\theta\to(\pi/2)^-}(\cos\theta)^{\cos\theta}$. The same trick applied above works here as well. Setting L to be this limit, we have

$$\begin{aligned}
\ln L &= \ln \lim_{\theta\to(\pi/2)^-}(\cos\theta)^{\cos\theta} \\
&= \lim_{\theta\to(\pi/2)^-} \ln(\cos\theta)^{\cos\theta} \\
&= \lim_{\theta\to(\pi/2)^-} \cos\theta\ln\cos\theta \\
&= \lim_{\theta\to(\pi/2)^-} \frac{\ln\cos\theta}{1/\cos\theta} \\
&\overset{l'H}{=} \lim_{\theta\to(\pi/2)^-} \frac{\tan\theta}{\sec\theta\tan\theta} = 0.
\end{aligned}$$

It follows that $\lim_{\theta\to(\pi/2)^-}(\cos\theta)^{\cos\theta} = 1$.

Exercises

1. Using l'Hôpital's rule if necessary, compute the limits indicated below:

   (a) $\lim_{x\to 1} \dfrac{x^3-1}{4x^3-x-3}$

   (b) $\lim_{x\to 1} \dfrac{\cos(\pi x/2)}{\sqrt[3]{(x-1)^2}}$

   (c) $\lim_{x\to\infty} \dfrac{2x^2-5x}{x^3-x+10}$

   (d) $\lim_{\theta\to 0} \dfrac{\sin 3\theta}{\sin 4\theta}$

   (e) $\lim_{\theta\to 0} \dfrac{\sin\theta^2}{\theta}$

   (f) $\lim_{\theta\to \pi/2} \dfrac{1-\sin\theta}{1+\cos 2\theta}$

   (g) $\lim_{x\to\infty} \dfrac{\ln(x+1)}{\log_2 x}$

   (h) $\lim_{x\to 0^+} (\ln x - \ln\sin x)$ (Hint: you need to convert this "$\infty - \infty$" indeterminate form to one of the forms discussed above!)

   (i) $\lim_{x\to\infty} (\ln 2x - \ln(x+1))$.

   (j) $\lim_{x\to\infty} \left(1+\dfrac{4}{x}\right)^x$

   (k) $\lim_{x\to\infty} \left(1+\dfrac{a}{x}\right)^x$

   (l) $\lim_{x\to 1} x^{1/(x-1)}$

   (m) $\lim_{x\to\infty} x^3 e^{-x}$

   (n) $\lim_{x\to 0^+} x^a e^{-x}, \quad a>0$

   (o) $\lim_{x\to 0^+} \ln x \ln(1-x)$

   (p) $\lim_{x\to 1^-} \ln x \ln(1-x)$ (Are (o) and (p) really different?)

2. Compute $\int_0^\infty xe^{-2x}\,dx$.

3. Let n be a non-negative integer. Using mathematical induction, show that $\int_0^\infty x^n e^{-x}\,dx = n!$.

4. (The (real) **Gamma function**) Let $z>0$ and define $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}\,dx$. Show that

   (a) $\Gamma(n) = (n-1)!$ for any positive integer n;

   (b) $\Gamma(z)$ exists (i.e., the improper integral converges) for all $z>0$.

5. (Convolution) Given functions f and g defined for all $x\in\mathcal{R}$, the **convolution** of f and g is defined by

$$f*g(x) \;=\; \int_{-\infty}^{\infty} f(t)g(x-t)\,dt,$$

provided the improper integral exists. Assuming that $f * g(x)$ exists for all x, show that "$*$" is commutative, i.e., that

$$f * g(x) = g * f(x), \quad \text{for all } x.$$

We shall meet the convolution again in our study of statistics (page 370).

6. Let $a > 0$ be a constant, and set

$$f(x) \;=\; \begin{cases} e^{-at} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases}$$

Show that if $g(x)$ is defined for all $x \in \mathcal{R}$, then

$$f * g(x) \;=\; e^{-ax} \int_{-\infty}^{x} g(t) e^{at}\, dx$$

provided the improper integral exists.

Now compute $f * g(x)$, where g is as above and where

(a) $f(x) = \sin bx$, where $b > 0$ is a constant.

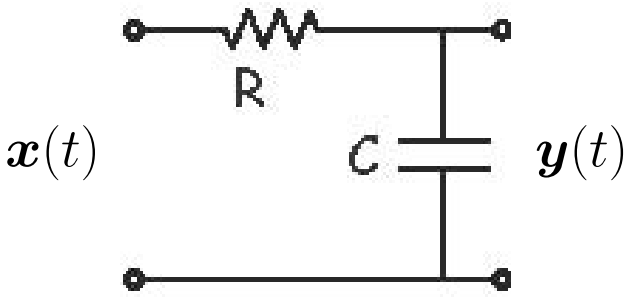
(b) $f(x) = x^2$.

(c) $f(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases}$

(d) $f(x) = \begin{cases} \sin bx & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$ where $b > 0$ is a constant.

7. (Convolution and the Low-Pass Filter) In electrical engineering one frequently has occasion to study the RC low-pass filter, whose schematic diagram is shown below. This is a "series" circuit with a resistor having resistance R Ω ("ohms") and a capacitor with capacitance C F ("Farads"). An input voltage of $x(t)$ volts is applied at the input terminals and the voltage $y(t)$ volts is observed at the output. The variable t represents time, measured in seconds.

An important theorem of electrical engineering is that if the input voltage is $x(t)$, then the output voltage is $y = x*h(t)$, where h (the "impulse response") is given explicitly by



$$h(t) = \begin{cases} \frac{1}{\tau}e^{-t/\tau} & \text{if } t \geq 0 \\ 0 & \text{if } t < 0, \end{cases}$$

and where $\tau = RC$.

Now assume that $R = 1000\,\Omega$ and that $C = 2\,\mu\text{F}$ $(= 2 \times 10^{-6}$ Farads). Let

$$x(t) = \begin{cases} \sin 2\pi f t & \text{if } t \geq 0 \\ 0 & \text{if } t < 0 \end{cases}$$

where $f > 0$ is the frequency of the signal (in "hertz" (Hz) or units of $(\text{sec})^{-1}$). In each case below, compute and graph the output voltage $y(t)$ as a function of time:

(a) $f = 100\,\text{Hz}$

(b) $f = 2$ kHz, or 2000 Hz

(c) $f = 100$ kHz

8. (For the courageous student![6]) Consider the function

$$f(t) = \begin{cases} \sin(1/t), & \text{if } t \neq 0 \\ 5 \text{ if } t = 0, \end{cases}$$

and set $F(x) = \int_{-1}^{x} f(t)\,dt$. Show that $F'(0) = 0$. (Hint: try carrying out the following steps:

[6] I am indebted to Robert Burckel for suggesting this problem.

(a) If $G(x) = \int_0^x f(t)\,dt$, then $F'(0) = G'(0)$.

(b) Show that G is an **odd** function, i.e., $G(-x) = -G(x)$.

(c) Use integration by parts to show that if $0 < y < x$, then

$$\begin{aligned}\int_y^x f(t)\,dt &= \int_y^x \frac{t^2 \sin(1/t)\,dt}{t^2} \\ &= x^2\cos(1/x) - y^2\cos(1/y) + \int_y^x 2t\,dt \le 3x^2.\end{aligned}$$

(d) Using part (c), show that for all x, $|G(x)| \le 3x^2$.

(e) Conclude from part (d) that $G'(0) = 0$.)

5.2 Numerical Series

Way back in **Algebra II** you learned that certain **infinite series** not only made sense, you could actually compute them. The primary (if not the only!) examples you learned were the **infinite geometric series**; one such example might have been

$$3 + \frac{3}{4} + \frac{3}{16} + \frac{3}{64} + \cdots = \sum_{n=0}^{\infty} \frac{3}{4^n}.$$

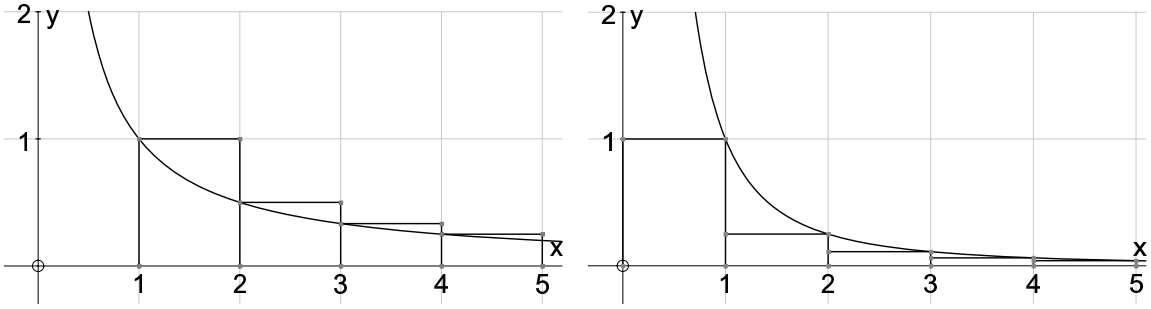
Furthemore, you even learned how to compute such infinite geometric series; in the above example since the **first term** is $a = 3$ and since the **ratio** is $r = \frac{1}{4}$, you quickly compute the sum:

$$3 + \frac{3}{4} + \frac{3}{16} + \frac{3}{64} + \cdots = \sum_{n=0}^{\infty} \frac{3}{4^n} \;=\; \boxed{\frac{a}{1-r} = \frac{3}{1-\frac{1}{4}} = 4}.$$

Perhaps unfortunately, most infinite series are not geometric but rather come in a variety of forms. I'll give two below; they seem similar but really exhibit very different behaviors:

Series 1: $$1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots$$

Series 2: $$1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \cdots .$$



To see how the above two series differ, we shall consider the above diagrams. The picture on the left shows that the area represented by the sum $1 + \frac{1}{2} + \frac{1}{3} + \cdots$ is **greater** than the area under the curve with equation $y = 1/x$ from 1 to ∞. Since this area is

$$\int_1^\infty \frac{dx}{x} \;=\; \ln x \Big|_1^\infty = \infty,$$

we see that the infinite series $1 + \frac{1}{2} + \frac{1}{3} + \cdots$ must **diverge** (to infinity). This divergent series is often called the **harmonic series**. (This terminology is justified by Exercise 20 on page 109.) Likewise, we see that the series $1 + \frac{1}{2^2} + \frac{1}{3^2} + \cdots$ can be represented by an area that is $\leq 1 + \int_1^\infty \frac{dx}{x^2} = 1 - \frac{1}{x}\Big|_1^\infty = 2$, which shows that this series cannot diverge to ∞ and so **converges** to some number.[7]

5.2.1 Convergence/divergence of non-negative term series

Series 2 in the above discussion illustrates an important principle of the real numbers. Namely, if a_0, a_1, a_2, ... is a **sequence** of real numbers such that

(i) $a_0 \leq a_1 \leq a_2 \leq \ldots$, and

(ii) there is an **upper bound** M for each element of the sequence, i.e., $a_n \leq M$ for each $n = 0, 1, 2, \ldots$,

[7]It turns out that this series converges to $\frac{\pi^2}{6}$; this is not particularly easy to show.

then the sequence **converges** to some limit L (which we might not be able to compute!): $\lim_{n\to\infty} a_n = L$.

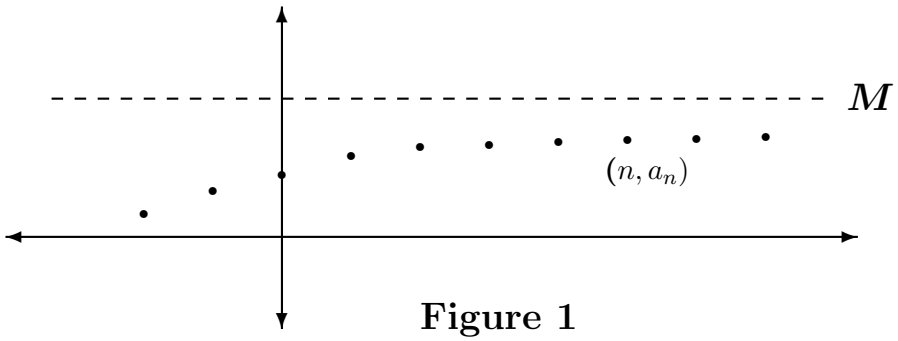


Figure 1

So what do **sequences** have to do with **infinite series**? Well, this is simple: if each term a_n in the infinite series $\sum_{n=0}^{\infty} a_n$ is non-negative, then the **sequence of partial sums** satisfies

$$a_0 \le a_0 + a_1 \le a_0 + a_1 + a_2 \le \cdots \le \sum_{n=0}^{k} a_n \le \cdots .$$

Furthermore, if we can establish that for some M each partial sum $S_k = \sum_{n=0}^{k} a_n$ satisfies $S_k \le M$ then we have a limit, say, $\lim_{k\to\infty} S_k = L$, in which case we write

$$\sum_{n=0}^{\infty} a_n = L.$$

In order to test a given infinite series $\sum_{n=0}^{\infty} a_n$ of non-negative terms for convergence, we need to keep in mind the following three basic facts.

Fact 1: In order for $\sum_{n=0}^{\infty} a_n$ to converge **it must happen that** $\lim_{n\to\infty} a_n = 0$. (Think about this: if the individual terms of the series don't get small, there's no hope that the series can converge. Furthermore, this fact remains true even when not all of the terms of the series are non-negative.)

Fact 2: If Fact 1 is true, we still need to show that there is some number M such that $\sum_{n=0}^{k} a_n \leq M$ for all k.

Fact 3: Even when we have verified Facts 1 and 2, we still might not (and usually won't) know the actual limit of the infinite series $\sum_{n=0}^{\infty} a_n$.

Warning about Fact 1: the requirement that $\lim_{n\to\infty} a_n = 0$ is a **necessary but not sufficient** condition for convergence. Indeed, in the above we saw that the series $\sum_{n=1}^{\infty} \frac{1}{n}$ **diverges** but that $\sum_{n=1}^{\infty} \frac{1}{n^2}$ **converges**.

EXERCISES

1. Apply **Fact 1** above to determine those series which definitely will not converge.

   (a) $\sum_{n=0}^{\infty} \frac{n}{n+1}$ (d) $\sum_{n=0}^{\infty} \frac{n}{(\ln n)^2}$ (g) $\sum_{n=2}^{\infty} \frac{(-1)^n n^2}{2^n}$

   (b) $\sum_{n=0}^{\infty} \frac{(-1)^n}{n}$ (e) $\sum_{n=1}^{\infty} \frac{\sin n}{n}$ (h) $\sum_{n=0}^{\infty} \frac{n!}{2^n}$

   (c) $\sum_{n=2}^{\infty} \frac{\ln n}{n}$ (f) $\sum_{n=0}^{\infty} (-1)^n \sin n$ (i) $\sum_{n=2}^{\infty} \frac{\ln n}{\ln(n^2+1)}$

2. Occassionally an infinite series can be computed by using a partial fraction decomposition. For example, note that $\frac{1}{n(n+1)} = \frac{1}{n} - \frac{1}{n+1}$ and so

$$\begin{aligned} \sum_{n=1}^{\infty} \frac{1}{n(n+1)} &= \sum_{n=1} \left(\frac{1}{n} - \frac{1}{n+1} \right) \\ &= \left(1 - \frac{1}{2}\right) + \left(\frac{1}{2} - \frac{1}{3}\right) + \left(\frac{1}{3} - \frac{1}{4}\right) + \cdots = 1. \end{aligned}$$

   Such a series is called a "telescoping series" because of all the internal cancellations. Use the above idea to compute

(a) $\displaystyle\sum_{n=0}^{\infty} \frac{1}{4n^2-1}$ (b) $\displaystyle\sum_{n=0}^{\infty} \frac{3}{9n^2-3n-2}$

3. Consider the series

$$\Sigma = \sum \left\{ \frac{1}{n} \,\middle|\, \text{the integer } n \text{ doesn't contain the digit } 0 \right\}.$$

Therefore, the series Σ contains reciprocals of integers, except that, for example, 10 is thrown out, as is 20, as is 100, 101, etc. No 0s are allowed! Determine whether this series converges. (Hint:

$$\begin{aligned} \Sigma &= 1 + \frac{1}{2} + \cdots + \frac{1}{9} \\ &+ \frac{1}{11} + \frac{1}{12} + \cdots + \frac{1}{19} + \frac{1}{21} + \cdots + \frac{1}{99} \\ &+ \frac{1}{111} + \frac{1}{112} + \cdots + \frac{1}{999} \\ &+ \cdots \\ &< 9 + \frac{9^2}{10} + \frac{9^3}{100} + \cdots .) \end{aligned}$$

4. (Formal definition of e) Consider the sequence $a_n = \left(1 + \frac{1}{n}\right)^n$, $n = 1, 2, \ldots$.

   (a) Use the binomial theorem to show that $a_n < a_{n+1}$, $n = 1, 2, \ldots$. (Note that in the expansions of a_n and a_{n+1}, the latter has one additional term. Moreover, the terms of a_n can be made to correspond to terms of a_{n+1} with each of the terms of the latter being larger.)

   (b) Show that, for each positive n, $a_n < 1+1+\frac{1}{2!}+\frac{1}{3!}+\cdots+\frac{1}{n!} < 3$.

   (c) Conclude that $\lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n$ exists. The limit is the familiar natural exponential base, e, and is often taken as the formal definition.

   (d) Show that for any real number x, $\lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n = e^x$. (Hint: note that $\lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n = \lim_{m\to\infty} \left(1 + \frac{1}{m}\right)^{mx}$.)

5. Prove that the limit $\lim_{n\to\infty}\left(\sum_{k=1}^{n}\frac{1}{k}-\ln n\right)$ exists; its limit is called **Euler's constant**[8] and is denoted γ (≈ 0.577). To prove this just draw a picture, observing that the sequence $\frac{1}{2} < \sum_{k=1}^{n}\frac{1}{k}-\ln n < 1$ for all n and that the sequence $a_n = \sum_{k=1}^{n}\frac{1}{k}-\ln n$ is an **decreasing** sequence.[9]

5.2.2 Tests for convergence of non-negative term series

In this subsection we'll gather together a few handy tests for convergence (or divergence). They are pretty intuitive, but still require practice.

The Limit Comparison Test

(i) Let $\sum_{n=0}^{\infty} a_n$ *be a* **convergent** *series of positve terms and let* $\sum_{n=0}^{\infty} b_n$ *be a second series of positive terms. If for some* R, $0 \le R < \infty$

$$\lim_{n\to\infty}\frac{b_n}{a_n} = R,$$

then $\sum_{n=0}^{\infty} b_n$ *also* **converges**. *(This is reasonable as it says that*

[8] or sometimes the **Euler-Mascheroni** constant

[9] Drawing a picture shows that

$$1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{n-1}-\ln n \ge \frac{1}{4}+\frac{1}{2}\left(\frac{1}{2}-\frac{1}{3}\right)+\frac{1}{2}\left(\frac{1}{3}-\frac{1}{4}\right)+\cdots+\frac{1}{2}\left(\frac{1}{n-1}-\frac{1}{n}\right)=\frac{1}{2}-\frac{1}{2n}.$$

Therefore,

$$1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{n}-\ln n \ge \frac{1}{2}+\frac{1}{2n} > \frac{1}{2}.$$

Next, that the sequence is decreasing follows from the simple observation that for all $n > 0$, $\frac{1}{n+1} < \ln\left(\frac{n+1}{n}\right)$.

Finally, I'd like to mention in passing that, unlike the famous mathematical constants π and e (which are not only irrational but actually transcendental), it is not even known whether γ is rational or irrational.

asymptotically the series $\sum_{n=0}^{\infty} b_n$ *is no larger than* R *times the convergent series* $\sum_{n=0}^{\infty} a_n$.*)*

(ii) Let $\sum_{n=0}^{\infty} a_n$ *be a* **divergent** *series of positve terms and let* $\sum_{n=0}^{\infty} b_n$ *be a second series of positive terms. If for some* $R,\ 0 < R \leq \infty$

$$\lim_{n\to\infty} \frac{b_n}{a_n} = R,$$

then $\sum_{n=0}^{\infty} b_n$ *also* **diverges**. *(This is reasonable as it says that asymptotically the series* $\sum_{n=0}^{\infty} b_n$ *is at least* R *times the divergent series* $\sum_{n=0}^{\infty} a_n$.*)*

Let's look at a few examples! Before going into these examples, note that we may use the facts that

$$\boxed{\sum_{n=1}^{\infty} \frac{1}{n} \text{ diverges and } \sum_{n=1}^{\infty} \frac{1}{n^2} \text{ converges.}}$$

EXAMPLE 1. The series $\sum_{n=2}^{\infty} \frac{1}{2n^2 - n + 2}$ converges. We test this against the convergent series $\sum_{n=2}^{\infty} \frac{1}{n^2}$. Indeed,

$$\lim_{n\to\infty} \frac{\left(\frac{1}{2n^2 - n + 2}\right)}{\left(\frac{1}{n^2}\right)} = \frac{1}{2},$$

(after some work), proving convergence.

EXAMPLE 2. The series $\sum_{n=0}^{\infty} \frac{1}{\sqrt{n+1}}$ diverges, as

$$\lim_{n\to\infty} \frac{\left(\dfrac{1}{\sqrt{n+1}}\right)}{\left(\dfrac{1}{n}\right)} = \infty,$$

showing that the terms of the series $\displaystyle\sum_{n=0}^{\infty} \frac{1}{\sqrt{n+1}}$ are asymptotically much bigger than the terms of the already divergent series $\displaystyle\sum_{n=1}^{\infty} \frac{1}{n}$. Therefore, by the Limit Comparison Test, $\displaystyle\sum_{n=0}^{\infty} \frac{1}{\sqrt{n+1}}$ diverges.

EXAMPLE 3. The series $\displaystyle\sum_{n=1}^{\infty} \frac{n^2+2n+3}{n^{9/2}}$ converges. We compare it with the convergent series $\displaystyle\sum_{n=1}^{\infty} \frac{1}{n^2}$:

$$\lim_{n\to\infty} \frac{\left(\dfrac{n^2+2n+3}{n^{9/2}}\right)}{\left(\dfrac{1}{n^2}\right)} = \lim_{n\to\infty} \frac{n^2+2n+3}{n^{7/2}} = 0,$$

proving convergence.

EXAMPLE 4. The series $\displaystyle\sum_{n=2}^{\infty} \frac{1}{(\ln n)^2}$ diverges. Watch this:

$$\begin{aligned}
\lim_{n\to\infty} \frac{\frac{1}{(\ln n)^2}}{\left(\frac{1}{n}\right)} &= \lim_{n\to\infty} \frac{n}{(\ln n)^2} \\
&= \lim_{x\to\infty} \frac{x}{(\ln x)^2} \\
&\overset{\text{l'Hôpital}}{=} \lim_{x\to\infty} \frac{\frac{d}{dx}x}{\frac{d}{dx}(\ln x)^2} \\
&= \lim_{x\to\infty} \frac{x}{(2\ln x)} \\
&\overset{\text{l'Hôpital}}{=} \lim_{x\to\infty} \frac{\frac{d}{dx}x}{\frac{d}{dx}(2\ln x)} \\
&= \lim_{x\to\infty} \frac{x}{2} = \infty.
\end{aligned}$$

This says that, asymptotically, the series $\sum_{n=2}^{\infty} \frac{1}{(\ln n)^2}$ is infinitely larger than the divergent harmonic series $\sum_{n=2}^{\infty} \frac{1}{n}$ implying divergence.

The next test will provide us with a rich assortment of series to test with. (So far, we've only been testing against the **convergent** series $\sum_{n=1}^{\infty} \frac{1}{n^2}$ and the **divergent** series $\sum_{n=1}^{\infty} \frac{1}{n}$.)

The p-Series Test. *Let p be a real number. Then*

$$\sum_{n=1}^{\infty} \frac{1}{n^p} \quad \begin{cases} \textbf{converges } \textit{if} & p > 1 \\ \textbf{diverges } \textit{if} & p \leq 1. \end{cases}$$

The p-series test is sometimes called the p-**test** for short; the proof of the above is simple; just as we proved that $\sum_{n=1}^{\infty} \frac{1}{n^2}$ converged by comparing it with $\int_1^{\infty} \frac{dx}{x^2}$ (which converges) and that $\sum_{n=1}^{\infty} \frac{1}{n}$ diverged by comparing with $\int_1^{\infty} \frac{dx}{x^2}$ (which diverges), we see that $\sum_{n=0}^{\infty} \frac{1}{n^p}$ will have

the same behavior as the improper integral $\int_1^\infty \frac{dx}{x^p}$. But, where $p \neq 1$, we have

$$\int_1^\infty \frac{dx}{x^p} = \frac{x^{1-p}}{1-p}\Big|_1^\infty = \begin{cases} \frac{1}{p-1} & \text{if } p > 1 \\ \infty & \text{if } p < 1. \end{cases}$$

We already know that $\sum_{n=1}^{\infty} \frac{1}{n}$ diverges, so we're done!

The ***p*-Test** works very well in conjunction with the **Limit Comparison Test**. The following two examples might help.

EXAMPLE 5. $\sum_{n=1}^{\infty} \frac{n^2+2n+3}{n^{7/2}}$ converges. We compare it with the series $\sum_{n=1}^{\infty} \frac{1}{n^{3/2}}$, which, by the p-test converges:

$$\lim_{n\to\infty} \frac{\left(\frac{n^2+2n+3}{n^{7/2}}\right)}{\left(\frac{1}{n^{3/2}}\right)} = \lim_{n\to\infty} \frac{n^2+2n+3}{n^2} = 1,$$

proving convergence.

EXAMPLE 6. $\sum_{n=1}^{\infty} \frac{n^2+2n+3}{n^{7/3}}$ diverges. We compare it with the series $\sum_{n=1}^{\infty} \frac{1}{\sqrt[3]{n}}$, which, by the p-test diverges:

$$\lim_{n\to\infty} \frac{\left(\frac{n^2+2n+3}{n^{7/3}}\right)}{\left(\frac{1}{\sqrt[3]{n}}\right)} = \lim_{n\to\infty} \frac{n^2+2n+3}{n^2} = 1,$$

proving divergence.

There is one more very useful test, one which works particularly well with expressions containing exponentials and/or factorials. This method is based on the fact that if $|r| < 1$, then the infinite geometric

series $a + ar + ar^2 + \cdots$ converges (to $\dfrac{a}{1-r}$). In this test we do not need to assume that the series consists only of non-negative terms.

The Ratio Test *Let* $\sum_{n=0}^{\infty} a_n$ *be an infinite series. Assume that*

$$\lim_{n\to\infty} \frac{a_{n+1}}{a_n} = R.$$

Then

(i) if $|R| < 1$, *then* $\sum_{n=0}^{\infty} a_n$ *converges;*

(ii) if $|R| > 1$, *then* $\sum_{n=0}^{\infty} a_n$ *diverges;*

(iii) if $|R| = 1$, *then this test is inconclusive.*

The reasoning behind the above is simple. First of all, in case (i) we see that $\sum_{n=0}^{\infty} a_n$ is asymptotically a geometric series with ratio $|R| < 1$ and hence converges (but we still probably won't know what the series converges to). In case (ii) then $\sum_{n=0}^{\infty} a_n$ will diverge since asymptotically each term is R times the previous one, which certainly implies that $\lim_{n\to\infty} a_n \neq 0$, preventing convergence. Note that in the two cases $\sum_{n=1}^{\infty} \dfrac{1}{n}$ and $\sum_{n=1}^{\infty} \dfrac{1}{n^2}$ we have $\lim_{n\to\infty} \dfrac{a_{n+1}}{a_n} = 1$,[10] which is why this case is inclusive.

We turn again to some examples.

EXAMPLE 7. Consider the series $\sum_{n=1}^{\infty} \dfrac{(n+1)^3}{n!}$. We have

[10] Indeed, we have in the first case

$$\lim_{n\to\infty} \frac{a_{n+1}}{a_n} = \lim_{n\to\infty} \frac{\left(\frac{1}{n+1}\right)}{\left(\frac{1}{n}\right)} = \lim_{n\to\infty} \frac{n}{n+1} = 1,$$

in the first case, and that

$$\lim_{n\to\infty} \frac{a_{n+1}}{a_n} = \lim_{n\to\infty} \frac{\left(\frac{1}{(n+1)^2}\right)}{\left(\frac{1}{n^2}\right)} = \lim_{n\to\infty} \frac{n}{n+1} = 1,$$

in the second case, despite the fact that the first series diverges and the second series converges.

$$\begin{aligned}\lim_{n\to\infty}\frac{a_{n+1}}{a_n} &= \lim_{n\to\infty}\frac{\left(\frac{(n+2)^3}{(n+1)!}\right)}{\left(\frac{(n+1)^3}{n!}\right)} \\ &= \lim_{n\to\infty}\frac{(n+2)^3}{(n+1)(n+1)^3} = 0.\end{aligned}$$

Therefore, $\sum_{n=1}^{\infty}\frac{(n+1)^3}{n!}$ converges.

EXAMPLE 8. Consider the series $\sum_{n=1}^{\infty}\frac{n^2}{3^n}$. We have

$$\begin{aligned}\lim_{n\to\infty}\frac{a_{n+1}}{a_n} &= \lim_{n\to\infty}\frac{\left(\frac{(n+1)^2}{3^{n+1}}\right)}{\left(\frac{n^2}{3^n}\right)} \\ &= \lim_{n\to\infty}\frac{(n+1)^2}{3n^2} = \frac{1}{3} < 1.\end{aligned}$$

It follows, therefore, that $\sum_{n=1}^{\infty}\frac{n^2}{3^n}$ also converges.

EXAMPLE 9. Consider the series $\sum_{n=1}^{\infty}\frac{n!}{2^n}$. We have

$$\begin{aligned}\lim_{n\to\infty}\frac{a_{n+1}}{a_n} &= \lim_{n\to\infty}\frac{\left(\frac{(n+1)!}{2^{n+1}}\right)}{\left(\frac{n!}{2^n}\right)} \\ &= \lim_{n\to\infty}\frac{n+1}{2} = \infty,\end{aligned}$$

which certainly proves divergence.

EXERCISES

1. Test each of the series below for convergence.

(a) $\sum_{n=1}^{\infty}\frac{n+2}{n^2+10n}$ (b) $\sum_{n=0}^{\infty}\frac{n^2-n+2}{n^4+n^2+1}$

(c) $\displaystyle\sum_{n=1}^{\infty} \frac{n^2}{\sqrt{n^7+2n}}$

(d) $\displaystyle\sum_{n=0}^{\infty} \frac{n^2+2n}{3^n}$

(e) $\displaystyle\sum_{n=0}^{\infty} \frac{(n+1)3^n}{n!}$

(f) $\displaystyle\sum_{n=1}^{\infty} \frac{2^n}{n^n}$

(g) $\displaystyle\sum_{n=1}^{\infty} \frac{n!}{n^n}$

(h) $\displaystyle\sum_{n=1}^{\infty} \frac{1}{\left(1+\frac{1}{n}\right)^n}$

2. As we have already seen, the series $\displaystyle\sum_{n=1}^{\infty} \frac{1}{n^2}$, converges. In fact, it is known that $\displaystyle\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$; a sketch of Euler's proof is given on page 228 of the Haese-Harris textbook.[11] Using this fact, argue that

$$\sum_{n=0}^{\infty} \frac{1}{(2n+1)^2} = \frac{\pi^2}{8}.$$

3. Prove the **sinusoidal** p-series test, namely that

$$\sum_{n=1}^{\infty} \sin\left(\frac{2\pi}{n^p}\right) \begin{cases} \text{converges} & \text{if } p > 1, \\ \text{diverges} & \text{if } p \leq 1. \end{cases}$$

(Of course, the 2π can be replaced by any constant. Exercise 2 on page 256 is, of course, relevant here!)

4. Recall Euler's ϕ-function ϕ (see page 63). Determine the behavior of the series $\displaystyle\sum_{n=1}^{\infty} \frac{1}{n\phi(n)}$. (See Exercise 16d on page 64.)

5. How about $\displaystyle\sum_{n=1}^{\infty} \frac{\phi(n)}{n^2}$?

6. Let F_0, F_1, F_2, ... be the terms of the Fibonacci sequence (see page 93). Show that $\displaystyle\sum_{n=0}^{\infty} \frac{F_n}{2^n}$ converges and compute this sum explicitly. (Hint: you'll probably need to work through Exercise 7 on page 106 first.)

[11] Peter Blythe, Peter Joseph, Paul Urban, David Martin, Robert Haese, and Michael Haese, MATHEMATICS FOR THE INTERNATIONAL STUDENT; MATHEMATICS HL (OPTIONS), Haese and Harris Publications, 2005, Adelaide, ISBN 1 876543 33 7

7. As in the above sequence, Let F_0, F_1, F_2, ... be the terms of the Fibonacci sequence. Show that $\sum_{k=0}^{\infty} \frac{1}{F_k} < \alpha$, where $\alpha = \frac{1+\sqrt{5}}{2}$ (the **golden ratio**). (Hint: show that if $k \geq 2$, then $F_k > \alpha^{k-1}$. and then use the ratio test.)[12]

8. Consider the generalized Fibonacci sequence (see Exercise 9 on page 106) defined by $u_0 = u_1 = 1$ and $u_{n+2} = u_{n+1} + u_n$. Show that if a, b are such that $u_n \to 0$ as $n \to \infty$, then $\sum_{n=0}^{\infty} u_n$ converges and compute this sum in terms of a and b.

5.2.3 Conditional and absolute convergence; alternating series

In this subsection we shall consider series of the form $\sum_{n=0}^{\infty} a_n$ where the individual terms a_n are not necessarily non-negative. We shall first make the following useful definition. An infinite series $\sum_{n=0}^{\infty} a_n$ is called **absolutely convergent** if the series $\sum_{n=0}^{\infty} |a_n|$ converges. This is important because of the following result.

Theorem. *If the series $\sum_{n=0}^{\infty} a_n$ is absolutely convergent, then it is convergent.*

Proof. Note that we clearly have

$$0 \leq a_n + |a_n| \leq 2|a_n|, \ n = 0, 1, 2, \ldots.$$

Since $\sum_{n=0}^{\infty} 2|a_n|$ converges, so does $\sum_{n=0}^{\infty} (a_n + |a_n|)$; call the limit L. Therefore, $\sum_{n=0}^{\infty} a_n = L - \sum_{n=0}^{\infty} |a_n|$, proving that $\sum_{n=0}^{\infty} a_n$ converges, as well.

[12]The above says, of course, that the infinite series of the reciprocals of the Fibonacci numbers converges. Its value is known to be an irrational number $\approx 3.35988566...$

We consider a couple of simple illustrations of the above theorem.

EXAMPLE 1. The series $\sum_{n=1}^{\infty} \frac{(-1)^n}{n^2}$ will converge by the above theorem, together with the ***p*-Test**.

EXAMPLE 2. The series $\sum_{n=0}^{\infty} \frac{(-1)^n}{n}$ does not converge absolutely; however, we'll see below that this series does converge.

An infinite series $\sum_{n=0}^{\infty} a_n$ which converges but is not absolutely convergent is called **conditionally convergent**. There are plenty of conditionally convergent series, as guaranteed by the following theorem.

Theorem. (**Alternating Series Test**) *Let* $a_0 \geq a_1 \geq a_2 \geq \cdots \geq 0$ *and satisfy* $\lim_{n\to\infty} a_n = 0$. *Then the "alternating series"* $\sum_{n=0}^{\infty} (-1)^n a_n$ *converges.*[13]

We'll conclude this section with two illustrations of the **Alternating Series Test**.

EXAMPLE 3. We know that the harmonic series $\sum_{n=1}^{\infty} \frac{1}{n}$ diverges; however, since the terms of this series decrease and tend to zero, the **Alternating Series Test** guarantees that $\sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n}$ converges. We'll show later on that this actually converges to $\ln 2$ (see page 302).

EXAMPLE 4. The series $\sum_{n=0}^{\infty} \frac{n}{n^2+1}$ can be shown to diverge by applying the **Limit Comparison Test** with a comparison with the harmonic

[13]The proof of this is pretty simple. First of all, note that the "even" partial sums satisfy

$$(a_0 - a_1) \leq (a_0 - a_1) + (a_2 - a_3) \leq (a_0 - a_1) + (a_2 - a_3) + (a_4 - a_5) \leq \cdots,$$

so it suffices to show that these are all bounded by some number (see Figure 1, page 266). However, note that

$$a_0 - \underbrace{((a_1 - a_2) + (a_3 - a_4) + (a_5 - a_6) + \cdots + (a_{2n-3} - a_{2n-2}))}_{\text{This is positive!}} - a_{2n-1} \leq a_0,$$

so we're done.

series (do this!). However, the terms decrease and tend to zero and so by the **Alternating Series Test** $\sum_{n=0}^{\infty} \frac{(-1)^n n}{n^2+1}$ converges.

EXERCISES

1. Test each of the series below for convergence.

   (a) $\sum_{n=1}^{\infty} (-1)^n \frac{n+2}{n^2+10n}$

   (b) $\sum_{n=2}^{\infty} \frac{(-1)^n}{\ln n}$

   (c) $\sum_{n=1}^{\infty} (-1)^n \frac{\ln n}{\ln(n^3+1)}$

   (d) $\sum_{n=1}^{\infty} \left(\frac{1}{n} - \frac{1}{n^2}\right)$

   (e) $\sum_{n=2}^{\infty} (-1)^n \frac{\ln \ln n}{\ln n}$

   (f) $\sum_{n=1}^{\infty} \frac{(-1)^n}{\left(1+\frac{1}{n}\right)^n}$

   (g) $\sum_{n=1}^{\infty} \frac{(-1)^n}{\sqrt{n}+\sqrt{n+1}}$

   (h) $\sum_{n=1}^{\infty} \frac{(-2)^n}{n!}$

2. Determine whether each of the series above converges **conditionally**, converges **absolutely** or diverges.

3. Prove that the series the improper integral $\int_{-\infty}^{\infty} \frac{\sin x}{x}\, dx$ converges.[14]

4. Prove that the improper integral $\int_0^{\infty} \cos x^2\, dx$ converges.[15] (Hint: try the substitution $u = x^2$ and see if you can apply the Alternating Series Test.)

5. Consider the infinite series $\sum_{n=0}^{\infty} \frac{\epsilon_n}{2^n}$, where each ϵ_n is ± 1. Show that any real number x, $-2 \leq x \leq 2$ can be represented by such a series by considering the steps below:

   (a) Write $\Sigma = \sum_{n=0}^{\infty} \frac{\epsilon_n}{2^n} = \Sigma_+ - \Sigma_-$, where Σ_+ is the sum of the positive terms in Σ and where Σ_- is $-$(negative terms in Σ).

[14] In fact, it converges to π.

[15] This can be shown to converge to $\frac{1}{2}\sqrt{\frac{\pi}{2}}$.

(See the footnote.[16]) By thinking in terms of binary decimal representations (See, e.g., Exercise 4 on page 92) argue that any real number y with $|y| \leq 2$ can be represented in the form Σ_+.

(b) If $\Sigma_+ = y$, show that $\Sigma_+ - \Sigma_- = 2(y-1)$.

(c) Conclude that any number x between -2 and 2 can be represented as $\Sigma_+ - \Sigma_-$ and hence as the infinite series $\displaystyle\sum_{n=0}^{\infty} \frac{\epsilon_n}{2^n}$. (This result is **not** true if the series is replaced by, say, one of the form $\displaystyle\sum_{n=0}^{\infty} \frac{\epsilon_n}{3^n}$. While the values of such a series would always lie between $-\frac{3}{2}$ and $\frac{3}{2}$, and despite the fact that uncountably many such numbers would occur, the set of such numbers is still very small.[17]

5.2.4 The Dirichlet test for convergence (optional discussion)

There is a very convenient test which can be thought of as a generalization of the alternating series test and often applies very nicely to testing for conditional convergence. For example, we may wish to test the series $\displaystyle\sum_{n=1}^{\infty} \frac{\cos n}{n}$ for convergence. It is not clear whether this series is absolutely convergent[18], nor is it an alternating series, so none of the methods presented thus far apply.

Let us first consider the following very useful lemma.

LEMMA. *Let (a_n) and (b_n) be two sequences and set $s_n = \displaystyle\sum_{k=1}^{n} a_k$. Then one has*

[16] There is an important result being used here, namely that if $\sum u_n$ is an absolutely convergent series, then its sum is unaffected by any rearrangement of its terms.

[17] The smallness of this set is best expressed by saying that it has "measure zero." Alternatively, if we were to select a real number randomly from the interval $\left[-\frac{3}{2}, \frac{3}{2}\right]$, then the probability of selecting a number of the form $\displaystyle\sum_{n=0}^{\infty} \frac{\epsilon_n}{3^n}$ is zero. Finally, if instead of allowing the numerators ϵ_n to be ± 1 we insisted that they be either 0 or 2, then what results is the so-called **Cantor Ternary Set** (which also has measure zero).

[18] It's not, but this takes some work to show.

$$\sum_{k=1}^{n} a_k b_k \;=\; s_n b_{n+1} + \sum_{k=1}^{n} s_k (b_k - b_{k+1}).$$

PROOF. Setting $s_0 = 0$ we obviously have $a_k = s_n - s_{k-1},\ k \geq 1$. From this, one has

$$\begin{aligned} \sum_{k=1}^{n} a_k b_k &= \sum_{k=1}^{n} (s_k - s_{k-1}) b_k \\ &= \sum_{k=1}^{n} s_k b_k - \sum_{k=1}^{n} s_{k-1} b_k \\ &= s_n b_{n+1} + \sum_{k=1}^{n} s_k (b_k - b_{k+1}). \end{aligned}$$

DIRICHLET'S THEOREM FOR CONVERGENCE. *Let (a_n) and (b_n) be two sequences of real numbers. If the partial sums $\left| \sum_{k=1}^{n} a_n \right|$ are all bounded by some constant M, and if*

$$b_1 > b_2 > b_3 > \cdots \geq 0 \quad \textit{with} \quad \lim_{n \to \infty} b_n = 0,$$

then the series $\sum_{k=1}^{\infty} a_k b_k$ converges.

PROOF. Setting $s_n = \sum_{k=1}^{n} a_n$ and $r_n = \sum_{k=1}^{n} a_k b_k$ we have from the above lemma that

$$r_n - r_m = s_n b_{n+1} - s_m b_{m+1} + \sum_{k=m+1}^{n} s_k (b_k - b_{k+1}),$$

where $n \geq m$ are positive indices. Taking absolute values and applying the **Triangle inequality** provides

$$\begin{aligned} |r_n - r_m| &\leq |s_n| b_{n+1} + |s_m| b_{m+1} + \sum_{k=m+1}^{n} |s_k| (b_k - b_{k+1}) \\ &\leq M b_{n+1} + M b_{m+1} + M \sum_{k=m+1}^{n} (b_k - b_{k+1}) \\ &= 2M b_{m+1}. \end{aligned}$$

Since $b_k \to 0$ as $k \to \infty$, we conclude that (r_n) is a Cauchy sequence of real numbers, hence converges.

As a corollary to the above, let's consider the convergence of the sequence alluded to above, viz., $\sum_{n=1}^{\infty} \frac{\cos n}{n}$. To do this, we start with the fact that for all integers k,

$$2\sin(1/2)\cos k = \sin(k+1/2) - \sin(k-1/2).$$

Therefore,

$$\begin{aligned} \left|2\sin(1/2)\right| \cdot \left|\sum_{k=1}^{n} \cos k\right| &= \left|\sum_{k=1}^{n} (\sin(k+1/2) - \sin(k-1/2))\right| \\ &= \left|\sin(n+1/2) - \sin(1/2)\right| \\ &\leq 2. \end{aligned}$$

Since $\sin(1/2) \neq 0$ we already see that $\sum_{k=1}^{n} \cos k$ is bounded, and Dirichlet's theorem applies, proving the convergence of $\sum_{n=1}^{\infty} \frac{\cos n}{n}$.

EXERCISES

1. Strenghten the result proved above by proving that the series $\sum_{n=1}^{\infty} \frac{\cos nx}{n^p}$ converges whenever $p > 0$, and x is not an integral multiple of 2π.
2. Prove that $\sum_{n=1}^{\infty} \frac{\sin nx}{n^p}$ converges whenever $p > 0$.

5.3 The Concept of a Power Series

Let us return to the familiar geometric series, with ratio r satisfying $|r| < 1$.

$$a + ar + ar^2 + \cdots \;=\; \frac{a}{1-r}.$$

Let's make a minor cosmetic change: rather than writing r in the above sum, we shall write x:

$$a + ax + ax^2 + \cdots \;=\; \frac{a}{1-x}, \quad |x| < 1.$$

In other words, if we set

$$f(x) \;=\; a + ax + ax^2 + \cdots \;=\; \sum_{n=0}^{\infty} ax^n \quad \text{and set} \quad g(x) \;=\; \frac{a}{1-x},$$

then the following facts emerge:

(a) The domain of f is $-1 < x < 1$, and the domain of g is $x \neq 1$.

(b) $f(x) = g(x)$ for all x in the interval $-1 < x < 1$.

We say, therefore, that $\sum_{n=0}^{\infty} ax^n$ is the **power series representation** of $g(x)$, **valid on the interval** $-1 < x < 1$.

So what is a power series anyway? Well, it's just an expression of the form

$$\sum_{n=0}^{\infty} a_n x^n,$$

where a_0, a_1, a_2, ... are just real constants. For any particular value of x this infinite sum **may or may not converge**; we'll have plenty to say about issues of convergence.

5.3.1 Radius and interval of convergence

Our primary tool in determining the convergence properies of a power series $\sum_{n=0}^{\infty} a_n x^n$ will be the **Ratio Test**. Recall that the series $\sum_{n=0}^{\infty} |a_n x^n|$ will converge if

$$\begin{aligned} 1 &> \lim_{n\to\infty} \frac{|a_{n+1}x^{n+1}|}{|a_n x^n|} \\ &= |x| \lim_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right|, \end{aligned}$$

which means that

$$\boxed{\sum_{n=0}^{\infty} a_n x^n \text{ is absolutely convergent for all x satisfying } |x| < \lim_{n\to\infty} \left|\frac{a_n}{a_{n+1}}\right|}.$$

The quantity $R = \lim_{n\to\infty} \left|\frac{a_n}{a_{n+1}}\right|$ is sometimes called the **radius of convergence** of the power series $\sum_{n=0}^{\infty} a_n x^n$. Again, as long as $-R < x < R$, we are guaranteed that $\sum_{n=0}^{\infty} a_n x^n$ is absolutely convergent and hence convergent.

A few simple examples should be instructive.

EXAMPLE 1. The power series $\sum_{n=0}^{\infty} \frac{(-1)^n x^n}{2n+1}$ has radius of convergence

$$R = \lim_{n\to\infty} \left|\frac{a_n}{a_{n+1}}\right| = \lim_{n\to\infty} \frac{\left(\frac{n}{2n+1}\right)}{\left(\frac{n+1}{2n+3}\right)} = \lim_{n\to\infty} \frac{n(2n+3)}{(n+1)(2n+1)} = 1.$$

This means that the above power series has radius of convergence 1 and so the series is absolutely convergent for $-1 < x < 1$.

EXAMPLE 2. The power series $\sum_{n=0}^{\infty} \frac{nx^n}{2^n}$ has radius of convergence

$$R \;=\; \lim_{n\to\infty}\left|\frac{a_n}{a_{n+1}}\right| \;=\; \lim_{n\to\infty}\frac{\left(\frac{n}{2^n}\right)}{\left(\frac{n+1}{2^{n+1}}\right)} \;=\; \lim_{n\to\infty}\frac{2n}{n+1} \;=\; 2,$$

so in this case the radius of convergence is 2, which guarantees that the power series converges for all x satisfying $-2 < x < 2$.

EXAMPLE 3. Consider the power series $\sum_{n=0}^{\infty}\frac{(-1)^n x^n}{n!}$. In this case the radius of convergence is similarly computed:

$$R \;=\; \lim_{n\to\infty}\left|\frac{a_n}{a_{n+1}}\right| \;=\; \lim_{n\to\infty}\frac{\left(\frac{1}{n!}\right)}{\left(\frac{1}{(n+1)!}\right)} \;=\; \lim_{n\to\infty} n+1 \;=\; \infty.$$

This infinite radius of convergence means that the power series $\sum_{n=0}^{\infty}\frac{(-1)^n x^n}{n!}$ actually converges for all real numbers x.

EXAMPLE 4. We consider here the series $\sum_{n=0}^{\infty}\frac{(x+2)^n}{n2^n}$, which has radius of convergence

$$R \;=\; \lim_{n\to\infty}\frac{(n+1)2^{n+1}}{n2^n} \;=\; 2.$$

This means that the series will converge where $|x+2| < 2$, i.e., where $-4 < x < 0$.

EXAMPLE 5. Here we consider the power series $\sum_{n=0}^{\infty}\frac{nx^{2n}}{2^n}$. The radius of convergence is

$$R \;=\; \lim_{n\to\infty}\frac{n}{2^n}\cdot\frac{2^{n+1}}{2(n+1)} = 2.$$

But this is a power series in x^2 and so will converge if $x^2 < 2$. This gives convergence on the interval $-\sqrt{2} < x < \sqrt{2}$.

In the examples above we computed intervals within which we are **guaranteed** convergence of the power series. Next, note that for values of x **outside** the radius of convergence we **cannot** have convergence, for then the limit of the ratios will be greater than 1, preventing the individual terms from approaching 0. This raises the question of **convergence at the endpoints**. We shall investigate this in the examples already considered above.

EXAMPLE 6. We have seen that the power series $\sum_{n=0}^{\infty} \frac{(-1)^n x^n}{2n+1}$ has radius of convergence $R = 1$ meaning that this series will converge in the interval $-1 < x < 1$. What if $x = -1$? What if $x = 1$? Well, we can take these up separately, using the methods of the previous two subsections. If $x = -1$, we have the series

$$\sum_{n=0}^{\infty} \frac{(-1)^n(-1)^n}{2n+1} = \sum_{n=0}^{\infty} \frac{1}{2n+1},$$

which **diverges** (use the **Limit Comparison Test** against the harmonic series). If $x = 1$, then the series becomes

$$\sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1},$$

which converges by the **Alternating Series Test**. We therefore know the full story and can state the **interval of convergence**:

$$\sum_{n=0}^{\infty} \frac{(-1)^n x^n}{2n+1} \quad \text{has interval of convergence} \quad -1 < x \leq 1.$$

EXAMPLE 7. We saw that the power series $\sum_{n=0}^{\infty} \frac{nx^n}{2^n}$ has radius of convergence $r = 2$. What about the behavior of the series when $x = \pm 2$. If $x = -2$ we have the series

$$\sum_{n=0}^{\infty} \frac{n(-2)^n}{2^n} = \sum_{n=0}^{\infty} (-1)^n n \quad \text{which diverges,}$$

whereas when $x = 2$, we have

$$\sum_{n=0}^{\infty} \frac{n2^n}{2^n} = \sum_{n=0}^{\infty} n \quad \text{which also diverges.}$$

Therefore,

$$\sum_{n=0}^{\infty} \frac{nx^n}{2^n} \quad \text{has interval of convergence} \quad -2 < x < 2.$$

EXAMPLE 8. The power series $\sum_{n=0}^{\infty} \frac{(-1)^n x^n}{n!}$ has infinite radius of convergence so there are no endpoints to check.

Before closing this section, we should mention that not all power series are of the form $\sum_{n=0}^{\infty} a_n x^n$; they may appear in a "translated format," say, one like $\sum_{n=0}^{\infty} a_n (x-a)^n$, where a is a constant. For example, consider the series in Example 4, on page 285. What would the interval of convergence look here? We already saw that this series was guaranteed to converge on the interval $-4 < x < 0$. If $x = -4$, then this series is the convergent alternating harmonic series. If $x = 0$, then the series becomes the divergent harmonic series. Summarizing, the interval of convergence is $-4 \leq x < 0$.

EXERCISES

1. Determine the radius of convergence of each of the power series below:

   (a) $\sum_{n=0}^{\infty} \frac{nx^n}{n+1}$

   (b) $\sum_{n=0}^{\infty} \frac{nx^n}{2^n}$

   (c) $\sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{n^2+1}$

   (d) $\sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{3^n}$

   (e) $\sum_{n=0}^{\infty} \frac{(x+2)^n}{2^n}$

   (f) $\sum_{n=0}^{\infty} \frac{(-1)^n (x+2)^n}{2^n}$

   (g) $\sum_{n=0}^{\infty} \frac{(2x)^n}{n!}$

   (h) $\sum_{n=1}^{\infty} \frac{x^n}{\left(1+\frac{1}{n}\right)^n}$

   (i) $\sum_{n=1}^{\infty} \frac{n \ln nx^n}{2^n}$

2. Determine the interval convergence of each of the power series below:

(a) $\sum_{n=0}^{\infty} \frac{nx^n}{n+1}$

(b) $\sum_{n=0}^{\infty} \frac{nx^n}{2^n}$

(c) $\sum_{n=1}^{\infty} \frac{x^n}{n2^n}$

(d) $\sum_{n=0}^{\infty} \frac{(-1)^n(x-2)^{2n}}{3^n}$

(e) $\sum_{n=1}^{\infty} \frac{(-1)^n(x+2)^n}{n2^n}$

(f) $\sum_{n=0}^{\infty} \frac{(-1)^n(x+2)^n}{2^n}$

(g) $\sum_{n=0}^{\infty} \frac{(2x)^n}{n!}$

(h) $\sum_{n=2}^{\infty} \frac{(-1)^n x^n}{n \ln n 2^n}$

(i) $\sum_{n=1}^{\infty} \frac{(-1)^n n \ln n x^n}{2^n}$

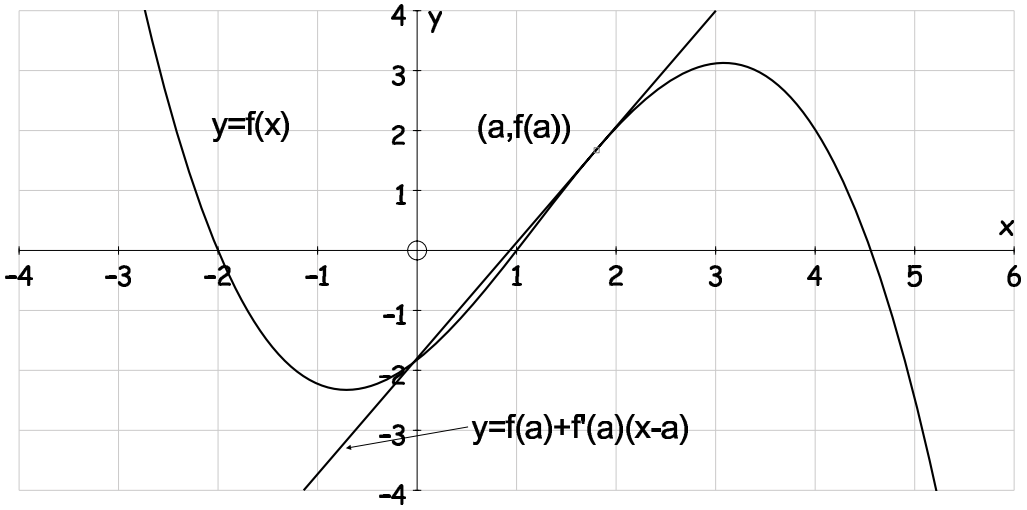
(j) $\sum_{n=0}^{\infty} \frac{(3x-2)^n}{2^n}$

5.4 Polynomial Approximations; Maclaurin and Taylor Expansions

Way back in our study of the **linearization of a function** we saw that it was occassionally convenient and useful to approximate a function by one of its tangent lines. More precisely, if f is a differentiable function, and if a is a value in its domain, then we have the approximation $f'(a) \approx \frac{f(x)-f(a)}{x-a}$, which results in

$$f(x) \approx f(a) + f'(a)(x-a) \quad \text{for } x \text{ near } a.$$

A graph of this situation should help remind the student of how good (or bad) such an approximation might be:



Notice that as long as x does not move too far away from the point a, then the above approximation is pretty good.

Let's take a second look at the above. Note that in approximating a function f by a **linear function** L near the point a, then

(i) The graph of L will pass through the point $(a, f(a))$, i.e., $L(a) = f(a)$, and

(ii) The slope of the line $y = L(x)$ will be the same as the derivative of f at $x = a$, i.e., $L'(a) = f'(a)$.

That is the say, the "best" linear function L to use in approximating f near a is one whose 0-th and first derivatives at $x = a$ are the same as for f:

$$L(a) = f(a) \quad \text{and} \quad L'(a) = f'(a).$$

So what if instead of using a straight line to approximate f we were to use a quadratic function Q? What, then, would be the natural requirements? Well, in analogy with what was said above we would require f and Q to have the same first three derivatives (0-th, first, and second) at $x = a$:

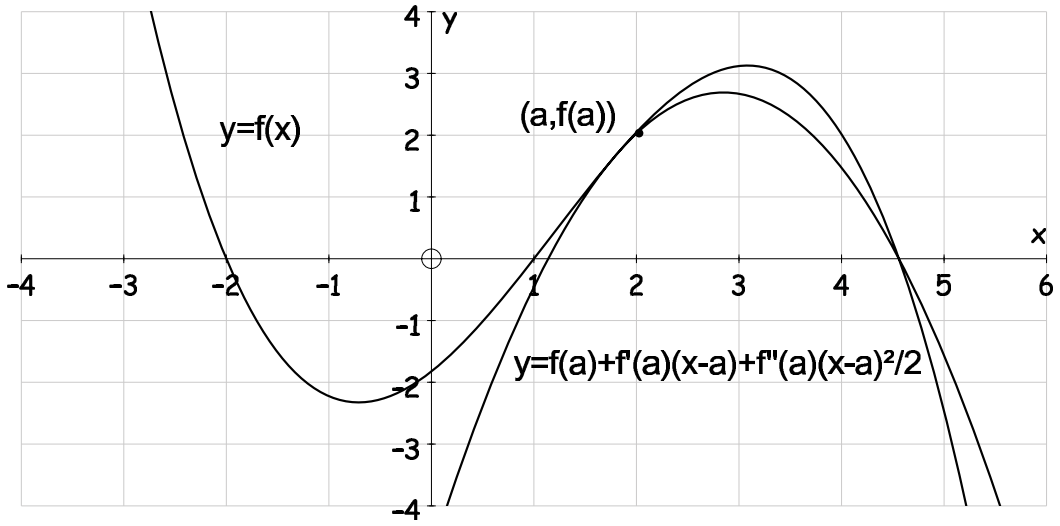
$$Q(a) = f(a), \quad Q'(a) = f'(a), \quad \text{and} \quad Q''(a) = f''(a).$$

Such a quadratic function is actually very easy to build: the result would be that

$$Q(x) \;=\; f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2.$$

(The reader should pause here to verify that the above quadratic function really does have the same first three derivatives as f at $x = a$.)

This "second-order" approximation is depicted here. Notice the improvement over the linear approximation.



In general, we may approximate a function with a polynomial $P_n(x)$ of degree n by insisting that this polynomial have all of its first $n+1$ derivatives at $x = a$ equal those of f:

$$P_n(a) = f(a), \;\; P_n'(a) = f'(a), \;\; P_n''(a) = f''(a), \;\; \cdots, \; P_n^{(n)}(a) = f^{(n)}(a),$$

where, in general, $f^{(k)}(x)$ denotes the k-th derivative of f at x. It is easy to see that the following gives a recipe for $P_n(x)$:

$$P_n(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a) + \frac{f'''(a)}{3!}(x-a)^3 + \cdots + f^{(n)}(a)(x-a)^n$$

$$= \sum_{k=0}^{n} \frac{f^{(k)}(a)}{n!}(x-a)^k.$$

We expect, then, to have a pretty good approximation

$$f(x) \approx \sum_{k=0}^{n} \frac{f^{(k)}(a)}{n!}(x-a)^k.$$

The polynomial $P_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{n!}(x-a)^k$ is called the **Taylor polynomial of degree n** for f at $x = a$. If $a = 0$, the above polynomial becomes

$$f(x) \approx \sum_{k=0}^{n} \frac{f^{(k)}(a)}{n!}x^k$$

and is usually called the **Maclaurin polynomial of degree n** for f.

What if, instead of stopping at a degree n polynomial, we continued the process indefinitely to obtain a power series? This is possible and we obtain

$\sum_{k=0}^{\infty} \frac{f^{(k)}(a)}{n!}(x-a)^k$ **Taylor series** for f at $x = a$, and

$\sum_{k=0}^{\infty} \frac{f^{(k)}(a)}{n!}x^k$ **Maclaurin series** for f.

Warning. It is very tempting to assume that the Taylor series for a function f will actually converge to $f(x)$ on its interval of convergence, that is,

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(a)}{n!}(x-a)^k.$$

For most of the functions we've considered here, this is true, but the general result can fail.[19] As a result, we shall adopt the notation

$$f(x) \sim \sum_{k=0}^{\infty} \frac{f^{(k)}(a)}{n!}(x-a)^k$$

to mean that $\boldsymbol{f(x)}$ **is represented by the power series** $\sum_{k=0}^{\infty} \frac{f^{(k)}(a)}{n!}(x-a)^k$; in Subsection 3.2 we'll worry about whether "$\sim$" can be replaced with "$=$". First, however, we shall delve into some computations.

5.4.1 Computations and tricks

In this subsection we'll give some computations of some Taylor and Maclaurin series, and provide some interesting shortcuts along the way.

EXAMPLE 1. Let $f(x) = \sin x$ and find its Maclaurin series expansion. This is simple as the derivatives (at $x = 0$) are easy to compute

$$f^{(0)}(x) = \sin 0 = 0,\ f'(0) = \cos 0 = 1,\ f''(0) = -\sin 0 = 0,$$

$$f'''(0) = -\cos 0 = -1, f^{(4)}(0) = \sin 0 = 0,$$

and the pattern repeats all over again. This immediately gives the Maclaurin series for $\sin x$:

$$\sin x \sim x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!}\,.$$

[19] As an example, consider the function f defined by setting

$$f(x) = \begin{cases} e^{-1/x^2} & \text{if } x \neq 0 \\ 0 & \text{if } x = 0. \end{cases}$$

One can show that all derivatives of f vanish at $x = 0$ and so cannot equal its Maclaurin series in any interval about $x = 0$.

EXAMPLE 2. **(A handy trick)** If we wish to compute the Maclaurin series for $\cos x$, we could certainly follow the same procedure as for the $\sin x$ in the above example. However, since $\cos x = \frac{d}{dx}\sin x$, we can likewise obtain the Maclaurin series for the $\cos x$ by differentiating that for the $\sin x$, this yields the series

$$\cos x \sim \frac{d}{dx}\sum_{n=0}^{\infty}(-1)^n\frac{x^{2n+1}}{(2n+1)!} = \sum_{n=0}^{\infty}(-1)^n\frac{x^{2n}}{(2n)!}.$$

EXAMPLE 3. We we wanted to compute the Maclaurin series for $\sin x^2$, then computing the higher-order derivatives of $\sin x^2$ would be *extremely tedious*! A more sensible alternative would be to simply replace x by x^2 in the Maclaurin series for $\sin x$:

$$\sin x^2 \sim x^2 - \frac{x^6}{3!} + \frac{x^{10}}{5!} - \frac{x^{14}}{7!} + \cdots = \sum_{n=0}^{\infty}(-1)^n\frac{x^{4n+2}}{(2n+1)!}\,.$$

EXAMPLE 4. **(A handy trick)** Since $\ln(1+x) = \int\frac{dx}{1+x}$ we may start with the geometric series

$$1 - x + x^2 - \cdots = \sum_{n=0}^{\infty}(-1)^n x^n = \frac{1}{1+x},$$

and then integrate each term to obtain the Maclaurin series for $\ln(1+x)$:

$$\ln(1+x) \sim x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots = \sum_{n=0}^{\infty}(-1)^n\int x^n\,dx = \sum_{n=0}^{\infty}\frac{x^{n+1}}{n+1}.$$

(Note that there is no constant occuring in the above integrations since when $x = 0$, $\ln(1+x) = \ln 1 = 0$.)

EXAMPLE 5. Since $\frac{d^n}{dx^n}e^x = e^x = 1$ at $x = 0$, we immediately have the Maclaurin series expansion for e^x:

$$e^x \sim\ 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots\ =\ \sum_{n=0}^{\infty}\frac{x^n}{n!}\ .$$

EXAMPLE 6. **(A handy trick)** In the above series we may substitute $-x^2$ for x and get the Maclaurin series for e^{-x^2}:

$$e^{-x^2} \sim 1 - x^2 + \frac{x^4}{2!} - \frac{x^6}{3!} + \cdots = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{n!} .$$

Note how easy this is compared to having to calculate $\frac{d^n}{dx^n}(e^{-x^2})$ (where the chain and product rules very much complicate things!).

EXAMPLE 7. **(A handy trick)** Let's find the Maclaurin series expansion of the function $f(x) = \dfrac{\sin x}{x}$. In this case, we certainly wouldn't want to compute successive derivatives of the quotient $\dfrac{\sin x}{x}$. However, remembering the Maclaurin series expansion of $\sin x$ and then dividing by x will accomplish the same thing much more easily; the resulting series is

$$\frac{\sin x}{x} \sim 1 - \frac{x^2}{3!} + \frac{x^4}{5!} - \cdots = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n+1)!}.$$

EXAMPLE 8. The Taylor series expansion of $\cos x$ about $x = \dfrac{\pi}{2}$. We have, where $f(x) = \cos x$, that $f\left(\frac{\pi}{2}\right) = \cos\left(\frac{\pi}{2}\right) = 0$, $f'\left(\frac{\pi}{2}\right) = -\sin\left(\frac{\pi}{2}\right) = -1$, $f''\left(\frac{\pi}{2}\right) = -\cos\left(\frac{\pi}{2}\right) = 0$, $f'''\left(\frac{\pi}{2}\right) = \sin\left(\frac{\pi}{2}\right) = 1$, after which point the cycle repeats. Therefore, the Taylor series is

$$\cos x \sim -\left(x - \tfrac{\pi}{2}\right) + \frac{\left(x - \frac{\pi}{2}\right)^3}{3!} - \frac{\left(x - \frac{\pi}{2}\right)^5}{5!} + \cdots = \sum_{n=1}^{\infty} (-1)^n \frac{\left(x - \frac{\pi}{2}\right)^{2n-1}}{(2n-1)!}.$$

EXAMPLE 9. **(A handy trick)** We know that

$$1 + x + x^2 + x^3 + \cdots = \sum_{n=0} x^n = \frac{1}{1-x} \quad \text{valid for } |x| < 1.$$

we can get further valid sums by differentiating the above:

$$\begin{aligned} 1 + 2x + 3x^2 + 4x^3 + \cdots &= \sum_{n=0}^{\infty} (n+1)x^n \\ &= \frac{d}{dx}\left(\frac{1}{1-x}\right) = \frac{1}{(1-x)^2}, \quad \text{valid for } |x| < 1. \end{aligned}$$

Further valid sums can be obtained by differentiating.

EXERCISES

1. Find the Maclaurin series expansion for each of the functions below:

   (a) $\dfrac{1}{1-x^2}$

   (b) $\dfrac{2x}{1-2x^2}$

   (c) $\dfrac{1}{(1-x)^2}$

   (d) $\dfrac{1}{(1-x^2)^2}$

   (e) $x^2 \sin x$

   (f) $\sin^2 x$ (Hint: Use a double-angle identity.)

   (g) $\dfrac{1}{(1-x)^3}$

   (h) $\ln(1+x^2)$

   (i) $\tan^{-1} 4x$

   (j) xe^{x^2}

2. Find the Maclaurin series expansion for the rational function $f(x) = \dfrac{x+1}{x^2+x+1}$. (Don't try to do this directly; use an appropriate trick.)

3. Sum the following series:

   (a) $\displaystyle\sum_{n=0}^{\infty} (x+1)^n$

   (b) $\displaystyle\sum_{n=1}^{\infty} n(x+1)^n$

   (c) $\displaystyle\sum_{n=0}^{\infty} \frac{(-1)^n x^n}{(n+1)!}$

   (d) $\displaystyle\sum_{n=1}^{\infty} \frac{(-1)^{n+1} x^n}{n}$

   (e) $\displaystyle\sum_{n=1}^{\infty} \frac{(-1)^{n+1} x^{2n}}{n}$

   (f) $\displaystyle\sum_{n=1}^{\infty} n^2 x^n$

4. Sum the following numerical series:

   (a) $\displaystyle\sum_{n=0}^{\infty} \frac{(-1)^n \pi^{2n+1}}{(2n+1)!}$

   (b) $\displaystyle\sum_{n=1}^{\infty} \frac{(-1)^{n+1}(e-1)^n}{n}$

   (c) $\displaystyle\sum_{n=0}^{\infty} \frac{(-1)^n (\ln 2)^n}{(n+1)!}$

   (d) $\displaystyle\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n 2^{2n}}$

   (e) $\displaystyle\sum_{n=1}^{\infty} \frac{n^2}{3^n}$

5. Consider the function defined by setting $f(x) = \ln(1 + \sin x)$.

   (a) Determine the Maclaurin series expansion for $f(x)$ through the x^4 term.

   (b) Using (a) determine the Maclaurin series expansion for the function $g(x) = \ln(1 - \sin x)$.

   (c) Using (a) and (b), determine the Maclaurin series expansion for $\ln \sec x$.

6. Consider the following integral

$$I = \int_0^1 \left(\int_0^1 \frac{dy}{1 - xy} \right) dx.$$

   (a) Using integration by parts, show that the internal integral is equal to $\dfrac{-\ln(1-x)}{x}$.

   (b) Determine the Maclaurin series expansion for this.

   (c) Use a term-by-term integration to show that

$$I = \sum_{n=1}^{\infty} \frac{1}{n^2},$$

   (which, as mentioned on page 276 is $= \frac{\pi^2}{6}$. See the footnote.[20])

7. Here's a self-contained proof that $\displaystyle\sum_{n=1}^{\infty} = \frac{\pi^2}{6}$. (See the footnote.[21])

   Step 1. To show that for any positive integer m,

$$\sum_{k=1}^{m} \cot^2 \frac{k\pi}{2m+1} = \frac{m(2m-1)}{3}.$$

   To complete step 1, carry out the following arguments.

[20] Using the variable change $u = \frac{1}{2}(y + x)$, $v = \frac{1}{2}(y - x)$, one can show in the above "double integral" is equal to $\frac{\pi^2}{6}$, giving an alternative proof to that alluded in HH, Exercise 15, page 228. Showing that the double integral has the correct value requires some work!

[21] This is distilled from I. Papadimitriou, *A Simple Proof of the Formula* $\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}$. *Amer. Math. Monthly* **80**, 424–425, 1973.

(i) By equating the imaginary parts of DeMoivre's formula

$$\cos n\theta + i \sin n\theta = (\cos\theta + i \sin\theta)^n = \sin^n\theta(\cot\theta + i)^n,$$

obtain the identity

$$\sin n\theta = \sin^n\theta \left\{ \binom{n}{1} \cot^{n-1}\theta - \binom{n}{3} \cot^{n-3}\theta + \binom{n}{5} \cot^{n-5}\theta - \cdots \right\}.$$

(ii) Let $n = 2m + 1$ and express the above as

$$\sin(2m+1)\theta = \sin^{2m+1}\theta P_m(\cot^2\theta), \quad 0 < \theta < \frac{\pi}{2},$$

where $P_m(x)$ is the polynomial of degree m given by

$$P_m(x) = \binom{2m+1}{1} x^m - \binom{2m+1}{3} x^{m-1} + \binom{2m+1}{5} x^{m-2} - \cdots .$$

(iii) Conclude that the real numbers

$$x_k = \cot^2\left(\frac{k\pi}{2m+1}\right), \quad 1 \le k \le m,$$

are zeros of $P_m(x)$, **and that they are all distinct**. Therefore, $x_1, x_2, \ldots, x_m$ comprise **all** of the zeros of $P_m(x)$.

(iv) Conclude from part (iii) that

$$\sum_{k=1}^{m} \cot^2\left(\frac{k\pi}{2m+1}\right) = \sum_{k=1}^{m} x_k = \binom{2m+1}{3} \Big/ \binom{2m+1}{1} = \frac{m(2m-1)}{3},$$

proving the claim of Step 1.

Step 2. Starting with the familiar inequality $\sin x < x < \tan x$ for $0 < x < \pi/2$, show that

$$\cot^2 x < \frac{1}{x^2} < 1 + \cot^2 x, \quad 0 < x < \frac{\pi}{2}.$$

Step 3. Put $x = \dfrac{k\pi}{2m+1}$, where k and m are positive integers and $1 \le k \le m$, and infer that

$$\sum_{k=1}^{m} \cot^2\left(\frac{k\pi}{2m+1}\right) < \frac{(2m+1)^2}{\pi^2} \sum_{k=1}^{m} \frac{1}{k^2} < m + \sum_{k=1}^{m} \cot^2\left(\frac{k\pi}{2m+1}\right).$$

Step 4. Use step 1 to write the above as

$$\frac{m(2m-1)}{3} < \frac{(2m+1)^2}{\pi^2} \sum_{k=1}^{m} \frac{1}{k^2} < m + \frac{m(2m-1)}{3}.$$

Step 5. Multiply the inequality of step 4 through by $\dfrac{\pi^2}{4m^2}$ and let $m \to \infty$. What do you get?

5.4.2 Error analysis and Taylor's theorem

In this final subsection we wish to address two important questions:

Question A: If $P_n(x)$ is the Maclaurin (or Taylor) polynomial of degree n for the function $f(x)$, how good is the approximation $f(x) \approx P_n(x)$? More precisely, how large can the error $|f(x) - P_n(x)|$ be?

Question B: When can we say that the Maclaurin or Taylor series of $f(x)$ actually converges to $f(x)$?

The answers to these questions are highly related.

The answer to both of these questions is actually contained in **Taylor's Theorem with Remainder**. Before stating this theorem, I want to indicate that this theorem is really just a generalization of the **Mean Value Theorem**, which I'll state below as a reminder.

Mean Value Theorem. *Let f be a differentiable function on some open interval I. If a and x are both in I, then there exists a real number c between a and x such that*

$$\frac{f(x) - f(a)}{x - a} = f'(c).$$

Put differently, there exists a number c between a and x such that

$$\boxed{f(x) = f(a) + f'(c)(x - a)}.$$

Having been reminded of the **Mean Value Theorem**, perhaps now **Taylor's Theorem with Remainder** won't seem so strange. Here it is.

Taylor's Theorem with Remainder. *Let f be an infinitely-differentiable function on an open interval I. If a and x are in I, and if n is a non-negative integer, then there exists a real number c between a and x such that*

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots$$

$$\cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n + \underbrace{\frac{f^{(n+1)}(c)}{(n+1)!}(x-a)^{n+1}}_{\text{this is the remainder}}.$$

PROOF.[22] We start by proving that, for all $n \geq 0$

$$f(x) = f(a)+f'(a)(x-a)+\frac{f''(a)}{2!}+\cdots+\frac{f^{(n)}(a)}{n!}(x-1)^n+\int_a^x \frac{f^{(n+1)}(t)}{n!}(x-t)^n\,dt.$$

Note that since

$$\int_a^x f'(t)\,dt = f(x) - f(a),$$

then a simple rearrangement gives

$$f(x) = f(a) + \int_a^x f'(t)\,dt,$$

which is the above statement when $n = 0$. We take now as our induction hypothesis, that

$$f(x) = f(a)+f'(a)(x-a)+\frac{f''(a)}{2!}(x-a)^2+\cdots+\frac{f^{(n-1)}(a)}{(n-1)!}(x-a)^n+\int_a^x \frac{f^{(n)}(t)}{(n-1)!}(x-t)^{n-1}$$

is true.

We evaluate the integral using integration by parts with the substitution

[22] Very few textbooks at this level provide a proof; however, since the two main ingredients are induction and integration by parts, I felt that giving the proof would be instructive reading for the serious student.

$$\begin{array}{ll} u = f^{(n)}(t) & dv = \dfrac{(x-t)^{(n-1)}}{(n-1)!}\,dt. \\ du = f^{(n+1)}(t)\,dt & v = -\dfrac{(x-t)^n}{n!} \end{array}$$

From the above, one obtains

$$\begin{aligned} \int_a^x \frac{f^{(n)}(t)}{(n-1)!}(x-t)^{(n-1)}\,dt &= -f^{(n)}(t)\frac{(x-t)^n}{n!}\Big|_a^x + \int_a^x \frac{f^{(n+1)}(t)}{n!}(x-t)^n\,dt \\ &= \frac{f^{(n)}(a)}{n!}(x-a)^n + \int_a^x \frac{f^{(n+1)}(t)}{n!}(x-t)^n\,dt \end{aligned}$$

Plugging this into the induction hypothesis shows that the original statement is correct for all n.

Next, note that if $F = F(t)$ is a continuous function of t, then one has that

$$\int_a^b F(t)\,dt = F(c)(b-a)$$

for some number c between a and b. (Indeed $F(c)$ is the average value of F on the interval $[a,b]$.) Using the substitution $u = (x-t)^{(n+1)}$, and applying the above observation, we have

$$\begin{aligned} (n+1)\int_a^x F(t)(x-t)^n\,dt &= \int_0^{(x-a)^{(n+1)}} F(x - \sqrt[n+1]{u})\,du \\ &= F(x - \sqrt[n+1]{\alpha})(x-a)^{n+1}, \end{aligned}$$

where α is between 0 and $(x-a)^{(n+1)}$. If we set $c = x - \sqrt[n+1]{\alpha}$ we see that c is between a and x and that

$$(n+1)\int_a^x F(t)(x-t)^n\,dt = F(c)(x-a)^{(n+1)}.$$

Now set $F(t) = \dfrac{f^{(n+1)}(t)}{n!}$ and apply the above to infer that

$$\int_a^x \frac{f^{(n+1)}(t)}{n!}(x-t)^n\,dt = \frac{f^{(n+1)}(c)}{(n+1)!}(x-a)^{n+1}.$$

This completes the proof.

The above remainder (i.e., error term) is called the **Lagrange form of the error**.

We'll conclude this subsection with some examples.

EXAMPLE 1. As a warm-up, let's prove that the Maclaurin series for $\cos x$ actually converges to $f(x) = \cos x$ for all x. We have, by **Taylor's Theorem with Remainder**, that

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots \pm \frac{x^{2n}}{(2n!)} + \frac{f^{(2n+1)}(c)}{(2n+1)!}x^{2n+1},$$

for some real number c between 0 and x. Since all derivatives of $\cos x$ are $\pm \sin x$ or $\pm \cos x$, we see that $\left|f^{(2n+1)}(c)\right| \leq 1$. This means that for fixed x, if we let $n \to \infty$, then the remainder

$$\frac{f^{(2n+1)}(c)}{(2n+1)!}\, x^{2n+1} \to 0,$$

proving that

$$\cos x \;=\; 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n)!}.$$

EXAMPLE 2. Here's a similar example. By **Taylor's Theorem with Remainder**, we have for $f(x) = \ln(1+x)$, that

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots \pm \frac{x^n}{n} + \frac{f^{(n+1)}(c)}{(n+1)!}x^{n+1},$$

for some real number c between 0 and x. It is easy to verify that the Maclaurin series for $\ln(1+x)$ has interval of convergence $-1 < x \leq 1$, so we need to insist that x is in this interval. Since $f^{(n+1)}(c) = \dfrac{n!}{(1+c)^{n+1}}$, we see that the error term satisfies

$$\left|\frac{f^{(n+1)}(c)}{(n+1)!}x^{n+1}\right| = \left|\, \frac{n!\, x^{n+1}}{(1+c)^{n+1}(n+1)!}\right| = \left|\, \frac{x^{n+1}}{(1+c)^{n+1}(n+1)}\right|$$

In this case, as long as $-\frac{1}{2} \leq x \leq 1$, then we are guaranteed that $\left|\frac{x^{n+1}}{(1+c)^{n+1}}\right| \leq 1$. Therefore, as $n \to \infty$ we have $\left|\frac{x^{n+1}}{(1+c)^{n+1}(n+1)}\right| \to 0$. Therefore, we at least know that for if $-\frac{1}{2} \leq x \leq 1$,

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots = \sum_{n=1}^{\infty} (-1)^{n-1} \frac{x^n}{n}.$$

In particular, this proves the fact anticipated on page 278, viz., that

$$\ln(2) = 1 - \frac{1}{2} + \frac{1}{3} - \cdots = \sum_{n=1}^{\infty} (-1)^{n-1} \frac{1}{n}.$$

EXAMPLE 3. One easily computes that the first two terms of the Maclaurin series expansion of $\sqrt{1+x}$ is $1 + \frac{x}{2}$. Let's give an upper bound on the error

$$\left|\sqrt{1+x} - \left(1 + \frac{x}{2}\right)\right|$$

when $|x| < 0.01$. By **Taylor's Theorem with Remainder**, we know that the absolute value of the error is given by $\left|f''(c)\frac{x^2}{2}\right|$, where c is between 0 and x, and where $f(x) = \sqrt{1+x}$. Since $f''(c) = \frac{-1}{4(1+c)^{3/2}}$, and since c is between 0 and x, we see that $1 + c \geq .99$ and so

$$\left|f''(c)\right| = \frac{1}{4(1+c)^{3/2}} \leq \frac{1}{4 \times .99^{3/2}} \leq .254.$$

This means that the error in the above approximation is no more than

$$\left|f''(c)\frac{x^2}{2}\right| \leq .254 \times \frac{(0.01)^2}{2} < .000013.$$

Another way of viewing this result is that, **accurate to four decimal places**, $\sqrt{1+x} = 1 + \frac{x}{2}$ whenever $|x| < 0.01$.

EXERCISES

1. Show that for all x, $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$.

2. Assume that you have a function f satisfying $f(0) = 5$ and for $n \geq 1$ $f^{(n)}(0) = \frac{(n-1)!}{2^n}$.

   (a) Write out $P_3(x)$, the third-degree Maclaurin polynomial approximation of $f(x)$.

   (b) Write out the Maclaurin series for $f(x)$, including the general term.

   (c) Use $P_3(x)$ to approximate $f(\frac{1}{2})$.

   (d) Assuming that $f^{(4)}(c) \leq \frac{1}{4}$ for all c satisfying $0 < c < \frac{1}{2}$, show that
   $|f(\frac{1}{2}) - P_3(\frac{1}{2})| < 10^{-3}$.

3. The function f has derivatives of all orders for all real numbers x. Assume $f(2) = -3$, $f'(2) = 5$, $f''(2) = 3$, and $f'''(2) = -8$.

   (a) Write the third-degree Taylor polynomial for f about $x = 2$ and use it to approximate $f(1.5)$.

   (b) The fourth derivative of f satisfies the inequality $|f^{(4)}(x)| \leq 3$ for all x in the closed interval $[1.5, 2]$. Use the Lagrange error bound on the approximation to $f(1.5)$ found in part (a) to explain why $f(1.5) \neq -5$.

   (c) Write the fourth-degree Taylor polynomial, $P(x)$, for $g(x) = f(x^2 + 2)$ about $x = 0$. Use P to explain why g must have a relative minimum at $x = 0$.

4. Let f be a function having derivatives of all orders for all real numbers. The third-degree Taylor polynomial for f about $x = 2$ is given by

$$P_3(x) \;=\; 7 - 9(x-2)^2 - 3(x-2)^3.$$

   (a) Find $f(2)$ and $f''(2)$.

   (b) Is there enough information given to determine whether f has a critical point at $x = 2$? If not, explain why not. If so, determine whether $f(2)$ is a relative maximum, a relative minimum, or neither, and justify your answer.

(c) The fourth derivative of f satisfies the inequality $\left|f^{(4)}(x)\right| \leq 6$ for all x in the closed interval $[0, 2]$. Use the Lagrange error bound on the approximation to $f(0)$ found in part (c) to explain why $f(0)$ is negative.

5. (a) Using mathematical induction, together with l'Hôpital's rule, prove that $\lim_{x\to\infty} \dfrac{P_n(x)}{e^x} = 0$ where $P_n(x)$ is a polynomial of degree n. Conclude that for any polynomial of degree n, $\lim_{x\to\pm\infty} \dfrac{P_n(x)}{e^{x^2}} = 0$.

(b) Show that $\lim_{x\to 0} \dfrac{P(\frac{1}{x})}{e^{\frac{1}{x^2}}} = 0$, where P is a polynomial. (Let $y = \frac{1}{x}$, and note that as $x \to 0,\ y \to \pm\infty$.)

(c) Let $f(x) = e^{-\frac{1}{x^2}}, x \neq 0$ and show by induction that $f^{(n)}(x) = Q_n\left(\frac{1}{x}\right) e^{-\frac{1}{x^2}}$, where Q_n is some polynomial (though not necessarily of degree n).

(d) Conclude from parts (b) and (c) that

$$\lim_{x\to 0} \frac{d^n}{dx^n}\left(e^{-1/x^2}\right) = 0$$

for all $n \geq 0$.

(e) What does all of this say about the Maclaurin series for e^{-1/x^2}?

5.5 Differential Equations

In this section we shall primarily consider **first-order ordinary**[23] **differential equations** (ODE), that is, differential equations of the form $y' = F(x, y)$. If the function F is linear in y, then the ODE is called a **linear** ordinary differential equation. A linear differential equation is therefore expressible in the form $y' = p(x)y + q(x)$, where p and q are functions defined on some common domain. In solving an ODE, we expect that an arbitrary constant will come as the result of an integration and would be determined by specifying an **initial value** of the

[23] to be distinguished from "partial" differential equations.

solution $y = y(x)$. This results in the **initial value problem** of the form

$$y' = p(x)y + q(x), \quad y(a) = y_0.$$

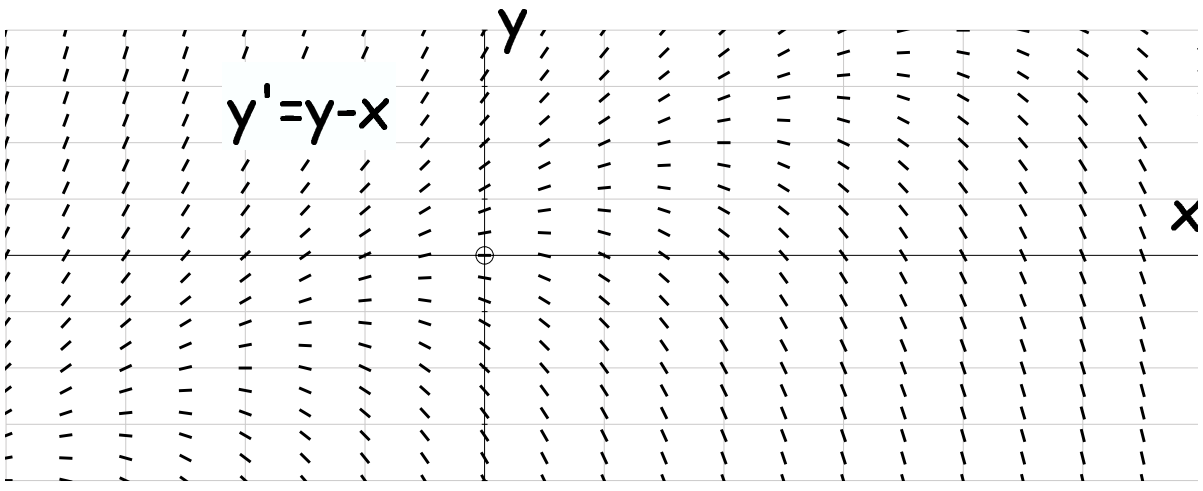
A good example of a commonly-encountered **nonlinear** ODE is the so-called **logistic differential equation**,

$$y' = ky(1 - y), \quad y(0) = y_0.$$

5.5.1 Slope fields

A good way of getting a preliminary feel for differential equations is through their **slope fields**. That is, given the differential equation (not necessarily linear) in the form $y' = F(x, y)$ we notice first that the slope at the point (x, y) of the solution curve $y = y(x)$ is given by $F(x, y)$. Thus, we can represent this by drawing at (x, y) a short line segment of slope $F(x, y)$. If we do this at enough points, then a visual image appears, called the **slope field** of the ODE. Some examples will clarify this; we shall be using the graphics software *Autograph* to generate slope fields.

EXAMPLE 1. Consider the ODE $y' = y - x$. The slope field is indicated below:

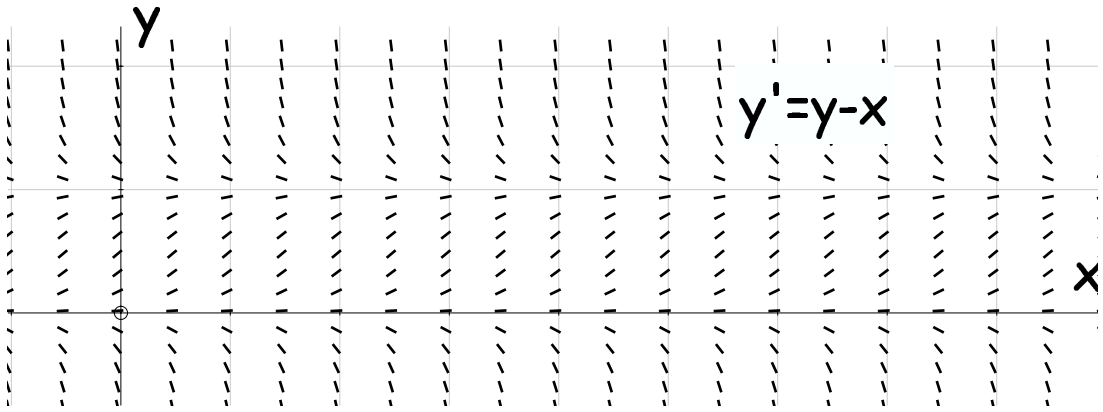


From the above slope field it appears that there might be a **linear** solution of the form $y = mx + b$. We can check this by substituting into the differential equation:

$$m = \frac{d}{dx}(mx + b) \;=\; mx + b - x,$$

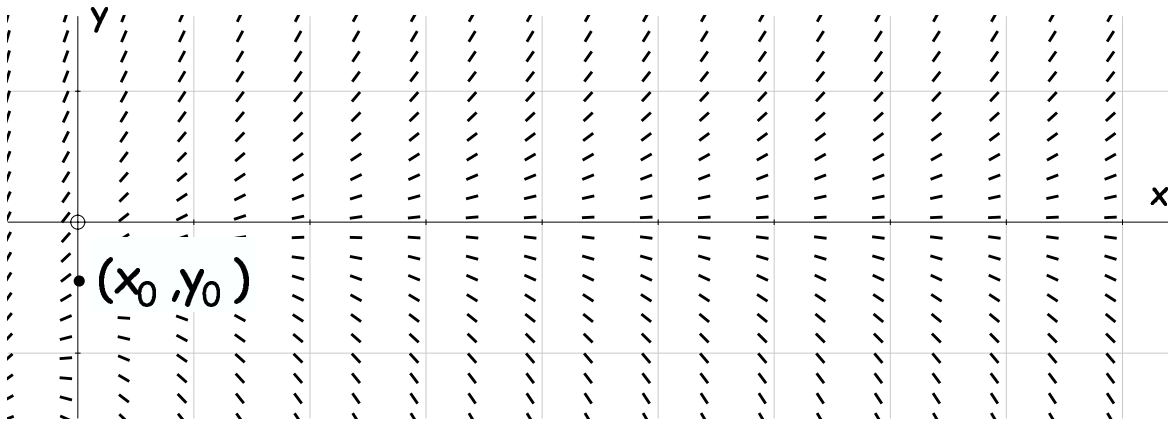
which immediately implies that $m = 1$ and $b = m = 1$.

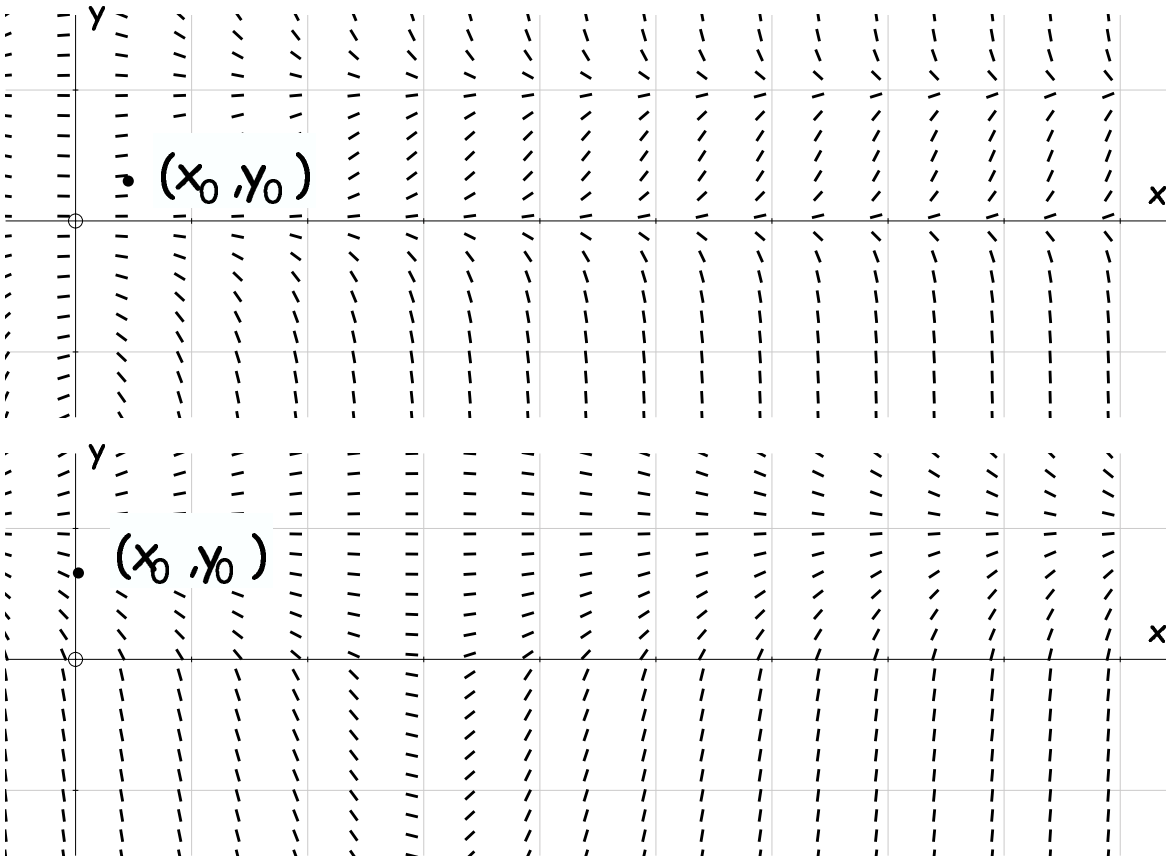
EXAMPLE 2. The logistic differential equation $y' = 3y(1 - y)$ has slope field given to the right. What we should be able to see from this picture is that if we have an initial condition of the form $y(0) = y_0 > 0$, then the solution $y = y(x)$ will satisfy $\lim_{x\to\infty} y = 1$.



EXERCISES

1. For each of the slope fields given below, sketch a solution curve $y = y(x)$ passing through the initial point (x_0, y_0).





2. Show that $y = Ke^{2x} - \frac{1}{4}(2x+1)$ is a solution of the linear ODE $y' = 2y + x$ for any value of the constant K.

3. Find a first-order linear ODE having $y = x^2 + 1$ as a solution. (There are many answers.)

4. In each case below, verify that the linear ODE has the given function as a solution.

   (a) $xy' + y = 3x^2, \quad y = x^2.$

   (b) $y' + 2xy = 0, \quad y = e^{-x^2}.$

   (c) $2x^2y'' + 3xy' - y = 0, \quad y = \sqrt{x}, \ x > 0.$

5. Consider the n-th order linear ODE with **constant coefficients**:

$$y^{(n)} + a_{n-1}y^{(n-1)} + \cdots + a_1y' + y \ = \ 0. \qquad (5.1)$$

   Assume that the associated **characteristic polynomial**

$$C(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

has a real zero α, i.e., that $C(\alpha) = 0$. Show that a solution of the ODE (5.1) is $y = e^{\alpha x}$.

5.5.2 Separable and homogeneous first-order ODE

Most students having had a first exposure to differential and integral calculus will have studied **separable** first-order differential equations. These are of the form

$$\frac{dy}{dx} \;=\; f(x)g(y)$$

whose solution is derived by an integration:

$$\int \frac{dy}{g(y)} \;=\; \int f(x)\,dx.$$

EXAMPLE 1. Solve the differential equation $\dfrac{dy}{dx} = -2yx$.

SOLUTION. From

$$\int \frac{dy}{y} \;=\; -\int 2x\,dx.$$

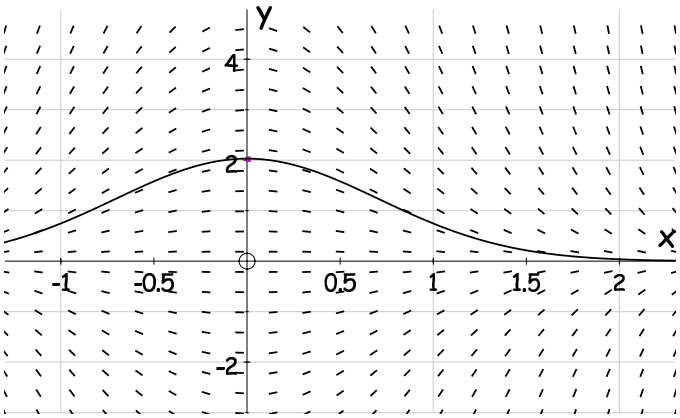
we obtain

$$\ln|y| \;=\; -x^2 + C,$$

where C is an arbitrary constant. Taking the natural exponential of both sides results in $|y| = e^{-x^2+C} = e^C e^{-x^2}$. However, if we define $K = e^C$, and if we allow K to take on negative values, then there is no longer any need to write $|y|$; we have, therefore the general solution

$$y \;=\; Ke^{-x^2},$$

which describes a rapidly-decreasing exponential function of x. The slope field, together with the particular solution with the initial condition $y(0) = 2$ is indicated to the right.



Some first-order ODE are not separable as they stand, but through a change of variables can be transformed into a separable ODE. Such is the case of ODE of the form

$$\frac{dy}{dx} \;=\; F\left(\frac{y}{x}\right), \tag{5.2}$$

for some function F. A change of independent variable

$$v \;=\; \frac{y}{x}$$

will accomplish this. To see this, we note that

$$y \;=\; vx, \quad \frac{dy}{dx} \;=\; x\frac{dv}{dx} + v;$$

with respect to x and v the ODE (5.2) becomes

$$x\frac{dv}{dx} + v \;=\; F(v).$$

The variables x and v separate easily, resulting in the ODE

$$\frac{1}{F(v) - v}\frac{dv}{dx} \;=\; \frac{1}{x},$$

which can be solved in principle as above.

EXAMPLE 2. The first-order ODE

$$2x^2\frac{dy}{dx} = x^2 + y^2$$

can be reduced to the form (5.2) by dividing both sides by $2x^2$:

$$\frac{dy}{dx} = \frac{x^2 + y^2}{2x^2} = \frac{1}{2}\left[1 + \left(\frac{y}{x}\right)\right].$$

Setting $v = \frac{y}{x}$ as above reduces the above ODE to

$$x\frac{dv}{dx} + v = \frac{1}{2}(1 + v^2);$$

that is,

$$2x\frac{dv}{dx} = v^2 - 2v + 1.$$

After separating the variables, we arrive at the equation

$$\int \frac{2\,dv}{(v-1)^2} = \int \frac{dx}{x}.$$

Integrating and simplifying yields

$$v = 1 - \frac{2}{\ln|x| + 2C}.$$

Replace v by y/x, set $c = 2C$ and arrive at the final general solution

$$y = x - \frac{2x}{\ln|x| + c}.$$

We define a function $F(x, y)$ to be **homogeneous** of **degree** k if for all real numbers t such that (tx, ty) is in the domain of F we have $F(tx, ty) = t^k F(x, y)$. Therefore, the function $F(x, y) = x^2 + y^2$ is homogeneous of degree 2, whereas the function $F(x, y) = \sqrt{x}/y$ is homogeneous of degree $-\frac{1}{2}$.

A first-order **homogeneous** ODE is of the form

$$M(x, y)\frac{dy}{dx} + N(x, y) = 0,$$

where $M(x, y)$ and $N(x, y)$ are both homogeneous of the **same degree**. These are important since they can always be reduced to the form (5.2). Indeed, suppose that M and N are both homogeneous of degree k. Then we work as follows:

$$\begin{aligned} \frac{dy}{dx} &= -\frac{N(x,y)}{M(x,y)} \\ &= -\frac{x^k N(1, y/x)}{x^k M(1, y/x)} \\ &= -\frac{N(1, y/x)}{M(1, y/x)} = F\left(\frac{y}{x}\right) \end{aligned}$$

which is of the form (5.2), as claimed. Note that Example 2 above is an example of a homogeneous first-order ODE.

EXERCISES

In the following problems, find both the general solution as well as the particular solution satisfying the initial condition.

1. $y' = 2xy^2, \quad y(0) = -1$

2. $yy' = 2x, \quad y(0) = 1$

3. $3y^2y' = (1 + y^2)\cos x, \quad y(0) = 1$

4. $2y' = y(y - 2), \quad y(0) = 1$

5. $xyy' = 2y^2 - x^2, \quad y(1) = 1$

6. $y' = \dfrac{y}{x} - 3\left(\dfrac{y}{x}\right)^{4/3}, \quad y(2) = 1$

7. $3xy^2y' = 4y^3 - x^3, \quad y(2) = 0$

5.5.3 Linear first-order ODE; integrating factors

In this subsection we shall consider the general first-order linear ODE:

$$y' + p(x)y \;=\; q(x). \tag{5.3}$$

As we'll see momentarily, these are, in principle, very easy to solve. The trick is to multiply both sides of (5.3) by the **integrating factor**

$$\mu(x) \;=\; e^{\int p(x)dx}.$$

Notice first that $\mu(x)$ satisfies $\mu'(x) = p(x)\mu(x)$. Therefore if we multiply (5.3) through by $\mu(x)$ we infer that

$$\frac{d}{dx}(\mu(x)y) = \mu(x)y' + p(x)\mu(x)y = \mu(x)\,q(x),$$

from which we may conclude that

$$\mu(x)\,y \;=\; \int \mu(x)\,q(x)dx.$$

EXAMPLE 1. Find the general solution of the first-order ODE

$$(x+1)\,y' - y = x, \quad x > -1.$$

First of all, in order to put this into the general form (5.3) we must divide everything by $x+1$:

$$y' - \frac{1}{x+1}y = \frac{x}{x+1}.$$

This implies that an integrating factor is

$$\mu(x) \;=\; e^{-\int \frac{dx}{x+1}} \;=\; \frac{1}{x+1}.$$

Multiply through by $\mu(x)$ and get

$$\begin{aligned}\frac{y}{x+1} &= \int \frac{x\,dx}{(x+1)^2} \\ &= \int \frac{(x+1-1)\,dx}{(x+1)^2} \\ &= \int \left(\frac{1}{x+1} - \frac{1}{(x+1)^2}\right) dx \\ &= \ln(x+1) + \frac{1}{x+1} + C\end{aligned}$$

It follows, therefore, that

$$y = (x+1)\ln(x+1) + c(x+1),$$

where c is an arbitrary constant.

Exercises

1. Solve the following first-order ODE.
   (a) $xy' + 2y = 2x^2, \quad y(1) = 0$
   (b) $2x^2y' + 4xy = e^{-x}, \quad y(2) = 1.$
   (c) $xy' + (x-2)y = 3x^3e^{-x}, \quad y(1) = 0$
   (d) $y' \ln x + \dfrac{y}{x} = x, \quad y(1) = 0$
   (e) $y' + (\cot x)y = 3\sin x \cos x, \quad y(0) = 1$
   (f) $x(x+1)y' - y = 2x^2(x+1), \quad y(2) = 0$

2. The first-order **Bernoulli** ODE are of the form

   $$y' + p(x)y = q(x)y^n,$$

   where n is any number other than 1. Show that the substitution $u = y^{1-n}$ brings the above Bernoulli equation into the first-order linear ODE

   $$\frac{1}{1-n}u' + p(x)u = q(x).$$

3. Solve the Bernoulli ODE

(a) $y' + \dfrac{3}{x}\, y = x^2 y^2, \quad x > 0$

(b) $2y' + \dfrac{1}{x+1}\, y + 2(x^2 - 1)\, y^3 = 0$

5.5.4 Euler's method

In this final subsection we shall discuss a rather intuitive numerical approach to solving a first-order ODE of the form $y' = F(x, y),\ y_0 = y(x_0)$. What we do here is to specify a **step size**, say h, and proceed to approximate $y(x_1)$, $y(x_2)$, $y(x_3), \ldots$, where $x_1 = x_0 + h$, $x_2 = x_1 + h = x_0 + 2h,$ and so on.

The idea is that we use the first-order approximation

$$y(x_1) \approx y(x_0) + y'(x_0)(x_1 - x_0) = y(x_0) + y'(x_0)h.$$

Notice that $y'(x_0) = F(x_0, y_0)$; we set $y_1 = y(x_0) + F(x_0, y_0)h$, giving the approximation $y(x_1) \approx y_1$. We continue:

$$\begin{aligned} y(x_2) &\approx y(x_1) + y'(x_1)(x_2 - x_1) && \text{(first-order approximation)} \\ &\approx y_1 + y'(x_1)h && \text{(since } y(x_1) \approx y_1\text{)} \\ &\approx y_1 + F(x_1, y_1)h && \text{(since } F(x_1, y(x_1)) \approx F(x_1, y_1)\text{)}. \end{aligned}$$

Continuing in this fashion, we see that the approximation $y(x_{n+1}) \approx y_{n+1}$ at the new point $x = x_{n+1}$ is computed from the previous approximation $y(x_n) \approx y_n$ at the point x_n via

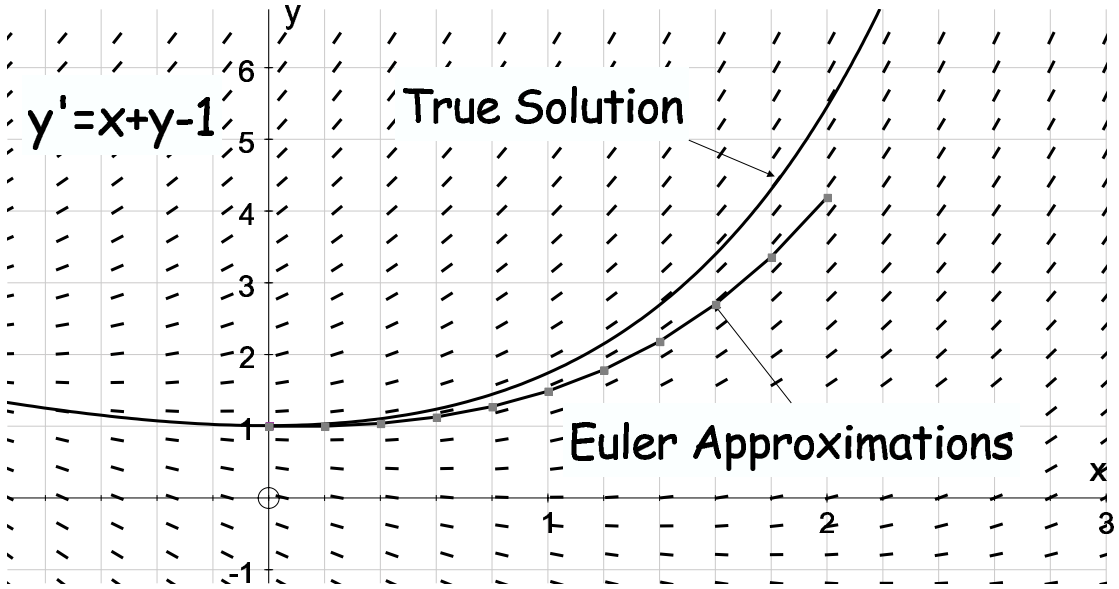
$$y(x_{n+1}) \ \approx \ y_{n+1} \ = \ y_n + F(x_n, y_n)h.$$

EXAMPLE 1. Approximate the solution of the ODE $y' = x + y - 1,\ y(0) = 1$ on the interval $[0, 2]$ using step size $h = 0.2$; note that $F(x, y) = x + y - 1$:

We can tabulate the results:

n	x_n	$y_n = y_{n-1}$ $+F(x_{n-1}, y_{n-1})$	n	x_n	$y_n = y_{n-1}$ $+F(x_{n-1}, y_{n-1})\,h$
0	0	1	6	1.2	1.788
1	.2	1	7	1.4	2.1832
2	.4	1.04	8	1.6	2.6998
3	.6	1.128	9	1.8	3.3598
4	.8	1.2736	10	2.0	4.1917
5	1.0	1.4883			

The figure below compares the exact solution with the approximations generated above.



EXERCISES

1. Give the exact solution of $y' = x + y - 1$, $y(0) = 1$. Tabulate these exact values against their approximations in the table below:

n	x_n	$y_n = y_{n-1} + F(x_{n-1}, y_{n-1})\,h$	$y(x_n)$
0	0	1	
1	.2	1	
2	.4	1.04	
3	.6	1.128	
4	.8	1.2736	
5	1.0	1.4883	
6	1.2	1.788	
7	1.4	2.1832	
8	1.6	2.6998	
9	1.8	3.3598	
10	2.0	4.1917	

2. Use the Euler method with $h = 0.1$ to find approximate values for the solution of the initial-value problem over the interval $[1, 2]$

$$xy' + y = 3x^2, \quad y(1) = -2.$$

Then solve exactly and compare against the approximations.

3. Do the same over the interval $[0, 1]$, $(h = 0.1)$ for

$$y' = 2xy + \frac{1}{y}, \quad y(0) = 1.$$

Chapter 6

Inferential Statistics

We shall assume that the student has had some previous exposure to elementary probability theory; here we'll just gather together some recollections.

The most important notion is that of a **random variable**; while we won't give a formal definition here we can still convey enough of its root meaning to engage in useful discussions. Suppose that we are to perform an experiment whose outcome is a numerical value X. That X is a **variable** follows from the fact that repeated experiments are unlikely to produce the same value of X each time. For example, if we are to toss a coin and let

$$X = \begin{cases} 1 & \text{if heads,} \\ 0 & \text{if tails,} \end{cases}$$

then we have a random variable. Notice that this variable X does not have a value until after the experiment has been performed!

The above is a good example of a **discrete random variable** in that there are only two possible values of X: $X = 0$ and $X = 1$. By contrast, consider the experiment in which I throw at dart at a two-dimensional target and let X measure the distance of the dart to the center (bull's eye). Here, X is still random (it depends on the throw), but can take on a whole continuum of values. Thus, in this case we call X a **continuous random variable**.

6.1 Discrete Random Variables

Let's start with an example which is probably familiar to everyone. We take a pair of fair dice and throw them, letting X be the sum of the dots showing. Of course, X is random as it depends on the outcome of the experiment. Furthermore X is discrete: it can only take on the integer values between 2 and 12. Finally, using elementary means it is possible to compute the probability that X assumes any one of these values. If we denote by $P(X = x)$ the probability that X assumes the value x, $2 \leq x \leq 12$ can be computed and tabulated as below:

x	2	3	4	5	6	7	8	9	10	11	12
$P(X = x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

The table above summarizes the **distribution** of the discrete random variable X. That is, it summarizes the individual probabilities $P(X = x)$, where x takes on any one of the allowable values. Furthermore, using the above distribution, we can compute probabilities of the form $P(x_1 \leq X \leq x_2)$; for example

$$P(2 \leq X \leq 5) = P(X = 2)+P(X = 3)+P(X = 4)+P(X = 5) \; = \; \frac{1}{36}+\frac{2}{36}+\frac{3}{36}+\frac{4}{36} = \frac{10}{36}.$$

It is reasonably clear that if X is an arbitrary discrete random variable whose possible outcomes are x_1, x_2, $x_3, \ldots$,

$$\sum_{i=1}^{\infty} P(X = x_i) \; = \; 1.$$

This of fundamental importance!

6.1.1 Mean, variance, and their properties

We define the **mean** μ_X (or **expectation** $E(X)$)[1] of the discrete random variable X by setting

[1]Some authors use the notation $\langle X \rangle$ for the mean of the discrete random variable X.

$$\mu_X \;=\; E(X) = \sum x_i P(X = x_i),$$

where the sum is over all possible values x_i which the random variable X can assume. As we'll see, the above is often an infinite series! This value can be interpreted as the average value of X over many observations of X. (We'll give a slightly more precise formulation of this in section **??**.) For example, if X is the random variable associated with the above dice game, then

$$\begin{aligned} E(X) \;&=\; 2 \times \frac{1}{36} + 3 \times \frac{2}{36} + 4 \times \frac{3}{36} + 5 \times \frac{5}{36} + 6 \times \frac{5}{36} + 7 \times \frac{6}{36} \\ &+\; 8 \times \frac{5}{36} + 9 \times \frac{4}{36} + 10 \times \frac{3}{36} + 11 \times \frac{2}{36} + 12 \times \frac{1}{36} \approx 7.14. \end{aligned}$$

Let X and Y be two discrete random variables; we wish to consider the mean $E(X+Y)$ of the sum $X+Y$. While it's probably intuitively plausible, if not downright obvious, that $E(X+Y) = E(X) + E(Y)$, this still deserves a proof.[2]

So we assume that X and Y are discrete random variables having means $E(X) = \mu_X$ and $E(Y) = \mu_Y$, respectively. Of fundamental importance to the ensuing analysis is that for any value x, then the probabilities $P(X = x)$ can be expressed in terms of conditional probabilities[3] on Y:

$$P(X = x) \;=\; \sum_{j=1}^{\infty} P(X = x \,|\, Y = y_j) P(Y = y_j). \tag{6.1}$$

Likewise, the probabilities $P(Y = y)$ can be similarly expressed in terms of conditional probabilities on X:

[2]Elementary textbooks typically only prove this under the simplifying assumption that X and Y are independent.

[3]Here, we have assumed that the students have already had some exposure to conditional probabilities. Recall that for any two events A and B the **probability of** A **conditioned on** B is given by

$$P(A \,|\, B) = \frac{P(A \text{ and } B)}{P(B)}.$$

$$P(Y = y) \;=\; \sum_{j=1}^{\infty} P(Y = y \,|\, X = x_i)P(X = x_i). \tag{6.2}$$

Having noted this, we now proceed:

$$\begin{aligned}
\mu_{X+Y} &= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}(x_i + y_j)P(X = x_i \text{ and } Y = y_j) \\
&= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} x_i P(X = x_i \text{ and } Y = y_j) \\
&\quad + \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} y_j P(X = x_i \text{ and } Y = y_j) \\
&= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} x_i P(X = x_i \,|\, Y = y_j)P(Y = y_j) \\
&\quad + \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} y_j P(Y = y_j \,|\, X = x_i)P(X = x_i) \\
&= \sum_{i=1}^{\infty} x_i \sum_{j=1}^{\infty} P(X = x_i \,|\, Y = y_j)P(Y = y_j) \\
&\quad + \sum_{j=1}^{\infty} y_j \sum_{i=1}^{\infty} P(Y = y_j \,|\, X = x_i)P(X = x_i) \\
&= \sum_{i=1}^{\infty} x_i P(X = x_i) + \sum_{j=1}^{\infty} y_i P(Y = y_i) \quad \text{by (6.1) and (6.2)} \\
&= \mu_X + \mu_Y,
\end{aligned}$$

proving that

$$E(X + Y) \;=\; E(X) + E(Y). \tag{6.3}$$

Next, note that if X is a random variable and if a and b are constants, then it's clear that $E(aX) = aE(X)$; from this we immediately infer (since b can be regarded itself as a random variable with mean b) that

$$E(aX + b) \;=\; aE(X) + b.$$

Next, we define the **variance** σ^2 (or Var(X)) of the random variable X having mean μ by setting $\sigma^2 = E((X-\mu)^2)$. The **standard deviation** σ is the non-negative square root of the variance. The mean and variance of a random variable are examples of **parameters** of a random variable.

We shall derive an alternate—and frequently useful—expression for the variance of the random variable X with mean μ. Namely, note that

$$\begin{aligned} \text{Var(X)} &= E((X-\mu)^2) \\ &= E(X^2 - 2\mu X + \mu^2) \\ &= E(X^2) - 2\mu E(X) + \mu^2 \quad \text{(by (6.3))} \\ &= E(X^2) - \mu^2. \end{aligned} \tag{6.4}$$

We turn now to the variance of the discrete random variable $X+Y$. In this case, however, we require that X and Y are **independent**. This means that for all values x and y we have

$$P(X = x \text{ and } Y = y) = P(X = x)P(Y = y).^4$$

In order to derive a useful formula for $\text{Var}(X+Y)$, we need the result that given X and Y are independent, then $E(XY) = E(X)E(Y)$; see Exercise 1, below. Using (6.4), we have

$$\begin{aligned} \text{Var}(X+Y) &= E((X+Y)^2) - \mu_{X+Y}^2 \\ &= E((X+Y)^2) - (\mu_X + \mu_Y)^2 \\ &= E(X^2 + 2XY + Y^2) - (\mu_X + \mu_Y)^2 \\ &= E(X^2) + E(2XY) + E(Y^2) - (\mu_X^2 + 2\mu_X\mu_Y + \mu_Y^2) \\ &= E(X^2) - \mu_X^2 + 2E(X)E(Y) - 2\mu_X\mu_Y + E(Y^2) - \mu_Y^2 \\ &= \text{Var}(X) + \text{Var}(Y). \end{aligned} \tag{6.5}$$

[4] An equivalent—and somewhat more intuitive—expression can be given in terms of conditional probabilities. Namely, two events A and B are equivalent precisely when $P(A \mid B) = P(A)$. In terms of discrete random variables X and Y, this translates into $P(X = x \mid Y = y) = P(X = x)$ for any possible values x of X and y of Y.

As you might expect, the above formula is *false* in general (i.e., when X and Y not independent); see Exercise 1, below. Using (6.5), we see immediately that if X is a discrete random variable, and if $Y = aX + b$, where a and b are real numbers, then we may regard b as a (constant) random variable, certainly independent of the random variable aX. Therefore,

$$\text{Var}(Y) = \text{Var}(aX + b) = \text{Var}(aX) + \text{Var}(b) = a^2\text{Var}(X),$$

where we have used the easily-proved facts that $\text{Var}(aX) = a^2\text{Var}(X)$ and where the variance of a constant random variable is zero (see Exercises 5 and 6, below).

We conclude this section with a brief summary of properties of mean and variance for discrete random variables.[5]

- If X is a random variable, and if a, b are real numbers, then $E(aX + b) = aE(X) + b$.
- If X is a random variable, and if a, b are real numbers, then $\text{Var}(aX + b) = a^2\text{Var}(X)$.
- If X and Y are random variables, then $E(X + Y) = E(X) + E(Y)$.
- If X and Y are *independent* random variables, then $E(XY) = E(X)E(Y)$.
- If X and Y are independent random variables, then $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$.

6.1.2 Weak law of large numbers (optional discussion)

In order to get a better feel for the meaning of the variance, we include the following two lemmas:

[5]These same properties are also true for continuous random variables!

LEMMA. (Markov's Inequality) *Let X be a non-negative discrete random variable. Then for any number $d > 0$, we have*

$$P(X \geq d) \leq \tfrac{1}{d}\, E(X).$$

PROOF. We define a new random variable Y by setting

$$Y \;=\; \begin{cases} d & \text{if } X \geq d \\ 0 & \text{otherwise.} \end{cases}.$$

Since $Y \leq X$, it follows that $E(X) \geq E(Y)$. Also note that Y has two possible values: 0 and d; furthermore,

$$E(Y) \;=\; dP(Y = d) \;=\; dP(X \geq d).$$

Since $E(X) \geq E(Y) = dP(X \geq d)$, the result following immediately.

LEMMA. (Chebyshev's Inequality) *Let X be a discrete random variable with mean μ and variance σ^2. Then for any $d > 0$ we have*

$$P(|X - \mu| \geq d) \;\leq\; \frac{\sigma^2}{d^2}.$$

PROOF. Define the random variable $Y = (X - \mu)^2$; it follows that $E(Y) = \sigma^2$. Applying Markov's inequality to Y we have

$$P(|X - \mu| \geq d) = P(Y \geq d^2) \leq \tfrac{1}{d^2} E(Y) = \frac{\sigma^2}{d^2},$$

as required.

We now assume that X_1, X_2, $\ldots$, X_n are random variables with the same mean μ; we denote the average of these random variables thus:

$$\overline{X} \;=\; \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

From what we've proved about the mean, we see already that $E(\overline{X}) = \mu$. In case the random variables $X_1, X_2, \ldots,$ have the same distribution, the **Weak Law of Large Numbers** says a bit more:

LEMMA. **(The Weak Law of Large Numbers)** *Assume that $X_1, X_2, \ldots, X_n, \ldots,$ is an infinite sequence of identically distributed random variables with mean μ (and having finite variance σ^2). Then for each $\epsilon > 0$*

$$\lim_{n\to\infty} P\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| > \epsilon\right) = 0.$$

PROOF. We set $S_n = X_1 + X_2 + \cdots + X_n$, and so $A_n = S_n/n$ has mean μ and variance σ^2/n. By Chebyshev's Inequality we have

$$P\left(\left|A_n - \mu\right| \geq \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2}.$$

Since $\epsilon > 0$ is fixed, the result is now obvious.

Notice that an equivalent formulation of the Weak Law of Large Numbers is the statement that for all $\epsilon > 0$ we have that

$$\lim_{n\to\infty} P\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| \leq \epsilon\right) = 1.$$

As you might expect, there is also a Strong Law of Large Numbers which is naively obtained by interchanging the limit and probability P; see the footnote.[6]

EXERCISES

1. Prove that if X and Y are discrete independent random variables, then $E(XY) = E(X)E(Y)$. Is this result still true if X and Y are **not** independent?

[6]That is to say, if $X_1, X_2, \ldots, X_n, \ldots,$ is an infinite sequence of identically distributed random variables with mean μ, then

$$P\left(\frac{X_1 + X_2 + \cdots + X_n}{n} \to \mu\right) = 1.$$

There is no requirement of finiteness of the variances.

2. Suppose that we draw two cards in succession, and without replacement, from a standard 52-card deck. Define the random variables X_1 and X_2 by setting

$$X_1 = \begin{cases} 1 & \text{if the first card drawn is red} \\ 0 & \text{if the first card drawn is black;} \end{cases}$$

similarly,

$$X_2 = \begin{cases} 1 & \text{if the second card drawn is red} \\ 0 & \text{if the second card drawn is black.} \end{cases}$$

   (a) Are X_1 and X_2 independent random variables?

   (b) Compute $P(X_1 = 1)$.

   (c) Compute $P(X_2 = 1)$.

3. Suppose that we have two dice and let D_1 be the result of rolling die 1 and D_2 the result of rolling die two. Show that the random variables $D_1 + D_2$ and D_1 are not independent. (This seems pretty obvious, right?)

4. We continue the assumptions of the above exercise and define the new random variable T by setting

$$T = \begin{cases} 1 & \text{if } D_1 + D_2 = 7 \\ 0 & \text{if } D_1 + D_2 \neq 7. \end{cases}$$

   Show that T and D_1 **are** independent random variables. (This takes a bit of work.)

5. Let X be a discrete random variable and let a be a real number. Prove that $\text{Var}(aX) = a^2\text{Var}(X)$.

6. Let X be a constant-valued random variable. Prove that $\text{Var}(X) = 0$. (This is very intuitive, right?)

7. John and Eric are to play the following game with a fair coin. John begins by tossing the coin; if the result is heads, he wins and the game is over. If the result is tails, he hands the coin over to

Eric, who then tosses the coin. If the result is heads, Eric wins; otherwise he returns the coin to John. They keep playing until someone wins by tossing a head.

(a) What is the probability that Eric wins on his first toss?

(b) What is the probability that John wins the game?

(c) What is the proability that Eric wins the game?

8. Let n be a fixed positive integer. Show that for a randomly-selected positive integer x, the probability that x and n are relatively prime is $\dfrac{\phi(n)}{n}$. (Hint: see Exercise 20 on page 64.)

9. Consider the following game. Toss a fair coin, until the first head is reached. The payoff is simply 2^n dollars, where n is the number of tosses needed until the first head is reached. Therefore, the payoffs are

No. of tosses	1	2	3	$\cdots$	n	$\cdots$
Payoff	\$2	\$4	\$8	$\cdots$	$\$2^n$	$\cdots$

How much would you be willing to pay this game? \$10? \$20? Ask a friend; how much would she be willing to play this game? Note that the expected value of this game is infinite![7]

6.1.3 The random harmonic series (optional discussion)

We close this section with an interesting example from analysis. We saw on page 265 the harmonic series $\displaystyle\sum_{n=1}^{\infty} \frac{1}{n}$ diverges and on page 278 we saw that the alternating harmonic series $\displaystyle\sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n}$ converges (to $\ln 2$; see page 302). Suppose now that $\epsilon_1, \epsilon_2, \ldots$ is a random sequence of $+1$s and -1s, where we regard each ϵ_k as a random variable with $P(\epsilon_k = 1) = P(\epsilon_k = -1) = 1/2$. Therefore each ϵ_k has mean 0 and variance 1. What is the probability that $\displaystyle\sum_{n=1}^{\infty} \frac{\epsilon_n}{n}$ converges?

[7] Thus, we have a paradox, often called the **St. Petersburg paradox**.

We can give an intuitive idea of how one can analyze this question, as follows. We start by setting $X_k = \dfrac{\epsilon_k}{k^{2/3}},\ k = 1, 2, \ldots,$ and note that $E(X_k) = 0$ and $\mathrm{Var}(X_k) = \dfrac{1}{k^{4/3}}$. Now set

$$S_n \ = \ \sum_{k=1}^{n} X_k \ = \ \sum_{k=1}^{n} \frac{\epsilon_k}{k^{2/3}}.$$

It follows immediately that S_n has mean 0 and (finite) variance $\sum_{k=1}^{n} \dfrac{1}{k^{4/3}} < 4$. (See footnote[8])

Under these circumstances it follows that the above infinite sum $\sum_{k=1}^{n} X_k$ actually converges with probability 1.[9] Furthermore, the same arguments can be applied to show that as long as $p > 1/2$, then the random series $\sum_{n=1}^{\infty} \dfrac{\epsilon_n}{n^p}$ also converges with probability 1.

We turn now to some relatively commonly-encountered discrete random variables, the **geometric**, the **binomial**, the **negative binomial**, the **hypergeometric**, and the **Poisson** random variables.

6.1.4 The geometric distribution

Consider the following game (experiment). We start with a coin whose probability of heads is p; therefore the probability of tails is $1 - p$. The game we play is to keep tossing the coin until a head is obtained. The random variable X is the number of trials until the game ends. The distribution for X as follows:

x	1	2	3	$\cdots$	n
$P(X = x)$	p	$p(1-p)$	$p(1-p)^2$	$\cdots$	$p(1-p)^{n-1}$

Therefore, the expectation of X is given by the infinite sum:

[8]Note that $\sum_{k=1}^{n} \dfrac{1}{k^{4/3}} < \sum_{k=1}^{\infty} \dfrac{1}{k^{4/3}} < 1 + \int_1^{\infty} x^{-4/3}\, dx = 4.$

[9]This can be inferred from the **Kolmogorov Three-Series Theorem**, see, e.g., Theorem 22.8 of P. Billingsley, *Probability and Measure, 2nd ed.* John Wiley & Sons, New York, 1986.

$$E(X) = \sum_{n=1}^{\infty} nP(X=n) = \sum_{n=1}^{\infty} np(1-p)^{n-1} = p\sum_{n=1}^{\infty} n(1-p)^{n-1}.$$

Note that

$$\begin{aligned}\sum_{n=1}^{\infty} n(1-p)^{n-1} &= \frac{d}{dx}(1+x+x^2+\cdots)\Big|_{x=1-p} \\ &= \frac{d}{dx}\left(\frac{1}{1-x}\right)\Big|_{x=1-p} \\ &= \frac{1}{(1-x)^2}\Big|_{x=1-p} \\ &= \frac{1}{p^2},\end{aligned}$$

which implies that the mean of the geometric random variable X is given by

$$E(X) = p\sum_{n=1}^{\infty} n(p-1)^{n-1} = \frac{1}{p}.$$

Notice that the smaller p becomes, the longer the game is expected to last.

Next, we turn to the variance of X. By Equation 6.4 we have

$$\begin{aligned}\text{Var(X)} &= E(X^2) - \frac{1}{p^2} \\ &= \sum_{n=1}^{\infty} n^2 P(X=n) - \frac{1}{p^2} \\ &= p\sum_{n=1}^{\infty} n^2(1-p)^{n-1} - \frac{1}{p^2}.\end{aligned}$$

Next,

$$\begin{aligned}
\sum_{n=1}^{\infty} n^2(1-p)^{n-1} &= \sum_{n=1}^{\infty} n(n-1)(1-p)^{n-1} + \sum_{n=1}^{\infty} n(1-p)^{n-1} \\
&= (1-p)\sum_{n=1}^{\infty} n(n-1)(1-p)^{n-2} + \sum_{n=1}^{\infty} n(1-p)^{n-1} \\
&= (1-p)\frac{d^2}{dx^2}(1+x+x^2+\cdots)\Big|_{x=1-p} + \frac{1}{p^2} \\
&= (1-p)\frac{d^2}{dx^2}\left(\frac{1}{1-x}\right)\Big|_{x=1-p} + \frac{1}{p^2} \\
&= \frac{2(1-p)}{p^3} + \frac{1}{p^2} = \frac{2-p}{p^3}
\end{aligned}$$

Therefore,

$$\mathrm{Var}(X) = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}$$

6.1.5 The binomial distribution

In this situation we perform n independent trials, where each trial has two outcomes—call them *success* and *failure*. We shall let p be the probability of success on any trial, so that the probability of failure on any trial is $1-p$. The random variable X is the total number of successes out of the n trials. This implies, of course, that the distribution of X is summarized by writing

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

The mean and variance of X are very easily computed once we realize that X can be expressed as a sum of n independent **Bernoulli** random variables. The Bernoulli random variable B is what models the tossing of a coin: it has outcomes 0 and 1 with probabilities $1-p$ and p, respectively. Very simple calculations shows that

$$E(B) = p \quad \text{and} \quad \mathrm{Var}(B) = p(1-p).$$

Next, if X is the binomial random variable with success probability p, then we may write

$$X \;=\; B_1 + B_2 + \cdots + B_n,$$

where each B_i is a Bernoulli random variable. It follows easily from what we already proved above that

$$E(X) \;=\; E(B_1) + E(B_2) + \cdots E(B_n) \;=\; np,$$

and

$$\mathrm{Var}(X) \;=\; \mathrm{Var}(B_1) + \mathrm{Var}(B_2) + \cdots + \mathrm{Var}(B_n) \;=\; np(1-p).$$

6.1.6 Generalizations of the geometric distribution

Generalization 1: The negative binomial distribution

Suppose that we are going to perform a number X of Bernoulli trials, each with success probability p, stopping after exactly r successes have occurred. Then it is clear that

$$P(X = x) \;=\; \binom{x-1}{r-1} p^r (1-p)^{x-r}.$$

In order to compute the mean and variance of X note that X is easily seen to be the sum of geometric random variables G_1, G_2, $\ldots$, G_r, where the success probability of each is p:

$$X \;=\; G_1 + G_2 + \cdots + G_r.$$

Using the results of (6.1.4) we have, for each $i = 1, 2, \ldots, r$, that

$$E(G_i) \;=\; \frac{1}{p}, \;\; \mathrm{Var}(G_i) \;=\; \frac{1-p}{p^2}.$$

It follows, therefore, that the mean and variance of the negative binomial random variable X are given by

$$E(X) = \frac{r}{p}, \quad \text{Var}(X) = \frac{r(1-p)}{p^2}.$$

The name "inverse binomial" would perhaps be more apt, as the following direct comparison with the binomial distribution reveals:

Binomial Random Variable X	Negative Binomial Random Variable Y
number of successes in n trials	number of trials needed for r successes
$E(X) = np, \ \text{Var}(X) = np(1-p)$	$E(Y) = \frac{r}{p}, \ \text{Var}(Y) = \frac{r(1-p)}{p^2}$

Generalization 2: The coupon problem

Suppose that in every cereal box there is a "prize," and that there are, in all, three possible prizes. Assume that in a randomly purchased box of cereal the probability of winning any one of the prizes is the same, namely 1/3. How many boxes of cereal would you expect to buy in order to have won all three prizes? It turns out that the natural analysis is to use a sum of geometric random variables.

We start by defining three independent random variables X_1, X_2, and X_3, as follows. X_1 is the number of trials to get the first new prize; note that X_1 is really not random, as the only possible value of X_1 is 1. Nonetheless, we may regard X_1 as a geometric random variable with probability $p = 1$. X_2 is the number of trials (boxes) needed to purchase in order to get the second new prize, after the first prize is already won. Clearly X_2 is also a geometric random variable, this time with $p = 2/3$. Finally, X_3 is the number of boxes needed to purchase to get the third (last) new prize after we already have two distinct prizes, and so X_3 is geometric with $p = 1/3$. Therefore, if X is the number of boxes purchased before collecting the complete set of three prizes, then $X = X_1 + X_2 + X_3$, which represents X as a sum of geometric random variables.

From the above, computing $E(X)$ is now routine:

$$E(X) \;=\; E(X_1+X_2+X_3) \;=\; E(X_1)+E(X_2)+E(X_3) = 1+\frac{3}{2}+3 = \frac{11}{2}.$$

The generalization of the above problem to that of finding the expected number of boxes needed to purchase before collecting all of n different prizes should now be routine! The answer in this case is

$$E(X) = 1 + \frac{n}{n-1} + \frac{n}{n-2} + \cdots + \frac{n}{2} + n = n\sum_{k=1}^{n}\frac{1}{k}.$$

Generalization 3: Fixed sequences of binary outcomes

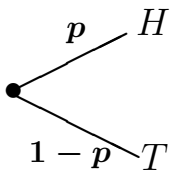
In section 6.1.4 we considered the experiment in which a coin is repeatedly tossed with the random variable X measuring the number of times before the first occurrence of a head. In the present section we modify this to ask such questions such as:

- *what is the expected number of trials before obtaining two heads in a row?*, or

- *what is the expected number of trials before seeing the sequence HT?*

What makes the above questions interesting is that on any two tosses of a fair coin, whereas the probability of obtaining the sequences HH and HT are the same, the expected waiting times before seeing these sequences differ. The methods employed here can, in principle, be applied to the study of any pre-determined sequence of "heads" and "tails."

In order to appreciate the method employed, let's again consider the geometric distribution. That is, assume that the probability of flipping a head (H) is p, and that X measures the number of trials before observing the first head. We may write $X = B+(1-B)(1+Y)$, where B is the Bernoulli random variable with $P(B = 1) = p$ and $P(B = 0) = 1 - p$, and where Y and X have the same distribution. (See the tree diagram to the right.)



($B = 1$ and the game is over)

($B = 0$ and we start the experiment over again with one trial already having been performed.)

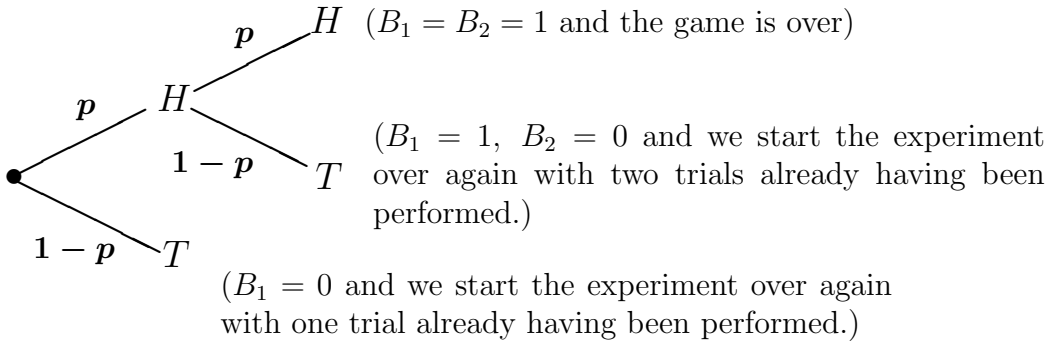
It follows, therefore that

$$\begin{aligned} E(X) &= E(B + (1 - B)(1 + Y)) \\ &= E(B) + E(1 - B)E(1 + Y) \text{ (since } B \text{ and } Y \text{ are independent)} \\ &= p + (1 - p)(1 + E(X)) \\ &= 1 + (1 - p)E(X). \end{aligned}$$

Solving for $E(X)$ quickly yields the correct result, viz., $E(X) = 1/p$.

The above method quickly generalizes to sequences. Let's consider tossing a coin with $P(\text{heads}) = p$, stopping after two consecutive heads are obtained. Letting X be the number of trials, Y have the same distribution as X and letting B_1 and B_2 be independent Bernoulli random variables, we may set

$$X = B_1 + B_2 + B_1(1 - B_2)(2 + Y) + (1 - B_1)(1 + Y). \qquad (6.6)$$



Computing the expectation of both sides of (6.6) quickly yields

$$E(X) = 2p^2 + p(1-p)(2 + E(X)) + (1-p)(1 + E(X)),$$

from which it follows that

$$E(X) \;=\; \frac{1+p}{p^2}.$$

Note that if the coin is **fair**, then the expected waiting time before seeing two heads in a row is 6.

Similar analyses can be applied to computing the expected waiting time before seeing the sequence HT (and similar) sequences, see Exercises 8, 9, and 10 on page 341.

6.1.7 The hypergeometric distribution

This distribution is modeled by a box containing N marbles, of which n of these are of a particular type ("successful" marbles) and so there are $N - n$ "unsuccessful" marbles. If we draw k marbles without replacement, and if X is the random variable which measures the number of successful marbles drawn, then X has distribution given by

$$P(X = m) \;=\; \frac{\binom{n}{m}\binom{N-n}{k-m}}{\binom{N}{k}}, \quad m = 0, 1, 2, \ldots, \max\{n, k\}.$$

From the above it follows that the mean of X is given by the sum

$$E(X) \;=\; \sum_{m=0}^{\max\{n,k\}} \frac{m\binom{n}{m}\binom{N-n}{k-m}}{\binom{N}{k}}.$$

We can calculate the above using simple differential calculus. We note first that

$$\left[\frac{d}{dx}(x+1)^n\right](x+1)^{N-n} = n(x+1)^{N-1} = \frac{n}{N}\frac{d}{dx}(x+1)^N.$$

Now watch this:

$$\begin{aligned}
\sum_{k=0}^{N}\left(\sum_{m=0}^{n} m\binom{k}{m}\binom{N-n}{k-m}\right)x^k &= \sum_{m=0}^{n} m\binom{n}{m}x^m \cdot \sum_{p=0}^{N-n}\binom{N-n}{p}x^p \quad \left(\begin{array}{l}\text{this takes}\\ \text{some thought!}\end{array}\right)\\
&= \left[x\frac{d}{dx}(x+1)^n\right](x+1)^{N-n}\\
&= \frac{xn}{N}\frac{d}{dx}(x+1)^N\\
&= \frac{n}{N}\sum_{k=0}^{N} k\binom{N}{k}x^k;
\end{aligned}$$

equating the coefficients of x^k yields

$$\sum_{m=0}^{n} m\binom{n}{m}\binom{N-n}{k-m} = \frac{nk}{N}\binom{N}{k}.$$

This immediately implies that the mean of the hypergeometric distribution is given by

$$E(X) \;=\; \frac{nk}{N}.$$

Turning to the variance, we have

$$E(X^2) \;=\; \sum_{m=0}^{n} \frac{m^2\binom{n}{m}\binom{N-n}{k-m}}{\binom{N}{k}}.$$

Next, we observe that

$$\begin{aligned} x^2\left[\frac{d^2}{dx^2}(x+1)^n\right](x+1)^{N-n} &= n(n-1)x^2(x+1)^{N-2} \\ &= x^2\frac{n(n-1)}{N(N-1)}\frac{d^2}{dx^2}(x+1)^N. \end{aligned}$$

Next comes the hard part (especially the first equality):

$$\begin{aligned} \sum_{k=0}^{N}\left(\sum_{m=0}^{n} m(m-1)\binom{n}{m}\binom{N-n}{k-m}\right) &= \sum_{m=0}^{n}(m-1)\binom{n}{m}x^m\cdot\sum_{p=0}^{N-n}\binom{N-n}{p}x^p \\ &= \left[x^2\frac{d^2}{dx^2}(x+1)^n\right](x+1)^{N-n} \\ &= x^2\frac{n(n-1)}{N(N-1)}\frac{d^2}{dx^2}(x+1)^N \\ &= \frac{n(n-1)}{N(N-1)}\sum_{k=0}^{N}k(k-1)\binom{N}{k}x^k. \end{aligned}$$

Just as we did at a similar juncture when computing $E(X)$, we equate the coefficients of x^k, which yields the equality

$$\sum_{m=0}^{n} m(m-1)\binom{n}{m}\binom{N-n}{k-m} = \frac{n(n-1)k(k-1)}{N(N-1)}\binom{N}{k}.$$

The left-hand sum separates into two sums; solving for the first sum gives

$$\begin{aligned} \sum_{m=0}^{n} m^2\binom{n}{m}\binom{N-n}{k-m} &= \frac{n(n-1)k(k-1)}{N(N-1)}\binom{N}{k} + \sum_{m=0}^{n} m\binom{n}{m}\binom{N-n}{k-m} \\ &= \frac{n(n-1)k(k-1)}{N(N-1)}\binom{N}{k} + \frac{nk}{N}\binom{N}{k}, \end{aligned}$$

which implies that

$$E(X^2) = \frac{n(n-1)k(k-1)}{N(N-1)} + \frac{nk}{N}.$$

Finally, from this we obtain the variance of the hypergeometric distribution:

$$\begin{aligned}
\mathrm{Var}(X) &= E(X^2) - E(X)^2 \\
&= \frac{n(n-1)k(k-1)}{N(N-1)} + \frac{nk}{N} - \left(\frac{nk}{N}\right)^2 \\
&= \frac{nk(N-n)(N-k)}{N^2(N-1)}.
\end{aligned}$$

6.1.8 The Poisson distribution

The **Poisson random variable** can be thought of as the limit of a binomial random variable in the following sense. First of all, assume that Y is the binomial random variable which measures the number of successes in n trials and where the probability of each trial is p. As we saw above, the mean of this random variable is $\mu_Y = np$. Now, rather than limiting the number of trials, we take the limit as $n \to \infty$ **but holding fixed the mean** $\mu = \mu_Y$. We call the resulting random variable the **Poisson random variable with mean** μ. If we denote this by X, then the distribution of X is computed as follows:

$$\begin{aligned}
P(X=k) &= \lim_{n\to\infty} \binom{n}{k} p^k (1-p)^{n-k} \\
&= \lim_{n\to\infty} \binom{n}{k} \left(\frac{\mu}{n}\right)^k \left(1-\frac{\mu}{n}\right)^{n-k} \qquad (\text{since } \mu = np) \\
&= \lim_{n\to\infty} \frac{n(n-1)\cdots(n-k+1)}{k!} \left(\frac{\mu}{n}\right)^k \left(1-\frac{\mu}{n}\right)^{n-k} \\
&= \lim_{n\to\infty} \frac{n(n-1)\cdots(n-k+1)}{n^k} \left(\frac{\mu^k}{k!}\right) \left(1-\frac{\mu}{n}\right)^{n-k} \\
&= \left(\frac{\mu^k}{k!}\right) \lim_{n\to\infty} \left(1-\frac{\mu}{n}\right)^{n-k} \quad \left(\text{since } \lim_{n\to\infty} \frac{n(n-1)\cdots(n-k+1)}{n^k} = 1\right) \\
&= \left(\frac{\mu^k}{k!}\right) \lim_{n\to\infty} \left(1-\frac{\mu}{n}\right)^{n} \left(1-\frac{\mu}{n}\right)^{-k} \\
&= \left(\frac{\mu^k}{k!}\right) \lim_{n\to\infty} \left(1-\frac{\mu}{n}\right)^{n}
\end{aligned}$$

Next, note that the limit above is a 1^∞ indeterminate form; taking the natural log and applying l'Hôpital's rule, we have

$$\lim_{n\to\infty}\left(1-\frac{\mu}{n}\right)^n \;=\; e^{-\mu}, \tag{6.7}$$

and so it follows that

$$P(X=k) \;=\; \frac{e^{-\mu}\mu^k}{k!}.$$

This gives the distribution of the Poisson random variable!

The Poisson distribution is often used to model events over time (or space). One typical application is to model traffic accidents (per year) at a particular intersection of two streets. For example, if our traffic data suggests that there are roughly 2.3 accidents/year at this intersection, then we can compute the probability that in a given year there will be less than 2 accidents or more than 4 accidents. These translate into the respective probabilities $P(X \le 1)$ and $P(X \ge 5)$. Specifically,

$$\begin{aligned} P(X\le 1) &= P(X=0)+P(X=1) \\ &= \frac{e^{-2.3}2.3^0}{0!}+\frac{e^{-2.3}2.3^1}{1!} \\ &= e^{-2.3}(1+2.3) = 3.3e^{-2.3} \approx .331. \end{aligned}$$

In the same vein,

$$\begin{aligned} P(X\ge 5) &= 1-P(X\le 4) \\ &= 1-\left(\frac{e^{-2.3}2.3^0}{0!}+\frac{e^{-2.3}2.3^1}{1!}+\frac{e^{-2.3}2.3^2}{2!}+\frac{e^{-2.3}2.3^3}{3!}+\frac{e^{-2.3}2.3^4}{4!}\right) \\ &= 1-e^{-2.3}\left(1+2.3+\frac{2.3^2}{2!}+\frac{2.3^3}{3!}+\frac{2.3^4}{4!}\right) \\ &\approx 0.081. \end{aligned}$$

We expect that the mean of the Poisson random variable is μ; however, a direct proof is possible as soon as we remember the Maclaurin series expansion for e^x (see Exercise 1 on page 302). We have that

$$\begin{aligned} E(X) &= \sum_{k=0}^{\infty} kP(X=k) \\ &= \sum_{k=0}^{\infty} k\frac{e^{-\mu}\mu^k}{k!} \\ &= \sum_{k=0}^{\infty} \frac{e^{-\mu}\mu^{k+1}}{k!} \\ &= \mu e^{-\mu}\sum_{k=0}^{\infty}\frac{\mu^k}{k!} = \mu e^{-\mu}e^{\mu} = \mu, \end{aligned}$$

as expected.

Similarly,

$$\begin{aligned} \mathrm{Var}(X) &= E(X^2) - \mu^2 \\ &= \sum_{k=0}^{\infty} k^2 P(X=k) - \mu^2 \\ &= \sum_{k=0}^{\infty} k^2\frac{e^{-\mu}\mu^k}{k!} - \mu^2 \\ &= e^{-\mu}\sum_{k=0}^{\infty} k\frac{\mu^{k+1}}{k!} - \mu^2 \\ &= \mu e^{-\mu}\sum_{k=0}^{\infty}(k+1)\frac{\mu^k}{k!} - \mu^2 \\ &= \mu e^{-\mu}\sum_{k=0}^{\infty} k\frac{\mu^k}{k!} + \mu e^{-\mu}\sum_{k=0}^{\infty}\frac{\mu^k}{k!} - \mu^2 \\ &= \mu^2 e^{-\mu}\sum_{k=0}^{\infty}\frac{\mu^k}{k!} + \mu - \mu^2 \\ &= \mu^2 + \mu - \mu^2 = \mu. \end{aligned}$$

That is to say, $\mathrm{Var}(X) = \mu = E(X)$.

Exercises

1. Suppose that you are going to toss a fair coin 200 times. Therefore, you know that the expected number of heads obtained is 100, and the variance is 50. If X is the actual number of heads, what does Chebyshev's Inequality say about the probability that X deviates from the mean of 50 by more than 15?

2. Suppose that you are playing an arcade game in which the probability of winning is $p = .2$.

   (a) If you play 100 times, how many games do you expect to win?

   (b) If you play 100 times, what is the probability that you will win more than 30 games?

   (c) If you play until you win exactly 20 games, how many games will you expect to play?

   (d) If you stop after winning 20 games, what is the probability that this happens no later than on the 90-th game?

3. Prove that the sum of two independent binomial random variables with the same success probability p is also binomial with success probability p.

4. Prove that the result of Exercise 3 is correct if " binomial" is replaced with "negative binomial."

5. Prove that the sum of two independent Poisson random variables is also Poisson.

6. Suppose that N men check their hats before dinner. However, the clerk then randomly permutes these hats before returning the hats to the N men. What is the expected number of men who will receive their own hats? (This is actually easier than it looks: let B_i be the Bernoulli random variable which is 1 if the man receives his own hat and 0 otherwise. While B_1, B_2, ..., B_N are not independent (why not?), the expectation of the sum is still the sum of the expectations.)

7. My motorcycle has a really lousy starter; under normal conditions my motorcycle will start with probability 1/3 when I try to start it. Given that I need to recharge my battery after every 200 attempts at starting my motorcycle, compute the probability that I will have to recharge my battery after one month. (Assume that I need to start my motorcycle twice each day.)

8. On page 334 we saw that if we toss a fair coin in succession, the expected waiting time before seeing two heads in a row is 6. Now play the same game, stopping after the sequence HT occurs. Show that expected length of this game is 4. Does this seem intuitive?

9. Do the same as in the above problem, comparing the waiting times before seeing the sequences THH versus THT. Are the waiting times the same?

10. (A bit harder) Show that on tossing a coin whose probability of heads is p the expected waiting time before seeing k heads in a row is $\dfrac{1-p^k}{(1-p)p^k}$.

11. As we have seen the binomial distribution is the result of witnessing one of two results, often referred to as success and failure. The **multinomial distribution** is where there is a finite number of outcomes, O_1, O_2, $\ldots, O_k$. For example we may consider the outcomes to be your final grade in this class: A, B, C, D, or F. Suppose that on any given trial the probability that outcome O_i results is p_i, $i = 1, 2, \ldots, k$ Naturally, we must have that $p_1 + p_2 + \cdots + p_k = 1$. Again, to continue my example, we might assume that my grades are assigned according to a more-or-less traditional distribution:

    A: 10%
    B: 20%
    C: 40%
    D: 20%
    F: 10%

    If we perform n trials, and we denote by X_i the number of times we witness outcome O_i, then the probabilities in question are of the form $P(X_1 = x_1, X_2 = x_2, \ldots, X_k =$

x_k), where $x_1 + x_2 + \cdots + x_k = n$. A little thought reveals that these probabilities are given by

$$P(X_1 = x_1,\, X_2 = x_2,\, \ldots, X_k = x_k) = \binom{n}{x_1, x_2, \ldots, x_k} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k},$$

where $\binom{n}{x_1, x_2, \ldots, x_k}$ is the **multinomial coefficient**

$$\binom{n}{x_1, x_2, \ldots, x_k} = \frac{n!}{x_1! x_2! \cdots x_k!}.$$

So, suppose that my grading distribution is as follows, and that I have 20 students. Compute the following probabilities:

(a) P(3 As, 6 Bs, 8 Cs, 2 Ds, and 1 F)

(b) P(3 or 4 As, 5 or 6 Bs, 8 Cs, 2 Ds, and 1 F)

(c) P(everyone passes (D or better)

(d) P(at most 5 people get As)

12. (Gambler's Ruin) Suppose that we have two players, player A and player B and that players A and B have between them N dollars. Players A and B now begin their game where player A tosses a fair coin, winning \$1 from B whenever she tosses a head and losing and losing \$1 (and giving it to B) whenever she tosses a tail. Holding N fixed, let $p_i = P(\text{A bankrupts B} \,|\, \text{A started with } i \text{ dollars})$. (It is clear that $p_0 = 0$ and that $p_N = 1$.)

(a) Let E_i be the event that A bankrupts B, given that A started with i dollars; then $P(E_i) = p_i$. Now argue that

$$\begin{aligned} p_i &= P(E_i \,|\, A \text{ wins the first game })P(A \text{ wins the first game }) \\ &\quad + P(E_i \,|\, B \text{ wins the first game })P(B \text{ wins the first game }) \\ &= \frac{1}{2}p_{i+1} + \frac{1}{2}p_{i-1}. \end{aligned}$$

(b) From the above, obtain $p_i = ip_1,\ i = 1, 2, \ldots, N$.

(c) That is, if A starts with a dollars and b starts with b dollars, then the probability that A bankrupts B is $\dfrac{a}{a+b}$.

The point of the above is that if A plays against someone with a lot of capital—like a casino—then the probability that A eventually goes bankrupt is very close to zero, even if the game is fair! This is known as **gambler's ruin**.

13. Generalize the results of Exercise 12 to the case when the probability of tossing head is p. That is, compute the probability that A bankrupts B, given that A starts with a dollars and B has $N - a$ dollars.

14. (An open-ended question) Note that the Poisson distribution with mean 2 and the geometric distribution with $p = .5$ both have the same mean and variance. How do these distributions compare to each other? Try drawing histograms of both. Note that the same can be said for the Poisson distribution with mean $2k$ and the negative binomial ($p = .5$, stopping at the k-th success).

15. Suppose we have a large urn containing 350 white balls and 650 blue balls. We select (without replacement) 20 balls from this urn. What is the probability that exactly 5 are white? Does this experiment differ significantly from an appropriately-chosen model based on the binomial distribution? What would the appropriate binomial approximation be?

16. Suppose that we have a large urn containing 1000 balls, exactly 50 of which are white (the rest are blue). Select 20 balls. Without knowing whether the selection was with or without replacement, estimate

    (a) the expected number of white balls in the sample;

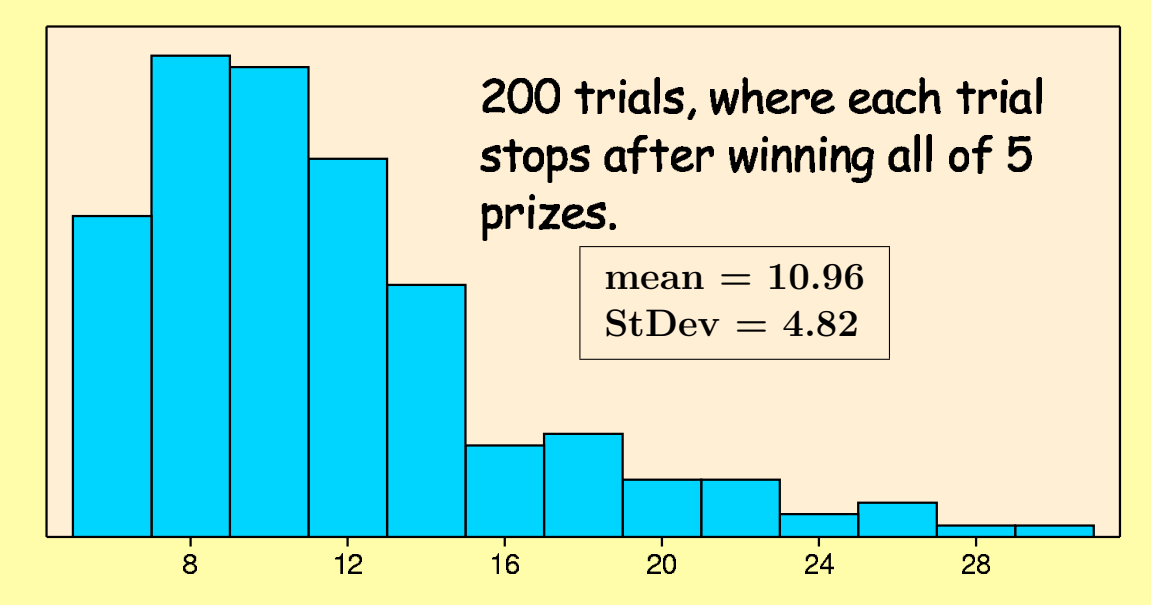
    (b) the probability that you selected at most 2 white balls (using a Poisson model);

    (c) the probability that you selected at most 2 white balls (using a hypergeometric model);

(d) the probability that you selected at most 2 white balls (using a binomial model).

17. Let X be the random variable associated with the coupon problem (see page 331), where n is the number of prizes involved. Compute the variance of X.

18. Consider the following Minitab-generated histogram of 200 trials, where one stops after winning all of $m = 5$ prizes.



(a) Is the sample mean close to the theoretical mean obtained above?

(b) How close does this histogram appear to that of a Poisson distribution (with the theoretical mean)?

(c) The TI code below will simulate playing the above game N times where M is the total number of prizes. The result of playing the N games is stored in list variable L_3. Have fun with it!

```
PROGRAM: PRIZES
  :Input "NO OF PRIZES: ", M      :While B< 1
  :Input "NO OF TRIALS: ", N      :C+1→ C
  :0 → A                          :randInt(1,M)→D
  :For(I,1,N)                     :L2(D)+1→ L2(D)
  :For(L,1,M)                     :1 → B
  :0 → L2(L)                      :For(J,1,M)
  :End                            :B*L2(J)→B
  1 → B                           :End
  :For(J,1,M)                     :End
  :B*L2(J)→ B                     :C→ L3(I)
  :End                            :End
  :0 → C                          :Stop
```

19. Let X be the binomial random variable with success probability p and where X measures the number of successes in n trials. Define the new random variable Y by setting $Y = 2X - n$. Show that $Y = y,\ -n \le y \le n$ can be interpreted as the total earnings after n games, where in each game we win \$1 with each success and we lose \$1 with each failure. Compute the mean and variance of Y.

20. Continuing with the random variable Y given above, let T be the random variable which measures the number of trials needed to first observe $Y = 1$. In other words, T is the number of trials needed in order to first observe one's cumulative earnings reach \$1. Therefore $P(T = 1) = p$, $P(T = 2) = 0$, $P(T = 3) = p^2(1 - p)$. Show that $P(T = 2k + 1) = C(k)p^{k+1}(1 - p)^k$, where $C(k) = \dfrac{1}{k+1}\binom{2k}{k}$, $n = 0, 1, 2, \ldots$, are the **Catalan numbers**.

21. We continue the thread of Exercise 20, above. Show that if $p = 1/2$—so the game is fair—then the expected time to first earn \$1 is *infinite*! We'll outline two approaches here: a short (clever?) approach and a more direct approach.

    (a) Let E be the expected waiting time and use a tree diagram as on page 333 to show that $E = \frac{1}{2} + \frac{1}{2}(2E + 1)$, which implies

that E must be infinite.

(b) Here we'll give a nuts and bolts direct approach.[10] Note first that the expectation E is given by

$$E = \sum_{k=0}^{\infty} \frac{(2k+1)C(k)}{2^{2k+1}}, \text{ where } C(k) = \frac{1}{k+1}\binom{2k}{k}, \ k = 0, 1, 2, \ldots.$$

(i) Show that $C(k) = \dfrac{2 \cdot 6 \cdot 10 \cdots (4k-2)}{(k+1)!}$, $k \geq 1$ (This is a simple induction).[11]

(ii) Conclude that $C(k) = \dfrac{1}{k+1} \prod_{m=1}^{k} \left(4 - \dfrac{2}{m}\right)$.

(iii) Conclude that $C(k)2^{-(2k-1)} = \dfrac{2}{k+1} \prod_{m=1}^{k} \left(1 - \dfrac{1}{2m}\right)$.

(iv) Using the fact that $\ln x > x - 1$, show that $\ln\left(1 - \frac{1}{2m}\right) > -\frac{1}{m}$, $m = 1, 2, \ldots$

(v) Conclude that

$$\begin{aligned} \prod_{l=1}^{k} \left(1 - \frac{1}{2l}\right) &= e^{\ln \prod_{l=1}^{k}\left(1-\frac{1}{2l}\right)} \\ &= e^{\sum_{l=1}^{k} \ln\left(1-\frac{1}{2l}\right)} \\ &> e^{-\sum_{l=1}^{k} \frac{1}{l}} \\ &> e^{-(1+\ln n)} = \frac{1}{ne} \quad \text{(see Exercise 5 on page 269)} \end{aligned}$$

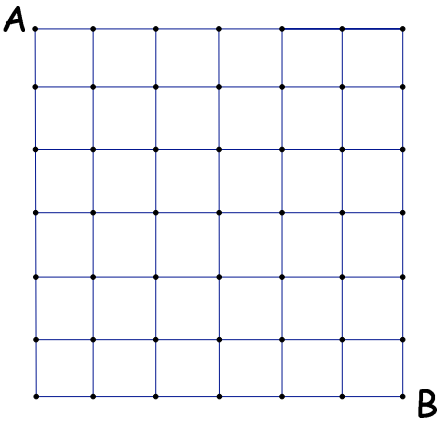
(vi) Finish the proof that $E = \infty$ by showing that the series for E is asymptotically a multiple of the divergent harmonic series.

22. Here are two more simple problems where Catalan numbers appear.

[10] I am indebted to my lifelong friend and colleague Robert Burckel for fleshing out most of the details.

[11] This still makes sense if $k = 0$ for then the numerator is the "empty product," and hence is 1.

(a) Suppose that we wish to move along a square grid (a 6×6 grid is shown to the right) where we start from the extreme northwest vertex (A) and move toward the extreme southeast vertex (B) in such a way that we always move "toward" the objective, i.e., each move is either to the right (east) or down (south).



A moment's thought reveals that there are $\binom{12}{6}$ such paths. What is the probability that a random path from A to B will always be above or on the diagonal drawn from A to B? (Answer: For the grid to the right the probability is $C(6)/\binom{12}{6} = 1/7$.) This result generalizes in the obvious way to $n \times n$ grids.

(b) Suppose this time that we have $2n$ people, each wishing to purchase a \$10 theater ticket. Exactly n of these people has only a \$10 bill, and the remaining n people has only a \$20 bill. The person selling tickets at the ticket window has no change. What is the probability that a random lineup of these $2n$ people will allow the ticket seller to make change from the incoming receipts? (This means, for instance, that the first person buying a ticket cannot be one of the people having only a \$20 bill.)

23. Suppose that we have a room with n politicians and that they are going to use the following "democratic" method for selecting a leader. They will distribute n identical coins, each having the probability of heads being p. The n politicians each toss their respective coins in unison; if a politician's coin comes up heads, and if all of the others come up tails, then this politician becomes the leader. Otherwise, they all toss their coins again, repeating until a leader has been chosen.

(a) Show that the probability of a leader being chosen in a given round is $np(1-p)^{n-1}$.

(b) Show that the maximum probability for a leader to be chosen in a given round occurs when $p = 1/n$.

(c) Show that if $n >> 0$, and if $p = 1/n$, then the probability that a leader is chosen in a given round is $\approx 1/e$. (See Equation 6.7, page 338.)

24. Suppose that in a certain location, the average annual rainfall is regarded as a continuous random variable, and that the rainfalls from year to year are independent of each other. Prove that in n years the expected number of record rainfall years is given by the harmonic series $1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}$.

6.2 Continuous Random Variables

Your TI graphing calculator has a random-number generator. It's called rand; find it! Invoking this produces a random number. Invoking this again produces another. And so on. What's important here is that rand represents a continuous (or nearly so!) random variable.

Let's look a bit closer at the output of rand. Note first that the random numbers generated are real numbers between 0 and 1. Next, note that the random numbers are **independent**, meaning that the value of one occurence of rand has no influence on any other occurence of rand.[12]

A somewhat more subtle observation is that rand is a **uniformly distributed** random variable. What does this mean? Does it mean that, for example

$$P(\text{rand} = .0214) \;=\; P(\text{rand} = 1/\pi)?$$

You will probably convince yourselves that this is not the meaning, as it is almost surely true that both sides of the above are 0, regardless of

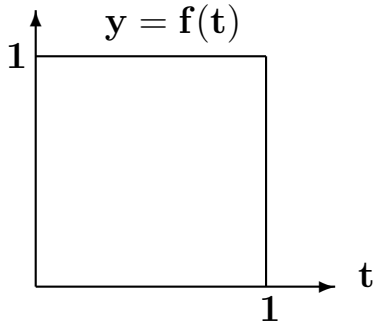
[12]Of course, this isn't the technical definition of "independence." A slighly more formal definition of independence of random variables X and Y is that

$$P(a \le X \le b \text{ and } c \le Y \le d) = P(a \le X \le b)P(c \le Y \le d).$$

the meaning of uniformity! What uniformity means is that for any two numbers x_1 and x_2 and any small number ϵ satisfying $x_1 \pm \epsilon$, $x_2 \pm \epsilon \in [0, 1]$

$$P(x_1 - \epsilon \leq \mathsf{rand} \leq x_1 + \epsilon) = P(x_2 - \epsilon \leq \mathsf{rand} \leq x_2 + \epsilon).$$

A much simpler description of the above is through the so-called **density function** for the random variable rand. This has the graph given below:

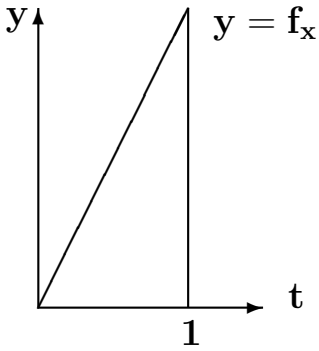


The way to interpret this—and any other density curve $y = f(x)$—is that the probability of finding a value of the corresponding random variable X between the values a and b is simply the **area under the density curve** from a to b:

$$P(a \leq X \leq b) \;=\; \int_a^b f(x)\,dx.$$

For the uniform distribution this means simply that, for example, $P(\mathsf{rand} \leq 2/3) = 2/3$, that $P(\mathsf{rand} > .25) = .75$, $P(.05 \leq \mathsf{rand} < .6) = .55$, and so on.

Let's consider another continuous random variable, defined in terms of its density function. Namely, let X be the random variable whose density function $y = f_x(t)$ is as given below:



Two important observations are in order.

(a) For any observation x of X, $0 \leq x \leq 1$.

(b) $\int_0^1 f(x)\,dx = 1$

(See Exercise 1, below.)

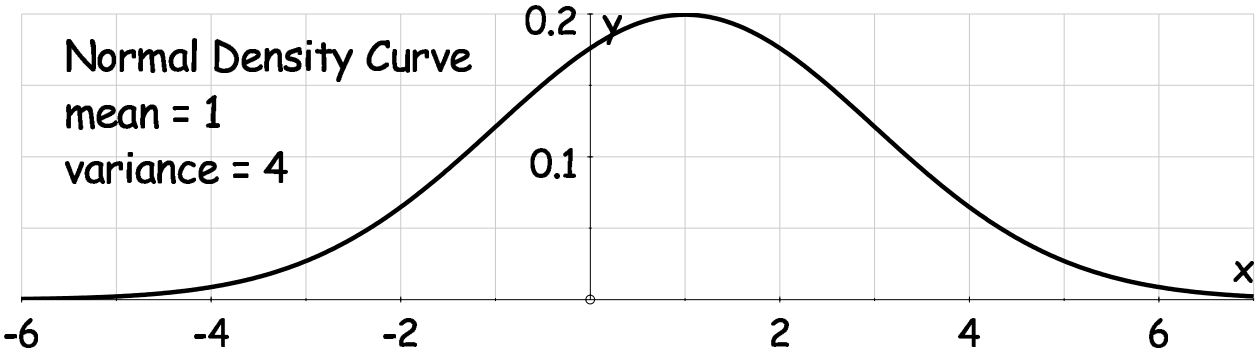
We see that the above density curve has quite a bit of "skew" to it; in particular it's clear that a random measurement of X is much more likely to produce a value greater than .5 than less than .5.

6.2.1 The normal distribution

The **normal density function** has the general form

$$f(x) \;=\; \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

where μ and σ are constants, or **parameters**[13] of the distribution. The graph is indicated below for $\mu = 1$ and $\sigma = 2$:



[13]We'll have much more to say about parameters of a distribution. In fact, much of our statistical study will revolve around the parameters.

In this case the **normal** random variable X can assume any real value. Furthermore, it is true—but not entirely trivial to show by elementary means—that

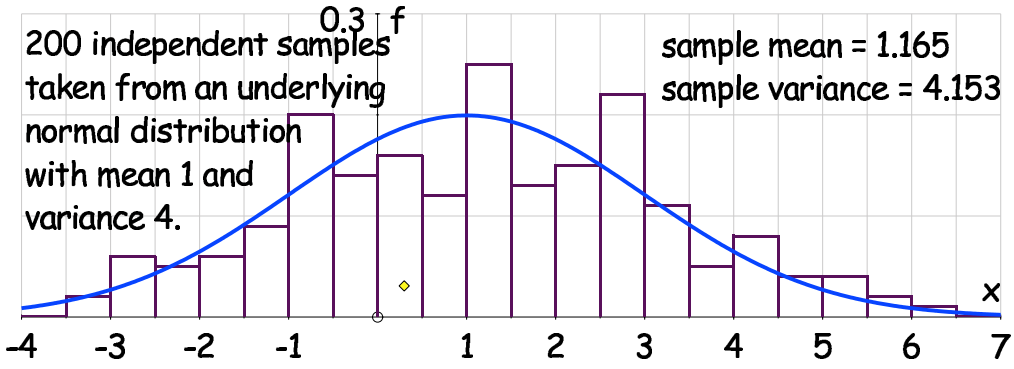
$$\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\,dx = 1,$$

which by now we should recognize as being a basic property of any density function.

EXAMPLE. Our graphing calculators allow for sampling from normal distributions, via the $\mathsf{randNorm}(\mu, \sigma, n)$, where n is the number of independent samples taken. The calculator operation

$$\mathsf{randNorm}(1, 2, 200) \to L_1$$

amounts to selecting 200 samples from a normally-distributed population having $\mu = 1$ and $\sigma = 2$. The same can be done in Autograph; the results of such a sample are indicated below:



6.2.2 Densities and simulations

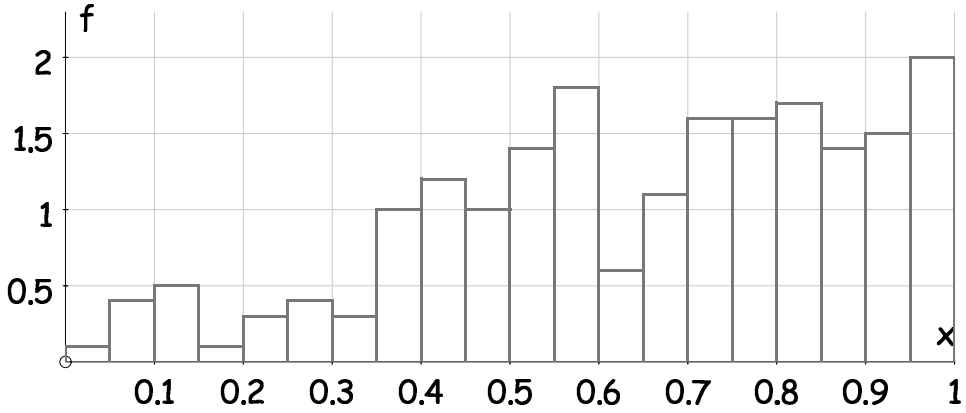
In the above we had quite a bit to say about density functions and about sampling from the uniform and normal distributions. We'll continue this theme here.

Let's begin with the following question. Suppose that X is the uniform random number generator on our TI calculators: $X = \mathsf{rand}$. Let's define a new random variable by setting $Y = \sqrt{X} = \sqrt{\mathsf{rand}}$. What

does the underlying density curve look like? Is it still uniform as in the case of X? Probably not, but let's take a closer look. Before getting too heavily into the mathematics let's start by collecting 200 samples of $\sqrt{\mathsf{rand}}$ and drawing a histogram. This will give us a general idea of what the underlying density curve might look like. Collecting the samples is easy:

$$\sqrt{\mathsf{rand}(200)} \to L_1$$

puts 200 samples from this distribution into the TI list variable L_1. Likewise, this sampling is easily done using more advanced softwares as Autograph or Minitab. Below is an Autograph-produced histogram of these 200 samples.



We would suspect on the basis of this histogram that the underlying density curve is not uniform but has considerable skew. Intuitively, we could have seen this simply by noting that for any real number x satisfying $0 < x < 1$ then $x < \sqrt{x}$; this is what creates the histogram's to skew to the left.

Can we make this more precise? Yes, and it's not too difficult. If we set $X = \mathsf{rand},\ Y = \sqrt{\mathsf{rand}}$, we have that for any value of $t,\ 0 \le t \le 1$,

$$P(Y \le t) = P(\sqrt{X} \le t) = P(X \le t^2) = t^2$$

(since X is uniformly distributed on $[0, 1]$.) In other words, if f_Y is the density function for Y, then it follows that

$$\int_0^t f_Y(x)\,dx = P(Y \le t) = t^2;$$

differentiating both sides with respect to t and applying the Fundamental Theorem of Calculus, we get

$$\boxed{f_Y(t) \;=\; 2t}\,.$$

Of course, this is the density function given a few pages earlier. In summary, the square root of the uniform random-number generator has a linear density function given by $f(t) = 2t$.

Assume, more generally, that we wish to transform data from the random-number generator $X = \mathsf{rand}$ so as to produce a new random variable Y having a given distribution function f_Y. If we denote this transformation by $Y = g(X)$, we have

$$\int_{-\infty}^{t} f_Y(x)dx = P(Y \leq t) = P(g(X) \leq t) = P(X \leq g^{-1}(t)) = g^{-1}(t),$$

which determines g^{-1} and hence the transformation g.

EXERCISES

1. If X is the random variable having the triangular density curve depicted on page 349, compute

   (a) $P(X \leq 1/3)$

   (b) $P(X \geq 2/3)$

   (c) $P(1/3 \leq X \leq 2/3)$

   (d) $P(X > .5)$

2. Suppose that you perform an experiment where you invoke the random-number generator twice and let Z be the sum of the two random numbers.

   (a) Compute $P(.5 \leq Z \leq 1.65)$ theoretically.

(b) Estimate $P(.5 \leq Z \leq 1.65)$ through a simulation, using the TI code as follows. (I would suggest taking $N \geq 100$ trials in this simulation.)

PROGRAM: SIMUL1
:$0 \to C$
:INPUT "N: ", N
:For(I,1,N)
:rand + rand $\to Z$
:$C + (.5 \leq Z)(Z \leq 1.65) \to C$
:END
:DISP "PROB: ", C/N
:STOP

The quantity C/N is the estimated probability!

(c) Construct a histogram for 100 observations of the random variable Z. Try the following code (using, say, $N = 100$):

PROGRAM: SIMUL2
:INPUT "N: ", N
:$\{0\} \to L_1$
:For(I,1,N)
:rand + rand $\to L_1(I)$
:END

Once you've done the above, you then use your graphing calculator to graph the histogram of the list variable L_1. (You'll need to decide on sensible window settings.)

3. Let B and C be uniform random variables on the interval $[-1, 1]$. (Therefore B and C are independent occurrences of $2\,\mathsf{rand} - 1$.) Compute

(a) the probability that the quadratic $x^2 + Bx + C = 0$ has two distinct real roots;

(b) the probability that the quadratic $x^2 + Bx + C = 0$ has a single multiple root;

(c) the probability that the quadratic $x^2 + Bx + C = 0$ has two real roots.

4. Do the same as above where you take 200 samples from a normal distribution having $\mu = 0$ and $\sigma = 1$. Create a histogram and draw the corresponding normal density curve simultaneously on your TI calculators.[14]

5. Define the random variable by setting $Z = \mathsf{rand}^2$.

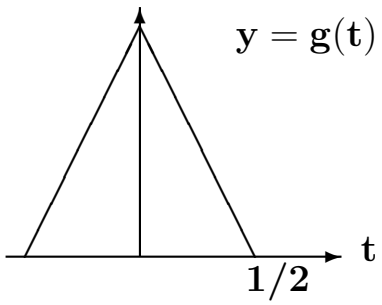
   (a) Determine the density function for Z. Before you start, why do you expect the density curve to be skewed to the **right**?

   (b) Collect 200 samples of Z and draw the corresponding histogram.

   (c) How well does your histogram in (b) conform with the density function you derived in (a)?

6. Consider the density function g defined by setting

$$g(t) = \begin{cases} 4t + 2 & \text{if } -\frac{1}{2} \le t \le 0 \\ -4t + 2 & \text{if } 0 \le t \le \frac{1}{2} \end{cases}$$



   (a) Show that $Y = \frac{1}{2}X_1 + \frac{1}{2}X_2 - \frac{1}{2}$, where $X_1 = \mathsf{rand}$, $X_2 = \mathsf{rand}$, X_1 and X_2 are independent. (Hint: just draw a picture in the X_1X_2-plane to compute $P(a \le Y \le b)$.)

[14] In drawing your histogram, you will need to make note of the widths of the histogram bars in order to get a good match between the histogram and the normal density curve. For example, if you use histogram bars each of width .5, then with 200 samples the total area under the histogram will be $.5 \times 200 = 100$. Therefore, in superimposing the normal density curve you'll need to multiply by 100 to get total area of 100. (Use $Y_1 = 100 * \text{normalpdf}(X, 0, 1)$.)

(b) Write a TI program to generate 200 samples of Y.

(c) Graph the histogram generated in (b) simultaneously with the density curve for Y.

7. Let $Z = \mathsf{rand}^2$ as in Exercise 5. Show that the density function for Z is given by

$$f(x) \;=\; \begin{cases} \frac{1}{2\sqrt{x}} & \text{if } 0 < x \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

8. We have seen that the density function for the normally-distributed random variable X having mean 0 and standard deviation 1 is

$$f(x) \;=\; \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

The **χ^2 random variable** with **one degree of freedom** is the random variable X^2 (whence the notation!). Using the ideas developed above, show that the density function for X^2 is given by

$$g(x) \;=\; \frac{1}{\sqrt{2\pi}}\, x^{-1/2} e^{-x/2}.$$

(More generally, the **χ^2 distribution with n degrees of freedom** is the distribution of the sum of n independent χ^2 random variables with one degree of freedom.)[15]

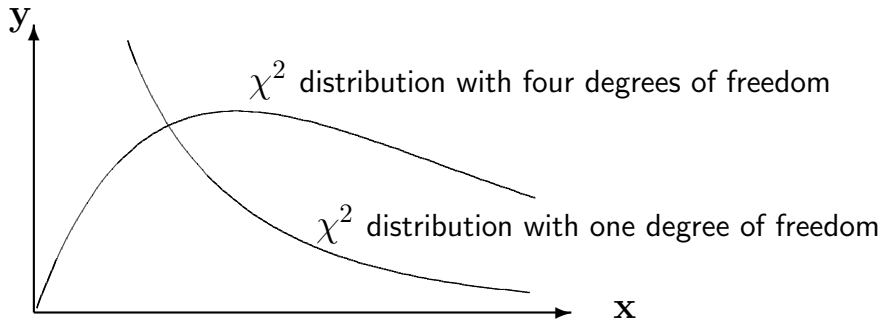
Below are the graphs of χ^2 with one and with four degrees of freedom.

[15]The density function for the χ^2 distribution with n degrees of freedom turns out to be

$$g(x) \;=\; \frac{x^{n/2-1} e^{-x/2}}{2^{n/2}\Gamma(\frac{n}{2})},$$

where $\Gamma(\frac{n}{2}) = (\frac{n}{2} - 1)!$ is n is even. If $n = 2k + 1$ is odd, then

$$\Gamma\left(\frac{n}{2}\right) = \Gamma\left(k + \frac{1}{2}\right) = \left(k - \frac{1}{2}\right)\left(k - \frac{3}{2}\right) \cdots \frac{3}{2} \cdot \frac{1}{2} \cdot \sqrt{\pi}.$$



9. (The **Maxwell-Boltzmann density function**) The Maxwell-Boltzmann distribution comes from a random variable of the form

$$Y = \sqrt{X_1^2 + X_2^2 + X_3^2},$$

where X_1, X_2, X_3 are independent normal random variables with mean 0 and variance a^2. Given that the density of the χ^2-random variable with three degrees of freedom, show that the density of Y is given by

$$f_Y(t) = \sqrt{\frac{2}{\pi}} \left(\frac{t^2 e^{-t^2/(2a^2)}}{a^3} \right).$$

This distribution is that of the speeds of individual molecules in ideal gases.[16]

10. Using integration by parts, show that $E(\chi^2) = 1$ and that $\text{Var}(\chi^2) = 2$ where χ^2 has one degree of freedom. Conclude that the expected value of the χ^2 random variable with n degrees of freedom is n and the variance is $2n$. We'll have much more to say about the χ^2 distribution and its application in Section 6.6.

11. Here's a lovely exercise.[17] Circles of radius 1 are constructed in the plane so that one has center $(2\,\mathsf{rand}, 0)$ and the other has center $(2\,\mathsf{rand}, 1)$. Compute the probability that these randomly-drawn

[16]It turns out that the the standard distribution a is given by $a = \dfrac{kT}{m}$, where T is the temperature (in degrees Kelvin), m is the molecular mass, and k is the Boltzmann constant

$$k = 1.3806603 \times 10^{-23} \text{m}^2 \cdot \text{kg/s}^2 \cdot \text{K}.$$

[17]This is essentially problem #21 on the 2008 AMC (American Mathematics Competitions) contest 12 B.

circles intersect. (Hint: let X_1 and X_2 be independent instances of rand and note that it suffices to compute $P(4(X_1 - X_2)^2 + 1 \leq 4)$.)

12. Let X be a random variable with density function $f(x)$, non-zero on the interval $a \leq x \leq b$. Compute the density function of $cX + d$, where c and d are constants with $a > 0$, in terms of f.

13. Define the random variable $Z =$ rand, and so $0 \leq Z \leq 1$.

    (a) Determine the density function of $Y = 10^Z$.

    (b) Notice that the random variable Y satisfies $1 \leq Y \leq 10$. Show that the probability that a random sample of Y has first digit 1 (i.e., satisfies $1 \leq Y < 2$) is $\log_{10} 2 \approx 30.1\%$. (This result is a simplified version of the so-called **Benford's Law**.)

    (c) Data arising from "natural" sources often satisfy the property that their logarithms are roughly uniformly distributed. One statement of Benford's Law is that—contrary to intuition—roughly 30% of the data will have first digit 1. We formalize this as follows. Suppose that we a random variable $1 \leq Y \leq 10^n$, where n is any positive integer, and assume that $Z = \log_{10} Y$ is uniformly distributed. Show that the probability that a random sample of Y has digit d, $1 \leq d \leq 0$ is $\log_{10}\left(1 + \frac{1}{d}\right)$.

6.2.3 The exponential distribution

The **exponential random variable** is best thought of as a continuous analog of the geometric random variable. This rather glib statement requires a bit of explanation.

Recall that if X is a geometric random variable with probability p, then $P(X = k)$ is the probability that our process (or game) will terminate after k independent trials. An immediate consequence of this fact is that the conditional probabilities $P(X = k + 1 \,|\, X \geq k) = p$, and hence is independent of k. In other words, if we have managed to survive k trials, then the probability of dying on the $k + 1$-st trial depends only on the parameter p and not on k. Similarly, we see that $P(X = k + \tau \,|\, X \geq k)$ will depend only on p and the integer τ, but not

on k. This says that during the game we don't "age"; our probability of dying at the next stage doesn't increase with age (k).

We now turn this process into a continuous process, where we can die at any time $t \geq 0$ and not just at integer times. We want the process to enjoy essentially the same condition as the geometric, namely that if X now denotes the present random variable, then the conditional probability $P(X = t + \tau \,|\, X \geq t)$ should depend on τ but not on t. In analogy with the above, this conditional probability represents the probability of living to time $t + \tau$, given that we have already lived t units of time.

We let f represent the density function of X; the above requirement says that

$$\frac{\int_t^{t+\tau} f(s)\,ds}{\int_t^{\infty} f(s)\,ds} \;=\; \text{function of } \tau \text{ alone .} \qquad (*)$$

We denote by F an antiderivative of f satisfying $F(\infty) = 0$.[18] Therefore,

$$1 \;=\; \int_0^{\infty} f(s)\,ds \;=\; F(s)\Big|_0^{\infty} = -F(0),$$

and so $F(0) = -1$.

Next, we can write (*) in the form

$$\frac{F(t+\tau) - F(t)}{-F(t)} \;=\; g(\tau),$$

for some function g. This implies that the quotient $\dfrac{F(t+\tau)}{F(t)}$ doesn't depend on t. Therefore, the derivative with respect to t of this quotient is 0:

$$\frac{F'(t+\tau)F(t) - F(t+\tau)F'(t)}{F(t)^2} = 0,$$

forcing

[18] Since $\int_0^{\infty} f(s)\,ds = 1$, we see that F cannot be unbounded at ∞.

$$F'(t+\tau)F(t) = F(t+\tau)F'(t).$$

But this can be written as

$$\frac{d}{dt}\ln F(t+\tau) = \frac{d}{dt}\ln F(t),$$

forcing

$$F(t+\tau) = -F(t)F(\tau),$$

for all t and τ. Finally, if we differentiate both sides of the above with respect to t and then set $t = 0$, we arrive at

$$F'(\tau) = -F'(0)F(\tau),$$

which, after setting $\lambda = F'(0)$, easily implies that $F(t) = -e^{-\lambda t}$ for all $t \geq 0$. Since f is the derivative of F, we conclude finally, that the density function of the exponential distribution must have the form

$$f(t) \;=\; \lambda e^{-\lambda t}, \quad t \geq 0.$$

Very easy integrations show that

$$E(X) \;=\; \frac{1}{\lambda} \;\text{ and }\; \mathrm{Var}(X) \;=\; \frac{1}{\lambda^2}.$$

The exponential distribution is often used in reliability engineering to describe units having a constant failure rate (i.e., age independent). Other applications include

- modeling the time to failure of an item (like a light bulb; see Exercise 2, below). The parameter λ is often called the **failure rate**;
- modeling the time to the next telephone call;
- distance between roadkill on a given highway;

- number of days between accidents at a given intersection.

EXERCISES

1. Prove the assertions made above concerning the exponential random variable X with density $f(t) = \lambda e^{-\lambda t}$, $t \geq 0$, viz., that $E(X) = 1/\lambda$ and that $\text{Var}(X) = 1/\lambda^2$.

2. Suppose that the useful life T of a light bulb produced by a particular company is given by the density function $f(t) = 0.01e^{-0.01t}$, where t is measured in hours. Therefore, the probability that this light bulb fails somewhere between times t_1 and t_2 is given by the integral $P(t_1 \leq T \leq t_2) = \int_{t_1}^{t_2} f(t)\, dt$.

   (a) The probability that the bulb will not burn out before t hours is a function of t and is often referred to as the **reliability** of the bulb.

   (b) For which value of t is the reliability of the bulb equal to $1/2$. Interpret this value of t.

3. Suppose that your small company had a single secretary and that she determined that one a given day, the measured time elapsed between 30 consecutive received telephone calls was (in minutes)

   6.8, 0.63, 5.3, 3.8, 3.5, 7.2, 16.0. 5.5, 7.2, 1.1, 1.4, 1.8, 0.28, 1.2, 1.6, 5.4, 5.4, 3.1, 1.3, 3.7, 7.5, 3.0, 0.03, 0.64, 1.5, 6.9, 0.01, 4.7, 1.4, 5.0.

   Assuming that this particular day was a typical day, use these data to estimate the mean wait time between phone calls. Assuming that the incoming phone calls roughly follow an exponential distribution with your estimated mean, compute the probability that after a given call your secretary will receive another call within two minutes.

4. Under the same assumptions as in the above exercise, roughly how long will it take for the tenth call during the day will be taken by your secretary?

5. You have determined that along a stretch of highway, you see on average one dead animal on the road every 2.1 km. Assuming an exponential distribution with this mean, what is the probability that after seeing the last road kill you will drive 8 km before seeing the next one.

6. (Harder question) Assume, as in the above problem that you see on average one dead animal every 2.1 km along the above-mentioned highway. What is the probability that you will drive at least 10 km before seeing the next two dead animals? (Hint: Let X_1 be the distance required to spot the first roadkill, and let X_2 be the distance required to spot the second roadkill. You're trying to compute $P(X_1 + X_2 \geq 10)$; try looking ahead to page 370.)

7. We can simulate the exponential distribution on a TI-series calculator, as follows. We wish to determine the transforming function g such that when applied to rand results in the exponential random variable Y with density function $f_Y(t) = \lambda e^{-\lambda t}$. Next, from page 353 we see that, in fact

$$\int_0^t f_Y(x)\,dx = \int_0^t \lambda e^{-\lambda x}\,dx = g^{-1}(t).$$

That is to say, $1 - e^{-\lambda t} = g^{-1}(t)$.

   (a) Show that this gives the transforming function $g(x) = \dfrac{-1}{\lambda}\ln(1-x)$.

   (b) On your TI calculators, extract 100 random samples of the exponential distribution with $\lambda = .5$ (so $\mu = 2$) via the command

$$\frac{-1}{.5}\ln(1 - \mathsf{rand}(100)) \to L_1.$$

   This will place 100 samples into the list variable L_1.

   (c) Draw a histogram of these 100 samples—does this look right?

8. (This is an extended exercise.) Continuing on the theme in Exercise 7, we can similarly use the TI calculator to generate samples

of a geometric random variable. Just as we were able above to transform rand into an exponential random variable, we shall (approximately) transform the TI random integer variable "randInt" into a geometric random variable.

First of all, the random integer generator has three inputs and has the form randInt$(n_{\min}, n_{\max}, N)$. The output consists of a sequence of N randomly and uniformly distributed integers between $n_{\min}$ and $n_{\max}$. We shall, for convenience, take $n_{\min} = 1$ and set $n = n_{\max}$. We let Y be the geometric random variable (with parameter p), and let X be a randomly-generated uniformly distributed integer $1 \leq X \leq n$. The goal is to find a function g such that $Y = g(X)$. This will allow us to use the TI calculator to generate samples of a geometric random variable (and therefore of a negative binomial random variable).

Note first that

$$P(Y \leq k) = p+p(1-p)+p(1-p)^2+\cdots+p(1-p)^{k-1} = 1-(1-p)^k,\ k \geq 0$$

and that

$$P(X \leq h) \ = \ \frac{h}{n}, \ \ 1 \leq h \leq n.$$

At this point we see a potential problem in transforming from the uniform variable to the geometric: the geometric random variable has an infinite number of possible outcomes (with decreasing probabilities) and the uniform random variable is an integer between 1 and n. Therefore, we would hope not to lose too much information by allowing n to be reasonably large ($n \approx 25$ seems pretty good). At any rate, we proceed in analogy with the analysis on page 353: assuming that $Y = g(X)$ we have

$$\begin{aligned} 1-(1-p)^k &= P(Y \le k) \\ &= P(g(X) \le k) \\ &= P(X \le g^{-1}(k)) \\ &= \frac{g^{-1}(k)}{n} \end{aligned}$$

Solving for g we get

$$g(h) = \frac{\ln(1-\frac{h}{n})}{\ln(1-p)}.$$

However, we immediately see that if $h = n = n_{\max}$ we see that $g(h)$ is undefined (and the TI will generate an error whenever $h = n$). This makes mathematical sense as there is no value of k for which $P(Y \le k) = 1$. One remedy is to let the value of n in the above expression for g be *slightly larger than* $n_{\max}$. Of course, having done this, the transformed values will no longer be integers, so we'll need to round to the nearest integer. On the TI calculator the value $\text{int}(x + .5)$ will have the effect of rounding to the nearest integer.

NOW DO THIS: Generate 100 samples of the geometric random variable with parameter $p = .25$ using the command

$$\frac{\ln\left(1-\frac{\text{randInt}(1,30,100)}{30.01}\right)}{\ln(1-.25)} \to L_1$$

followed by the command

$$\text{int}(L_1 + .5) \to L_1.$$

This will store 100 randomly-generated integer samples in the list variable L_1. You should check to see if they appear to follow the geometric distribution with parameter $p = .25$. (Start by comparing the mean of your data with the theoretical mean of the geometric random variable!)

9. Let X be an exponential random variable with failure rate λ, and let $Y = X^{1/\alpha},\ \alpha > 0$. Using the idea developed on page 353, compute the density function for Y. This gives the so-called **Weibull distribution**.

6.3 Parameters and Statistics

Suppose that we have a continuous random variable X having density function f_X. Associated with this random variable are a few **parameters**, the **mean** (and also the **median** and the **mode**) and the **variance** of X. In analogy with discrete random variables they are defined as follows.

Mean of X. We set

$$E(X) \ = \ \mu_X \ = \ \int_{-\infty}^{\infty} x f_X(x)\, dx.$$

Median of X. This is just the half-way point of the distribution, that is, if m is the median, we have $P(X \leq m) = \frac{1}{2} = P(X \geq m)$. In terms of the density function, this is just the value m for which

$$\int_{-\infty}^{m} f_X(x)\, dx \ = \ \frac{1}{2}.$$

Mode of X. This is just the value of x at which the density function assumes its maximum. (Note, then, that the mode might not be unique: a distribution might be "bimodal" or even "multimodal.")

The **mean**, **median**, and **mode** measure "central tendency."

Variance of X. We set

$$\mathrm{Var}(X) \ = \ \sigma_X^2 \ = \ E((X - \mu_X)^2).$$

As we shall see below,

$$\mathrm{Var}(X) \;=\; \int_{-\infty}^{\infty} (x-\mu_X)^2 f_X(x)\,dx.$$

(though most texts gloss over this point).

The positive square root σ_X of $\mathrm{Var}(X)$ is called the **standard deviation** of X.

6.3.1 Some theory

If we adapt the arguments beginning on page 319 to continuous random variables, we can give a heuristic argument that the expectation of the sum of two continuous random variables is the sum of the expectations. The basic idea is that there is a **joint density** function f_{XY} which gives probabilities such as

$$P(a \le X \le b \text{ and } c \le Y \le d) = \int_a^b \int_c^d f_{XY}(x,y)\,dxdy.$$

These can be represented in terms of conditional probabilities in the usual way: $f_{XY}(x,y) = f_X(x|y)f_Y(y)$; furthermore, one has

$$f_X(x) = \int_{-\infty}^{\infty} f_X(x|y)\,dy.$$

Accepting all of this stuff, one proceeds exactly as on pages 319–320:

$$\begin{aligned}
\mu_{X+Y} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x+y) f_{XY}(x,y)\,dx\,dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{XY}(x,y)\,dx\,dx + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f_{XY}(x,y)\,dx\,dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_X(x|y) f_Y(y)\,dy\,dx + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f_Y(y|x) f_X(x)\,dx\,dy \\
&= \int_{-\infty}^{\infty} x f_X(x)\,dx + \int_{-\infty}^{\infty} y f_Y(y)\,dy \\
&= \mu_X + \mu_Y.
\end{aligned}$$

A similar argument, together with mathematical induction can be used to show that

$$\boxed{E(X_1 + X_2 + \cdots + X_k) = E(X_1) + E(X_2) + \cdots + E(X_k)}.$$

If the random variables X and Y are independent, then we may write the density function $f_{XY}(x, y)$ as a product: $f_{XY}(x, y) = f_X(x)f_Y(y)$, from which it follows immediately that

$$E(XY) \;=\; E(X)E(Y), \quad \text{where } X \text{ and } Y \text{ are independent.}$$

In particular, this shows the following very important result. Assume that we are to take n independent samples from a given population having mean μ. If $\overline{X}$ denotes the average of these samples, then $\overline{X}$ is a itself a random variable and

$$\overline{X} \;=\; \frac{X_1 + X_2 + \cdots + X_n}{n},$$

where X_1, X_2, $\ldots$, X_n are independent random variables from this population. We have, therefore, that

$$E(\overline{X}) = \frac{E(X_1) + E(X_2) + \cdots + E(X_n)}{n} \;=\; \mu.$$

We now turn our attention to variance. However, a couple of preliminary observations are in order. First of all, let X be a continuous random variable, let a be a real constant, and set $Y = X + a$. We wish first to compare the density functions f_Y and f_X. Perhaps it's already obvious that $f_Y(x) = f_X(x - a)$, but a formal proof might be instructive. We have

$$\int_{-\infty}^{t} f_Y(x)\,dx = P(Y \le t) = P(X+a \le t) = P(X \le t-a) = \int_{-\infty}^{t-a} f_X(x)\,dx.$$

But a simple change of variable shows that

$$\int_{-\infty}^{t-a} f_X(x)\,dx \;=\; \int_{-\infty}^{t} f_X(x-a)\,dx.$$

In other words, for all real numbers t, we have

$$\int_{-\infty}^{t} f_Y(x)\,dx \;=\; \int_{-\infty}^{t} f_X(x-a)\,dx.$$

This implies (e.g., by the Fundamental Theorem of Calculus) that

$$f_{X+a}(x) \;=\; f_X(x-a) \qquad (*)$$

for all $x \in \mathcal{R}$.

Next, we would like to compute the density function for the random variable $Y = X^2$ in terms of that of X. To do this, note that

$$\int_0^t f_{X^2}(x)\,dx \;=\; P(X^2 < t) \;=\; P(-\sqrt{t} < X < \sqrt{t}) \;=\; \int_{-\sqrt{t}}^{\sqrt{t}} f(x)\,dx.$$

An application of the Fundamental Theorem of Calculus gives

$$f_{X^2}(x) \;=\; \frac{1}{2\sqrt{x}}\left(f_X(\sqrt{x}) - f_X(-\sqrt{x})\right). \qquad (**)$$

Using equations (*) and (**), we can compute the variance of the continuous random variable X having mean μ, as follows. We have

$$\begin{aligned}
\mathrm{Var}(X) &= E((X-\mu)^2) \\
&= \int_0^\infty x f_{(X-\mu)^2}(x)\,dx \\
&\overset{\text{by (**)}}{=} \frac{1}{2}\int_0^\infty \sqrt{x} f_{X-\mu}(\sqrt{x})\,dx \;-\; \frac{1}{2}\int_0^\infty \sqrt{x}\, f_{X-\mu}(-\sqrt{x})\,dx \\
&\overset{(u=\sqrt{x})}{=} \int_0^\infty u^2 f_{X-\mu}(u)\,du + \int_{-\infty}^0 u^2 f_{X-\mu}(u)\,du = \int_{-\infty}^\infty u^2 f_{X-\mu}(u)\,du \\
&\overset{\text{by (*)}}{=} \int_{-\infty}^\infty u^2 f_X(u+\mu)\,du \\
&= \int_{-\infty}^\infty (u-\mu)^2 f_X(u)\,du, \\
&= \int_{-\infty}^\infty (x-\mu)^2 f_X(x)\,dx.
\end{aligned}$$

This proves the assertion made on page 365. Next, we have

$$\begin{aligned} E(X^2) &= \int_0^\infty x f_{X^2}(x)\,dx \\ &= \frac{1}{2}\int_0^\infty \sqrt{x} f_X(x)\,dx \;-\; \frac{1}{2}\int_0^\infty \sqrt{x}\, f_X(-\sqrt{x})\,dx \\ &= \int_{-\infty}^\infty u^2 f_X(u)\,dx \\ &= \int_{-\infty}^\infty x^2 f_X(x)\,dx \end{aligned}$$

Finally,

$$\begin{aligned} \mathrm{Var}(X) &= \int_{-\infty}^\infty (x-\mu)^2 f_X(x)\,dx \\ &= \int_{-\infty}^\infty (x^2 - 2x\mu + \mu^2) f_X(x)\,dx \\ &= \int_{-\infty}^\infty x^2 f_X(x)\,dx - 2\mu\int_{-\infty}^\infty x f_X(x)\,dx + \mu^2\int_{-\infty}^\infty f_X(x)\,dx \\ &= E(X^2) - \mu^2, \end{aligned}$$

exactly as for discrete random variables (page 321).

We need one final theoretical result concerning variance. Assume that we take two independent measurements X and Y from a given population both having mean μ. What is the variance of $X+Y$? This will require the two results sketched on page 366, namely

(i) that $E(X+Y) = E(X) + E(Y)$, *whether or not* X and Y, and

(ii) that if X and Y are independent, then $E(XY) = E(X)E(Y)$.

Using these two facts, we can proceed as follows:

$$\begin{aligned} \mathrm{Var}(X+Y) &= E((X+Y-2\mu)^2) \\ &= E(X^2 + Y^2 + 2XY - 4\mu X - 4\mu Y + 4\mu^2) \\ &= E(X^2) + E(Y^2) + 2\mu^2 - 4\mu^2 - 4\mu^2 + 4\mu^2 \\ &= E(X^2) - \mu^2 + E(Y^2) - \mu^2 \;=\; \mathrm{Var}(X) + \mathrm{Var}(Y). \end{aligned}$$

Convolution and the sum of independent random variables. Assume that X and Y are independent random variables with density functions f_X and f_Y, respectively. We shall determine the distribution of $X + Y$ in terms of f_X andf_Y.

To do this we observe that

$$\begin{aligned} f_{X+Y}(t) &= \frac{d}{dt}P(X+Y \leq t) \\ &= \frac{d}{dt}P(Y \leq t-X) \\ &= \frac{d}{dt}\int_{-\infty}^{\infty}\int_{-\infty}^{t-x} f_X(x)f_Y(y)\,dy\,dx \\ &= \int_{-\infty}^{\infty} f_X(x)f_Y(t-x)\,dx. \end{aligned}$$

The last expression above is called the **convolution** of the density functions.[19] We write this more simply as

$$f_{X+Y}(t) = f_X * f_Y(t),$$

where for any real-valued[20] functions f and g, the convolution is defined by setting

$$f * g(t) = \int_{-\infty}^{\infty} f(x)g(t-x)dx.$$

From the above we can easily compute the distribution of the difference $X - Y$ of the independent random variables X and Y. Note first that the distribution of $-Y$ is clearly the function $f_{-Y}(t) = f_Y(-t),\ t \in \mathbb{R}$. This implies that the distribution of f_{X-Y} is given by

$$f_{X-Y}(t) = f_X * f_{-Y}(t) = \int_{-\infty}^{\infty} f_X(x)f_{-Y}(t-x)dx = \int_{-\infty}^{\infty} f_X(x)f_Y(x-t)dx.$$

[19] Of course, the notion of **convolution** was already introduced in Exercise 5 on page 261.

[20] Actually, there are additional hypotheses required to guarantee the existence of the convolution product.

Next, continuing to assume that X an Y are independent random variables, we proceed to compute $E(X+Y)$. We have

$$\begin{aligned} E(X+Y) &= \int_{-\infty}^{\infty} x(f_X * f_Y)(x)\,dx \\ &= \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f_X(t) f_Y(x-t)\,dt\,dx \\ &= \int_{-\infty}^{\infty} f_X(t) \int_{-\infty}^{\infty} x f_Y(x-t)\,dt\,dx \\ &= \int_{-\infty}^{\infty} f_X(t) \int_{-\infty}^{\infty} (t+x) f_Y(x)\,dx\,dt \\ &= \int_{-\infty}^{\infty} f_X(t)(t+E(Y))\,dt \\ &= E(X)+E(Y). \end{aligned}$$

EXERCISES

1. Compute the mean and the variance of the random variable rand. (Recall that rand has density function

$$f(x) = \begin{cases} 1 & \text{if } 0 \le x \le 1, \\ 0 & \text{otherwise.)} \end{cases}$$

2. Compute the mean and the variance of the random variable $\sqrt{\mathsf{rand}}$. (Recall that $\sqrt{\mathsf{rand}}$ has density function

$$f(x) = \begin{cases} 2x & \text{if } 0 \le x \le 1, \\ 0 & \text{otherwise.)} \end{cases}$$

3. Compute the mean and the variance of the random variable rand^2. (See Exercise 7 on page 356.)

4. Compute the mean and the variance of the random variable having density function given in Exercise 6 on page 355.

5. In Exercise 6 on page 362 you were asked essentially to investigate the distribution of $X_1 + X_2$ where X_1 and X_2 were independent exponential random variables, each with mean $\mu = 1/\lambda$. Given that the density function of each is $f(x) = \lambda e^{-\lambda x}$ and given that the sum has as density the convolution of f with itself (see page 370), compute this density.

6. In Exercise 8 on page 356 we showed that the density function for the χ^2 random variable with one degree of freedom is $f(x) = \frac{1}{\sqrt{2\pi}} x^{-1/2} e^{-x/2}$. Using the fact that the χ^2 with two degrees of freedom is the sum of independent χ^2 random variables with one degree of freedom, and given that the density function for the sum of independent random variables is the convolution of the two corresponding density functions, compute the density function for the χ^2 random variable with two degrees of freedom. (See the footnote on page 356.)

7. Let f be an **even** real-valued function such that $\int_{-\infty}^{\infty} f(x)dx$ exists. Show that $f * f$ is also an even real-valued function.

8. Consider the function defined by setting

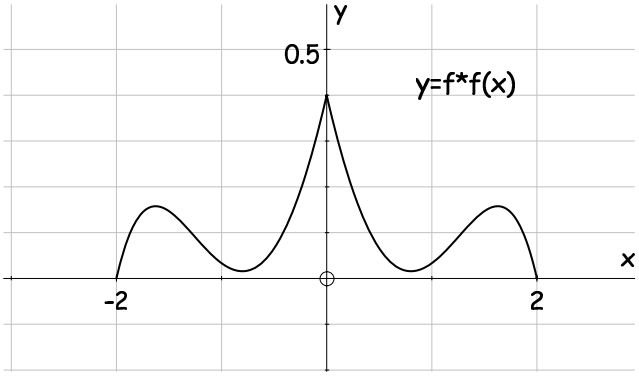
$$f(x) = \begin{cases} x^2 & \text{if } -1 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

   (a) Show that

$$f * f(x) = \begin{cases} \int_{-1}^{x+1} y^2(x-y)^2\,dy & \text{if } -1 \le x \le 0, \\ \int_{x-1}^{1} y^2(x-y)^2\,dy & \text{if } 0 \le x \le 1. \end{cases}$$

(b) Conclude that $f * f$ is not differentiable at $x = 0$.

(c) Show that the graph of $f * f$ is as depicted to the right.



(d) Also, compute $\int_{-\infty}^{\infty} f * f(x)dx$, and compare with $\int_{-\infty}^{\infty} f(x)dx$. Does this make sense? Can you formulate a general statement?

6.3.2 Statistics: sample mean and variance

In all of the above discussions, we have either been dealing with random variables whose distributions are known, and hence its mean and variance can (in principle) be computed, or we have been deriving theoretical aspects of the mean and variance of a random variable. While interesting and important, these are intellectual luxuries that usually don't present themselves in the real world. If, for example, I was charged with the analysis of the mean number of on-the-job injuries in a company in a given year, I would be tempted to model this with a Poisson distribution. Even if this were a good assumption, I probably wouldn't know the mean of this distribution. Arriving at a "good" estimate of the mean and determining whether the Poisson model is a "good" model are both statistical questions.

Estimation of a random variable's mean will be the main focus of the remainder of the present chapter, with a final section on the "goodness of fit" of a model.

We turn now to **statistics**. First of all, any particular values (outcomes) of a random variable or random variables are collectively known as **data**. A **statistic** is any function of the data. Two particularly important statistics are as follows. A **sample** (of size n) from a distribution with random variable X is a set of n independent measurements

$x_1, x_2, \ldots, x_n$ of this random variable. Associated with this sample are

The **sample mean**: this is defined by setting

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

The basic reason for considering the sample mean is the following. Suppose that we have taken the samples x_1, x_2, ..., x_n from a population whose mean is μ. Would we expect that $\overline{x} \approx \mu$? Fortunately, the answer is in agreement with our intuition; that we **really do** expect that the sample mean to approximate the theoretical (or population) mean. The reason, simply is that if we form the random variable

$$\overline{X} = \frac{X_1 + X_2 \cdots + X_n}{n},$$

then it is clear that $E(\overline{X}) = \mu$. (Indeed, we already noted this fact back on page 324.) That is to say, when we take n independent samples from a population, then we "expect" to get back the theoretical mean μ. Another way to state this is to say that $\overline{x}$ is an **unbiased estimate** of the population mean μ.

Next, notice that since X_1, X_2, $\cdots$, X_n are independent, we have that

$$\begin{aligned}
\mathrm{Var}(\overline{X}) &= \mathrm{Var}\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) \\
&= \frac{1}{n^2}\mathrm{Var}(X_1 + X_2 + \cdots + X_n) \\
&= \frac{1}{n^2}(\mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n)) \\
&= \frac{\sigma^2}{n}.
\end{aligned}$$

This shows why it's best to take "large" samples: the "sampling statistic" $\overline{X}$ has variance which tends to zero as the sample size tends to infinity.

The **sample variance**: this is defined by setting

$$s_x^2 \;=\; \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2.$$

The **sample standard deviation** $s_x = \sqrt{s_x^2}$.

If X_1, X_2, $\ldots$, X_n represent independent random variables having the same distribution, then setting

$$S_x^2 \;=\; \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$$

is a random variable. Once the sample has been taken, this random variable has taken on a value, $S_x = s_x$ and is, of course, no longer random. The relationship between S_x and s_x is the same as the relationship between $\overline{X}$ (random variable before collecting the sample) and $\overline{x}$ (the computed average of the sample).

You might wonder why we divide by $n-1$ rather than n, which perhaps seems more intuitive. The reason, ultimately, is that

$$E(S_x^2) \;=\; E\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2\right) = \sigma^2.$$

A sketch of a proof is given in the footnote.[21] (We remark in passing that many authors do define the sample variance as above, except that

[21] First of all, note that, by definition

$$E((X_i-\mu)^2) = \sigma^2,$$

from which it follows that

$$E\left(\sum_{i=1}^{n}(X_i-\mu)^2\right) = n\sigma^2.$$

Now watch this:

the sum is divided by n instead of $n-1$. While the resulting statistic is a biased estimate of the population variance, it does enjoy the property of being what's called a **maximum-likelihood estimate** of the population variance. A fuller treatment of this can be found in any reasonably advanced statistics textbook.)

Naturally, if we take a sample of size n from a population having mean μ and variance σ^2, we would expect that the sample mean and variance would at least approximate μ and σ^2, respectively. In practice, however, given a population we rarely know the population mean and variance; use use the statistics $\overline{x}$ and s_x^2 in order to estimate them (or to make hypotheses about them).

$$\begin{aligned}
\sum_{i=1}^{n}(X_i-\mu)^2 &= \sum_{i=1}^{n}\left[(X_1-\overline{X})+(\overline{X}-\mu)\right]^2 \\
&= \sum_{i=1}^{n}\left[(X_i-\overline{X})^2+2(X_i-\overline{X})(\overline{X}-\mu)+(\overline{X}-\mu)^2\right] \\
&= \sum_{i=1}^{n}(X_i-\overline{X})^2+n(\overline{X}-\mu)^2 \quad (\text{ since } \textstyle\sum(X_i-\overline{X})=0.)
\end{aligned}$$

Next, since $E(\overline{X})=\mu$, we have $E\left(n(\overline{X}-\mu)^2\right)=nE\left((\overline{X}-\mu)^2\right)=n\text{Var}(\overline{X})=\sigma^2$. Therefore, we take the expectation of the above random variables:

$$\begin{aligned}
n\sigma^2 &= E\left(\sum_{i=1}^{n}(X_i-\mu)^2\right) \\
&= E\left(\sum_{i=1}^{n}(X_i-\overline{X})^2+n(\overline{X}-\mu)^2\right) \\
&= E\left(\sum_{i=1}^{n}(X_i-\overline{X})^2\right)+E\left(n(\overline{X}-\mu)^2\right) \\
&= E\left(\sum_{i=1}^{n}(X_i-\overline{X})^2\right)+\sigma^2
\end{aligned}$$

from which we see that

$$E\left(\sum_{i=1}^{n}(X_i-\overline{X})^2\right) = (n-1)\sigma^2.$$

Therefore, we finally arrive at the desired conclusion:

$$E\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2\right) = \sigma^2.$$

6.3.3 The distribution of $\overline{X}$ and the Central Limit Theorem

The result of this section is key to all of sampling theory. As we might guess, one of the most important statistics we're apt to encounter is the mean $\overline{x}$ of n independent samples taken from some population. Underlying this is the random variable $\overline{X}$ with parameters

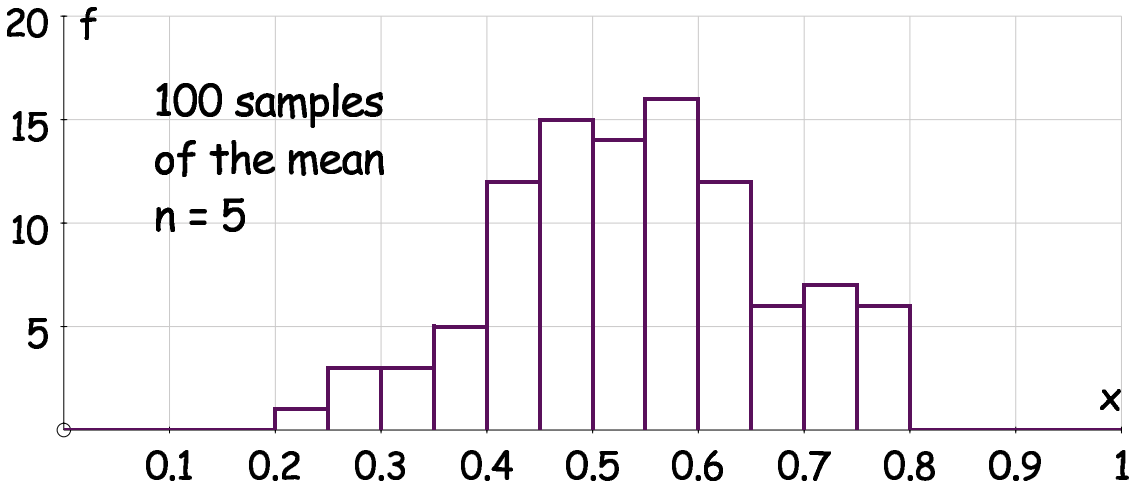
$$E(\overline{X}) \;=\; \mu, \quad \text{and} \quad \text{Var}(\overline{X}) \;=\; \frac{\sigma^2}{n}.$$

Let's start by getting our hands dirty.

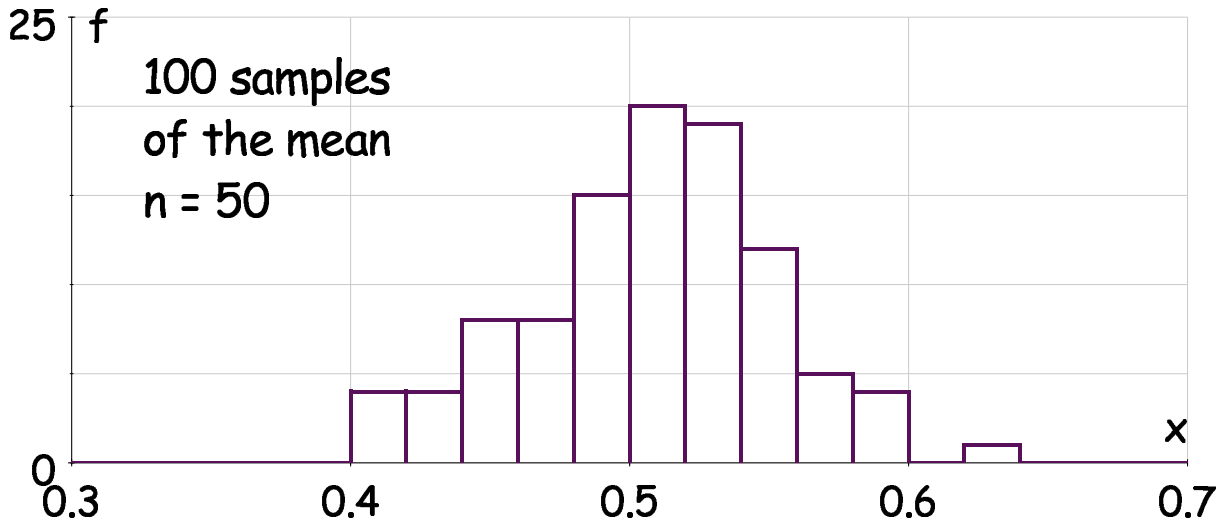
SIMULATION 1. Let's take 100 samples of the mean (where each mean is computed from 5 observations) from the uniform distribution having density function

$$f(x) \;=\; \begin{cases} 1 & \text{if } 0 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

We display the corresponding histogram:



SIMULATION 2. Here, let's take 100 samples of the mean (where each mean is computed from 50 observations) from the uniform distribution above. The resulting histogram is as below.



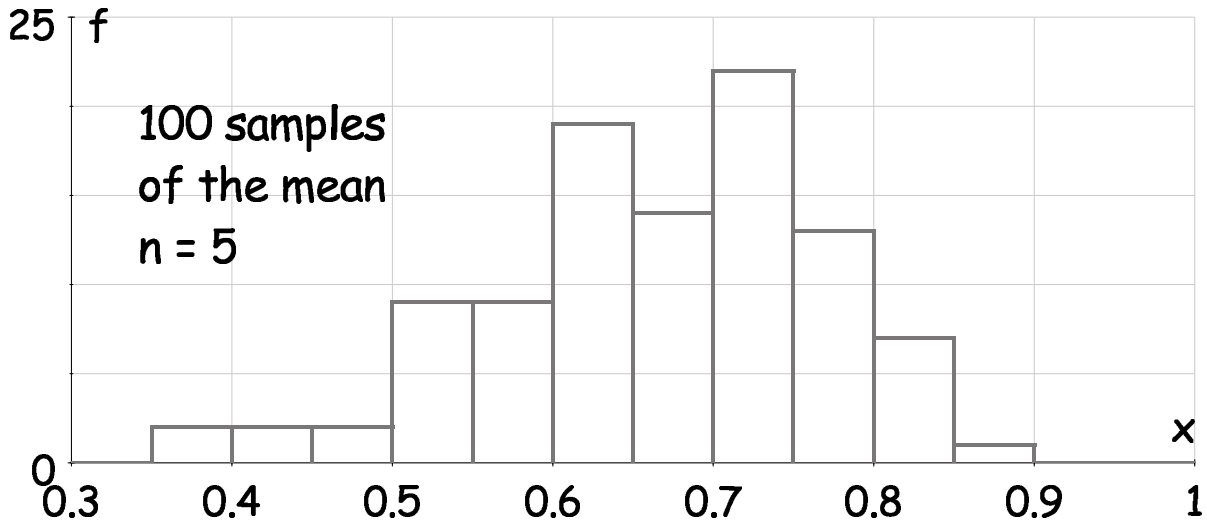
There are two important observations to make here. First of all, even though we haven't sampled from a normal distribution, the sample means appear to be somewhat normally distributed (more so in the $n = 50$ case). Next, notice that the range of the samples of the mean for $n = 50$ is much less than for $n = 5$. This is because the standard deviations for these two sample means are respectively $\frac{\sigma}{\sqrt{5}}$ and $\frac{\sigma}{\sqrt{50}}$, where σ is the standard deviation of the given uniform distribution.[22]

SIMULATION 3. Let's take 100 samples of the mean (where each mean is computed from 5 observations) from the distribution having density function

$$f(x) = \begin{cases} 2x & \text{if } 0 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

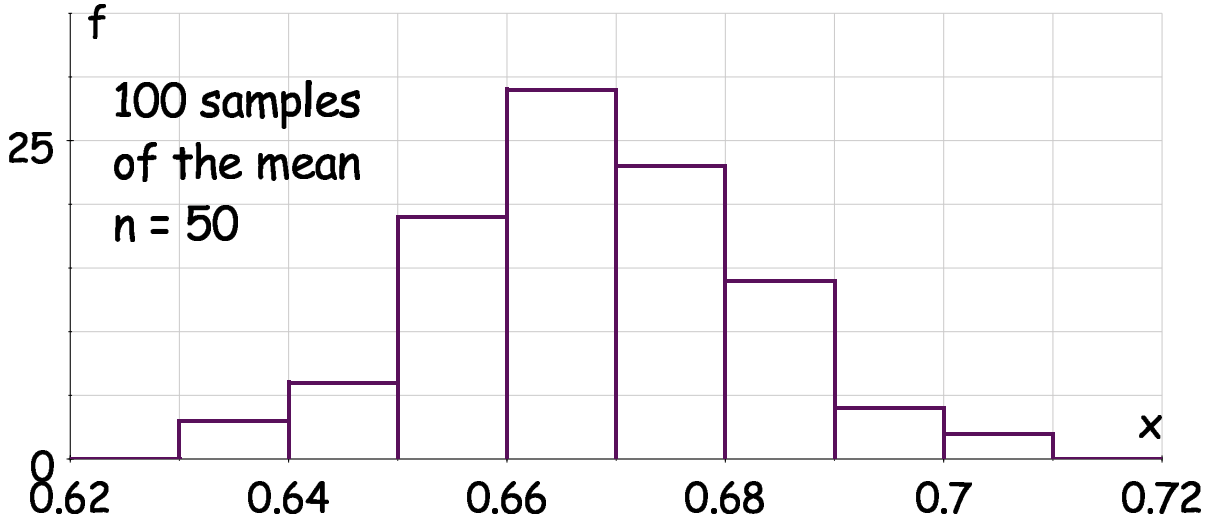
(Recall that this is the density function for $\sqrt{\mathsf{rand}}$.) We display the corresponding histogram.

[22]The variance of this distribution was to be computed in Exercise 1 on page 371; the result is $\sigma^2 = \frac{1}{12}$.



SIMULATION 4. Let's take 100 samples of the mean (where each mean is computed from 50 observations) from distribution having the same density function as above.

We display the corresponding histogram.



Again, note the tendency toward a normal distribution with a relatively narrow spread (small standard distribution).

The above is codified in the "Central Limit Theorem:"

Central Limit Theorem. *The sample mean $\overline{X}$ taken from n samples of a distribution with mean μ and variance σ^2 has a distribution which*

as $n \to \infty$ becomes arbitrarily close to the normal distribution with mean μ and variance $\dfrac{\sigma^2}{n}$.

Perhaps a better way to state the Central Limit Theorem is as follows. If Z is the normal random variable with mean 0 and variance 1, then for any real number z,

$$\lim_{n\to\infty} P\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < z\right) = P(Z < z).$$

6.4 Confidence Intervals for the Mean of a Population

A major role of statistics is to provide reasonable methods by which we can make inferences about the parameters of a population. This is important as we typically never know the parameters of a given population.[23] When giving an estimate of the mean of a population, one often gives an interval estimate, together with a **level of confidence**. So, for example, I might collect a sample from a population and measure that the mean $\overline{x}$ is 24.56. Reporting this estimate by itself is not terribly useful, as it is highly unlikely that this estimate coincides with the population mean. So the natural question is "how far off can this estimate be?" Again, not knowing the population mean, this question is impossible to answer. In practice what is done is to report a **confidence interval** together with a **confidence level**. Therefore, in continuing the above hypothetical example, I might report that

> "The 95% confidence interval for the mean is 24.56 ± 2.11."
>
> or that
>
> "The mean falls within 24.56 ± 2.11 with 95% confidence."

[23] In fact we almost never even know the population's underlying distribution. However, thanks to the Central Limit Theorem, as long as we take large enough samples, we can be assured of being "pretty close" to a normal distribution.

A very common misconception is that the above two statements mean that the population mean lies within the above reported interval with probability 95%. However, this is meaningless: *either the population mean does or doesn't lie in the above interval*; there is no randomness associated with the interval reported! As we'll see, the randomness is associated with the process of arriving at the interval itself. If 100 statisticians go out and compute 95% confidence intervals for the mean, then roughly 95 of the computed confidence intervals will actually contain the true population mean. Unfortunately, we won't know which ones actually contain the true mean!

6.4.1 Confidence intervals for the mean; known population variance

While it is highly unreasonable to assume that we would know the variance of a population but not know the mean, the ensuing discussion will help to serve as a basis for more practical (and realistic) methods to follow. Therefore, we assume that we wish to estimate the mean μ of a population whose variance σ^2 is known. It follows then, that if $\overline{X}$ is the random variable representing the mean of n independently-selected samples, then

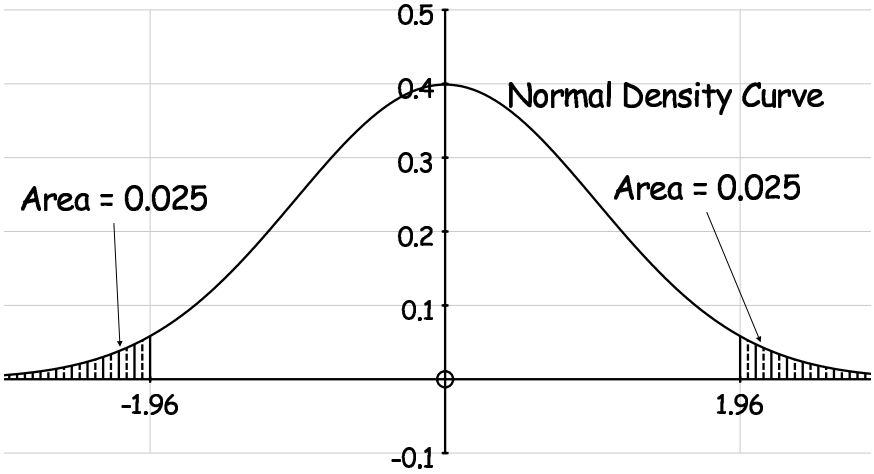
- the variance of $\overline{X}$ is $\dfrac{\sigma^2}{n}$, and
- (if n is "large")[24] the random variable $\overline{X}$ is approximately normally distributed.

In the ensuing discussion, we shall assume either that we are sampling from an (approximately) normal population or that n is relatively large. In either case, $\overline{X}$ will be (approximately) normally distributed. We have that

$$E(\overline{X}) = \mu, \quad \text{and} \quad \text{Var}(\overline{X}) = \frac{\sigma^2}{n};$$

[24] A typical benchmark is to use sample sizes of $n \geq 30$ in order for the normality assumption to be reasonable. On the other hand, if we know—or can assume—that we are sampling from a normal population in the first place, then $\overline{X}$ will be normally distributed for any n.

therefore the random variable $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ is normally distributed with mean 0 and standard deviation 1. The values $z \approx \pm 1.96$ are the values such that a normally-distributed random variable Z with mean 0 and variance 1 will satisfy $P(-1.96 \leq Z \leq 1.96) = 0.95$; see figure below



In other words, we have

$$P(-1.96 \leq \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leq 1.96) = 0.95.$$

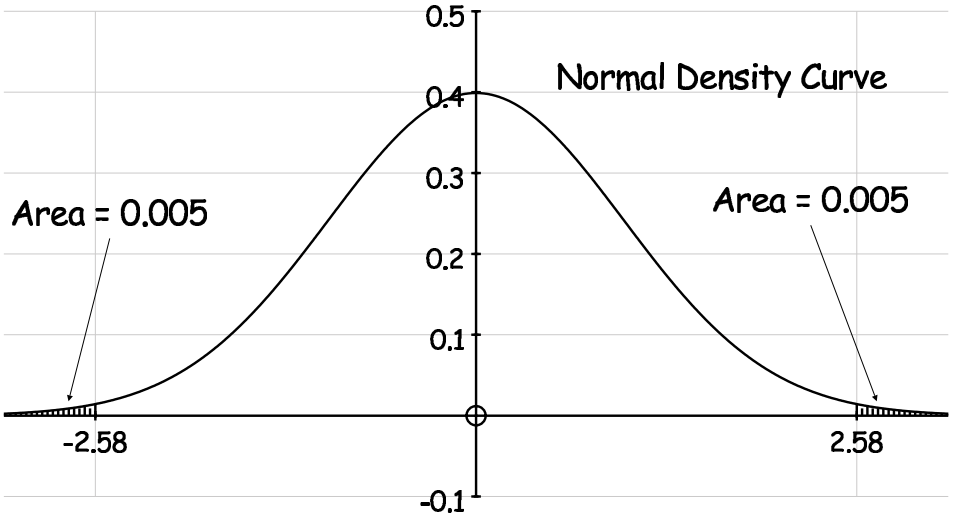
We may rearrange this and write

$$P(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}).$$

Once we have calculated the mean $\overline{x}$ of n independent samples, we obtain a specific interval $\left[\overline{x} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{x} - 1.96\frac{\sigma}{\sqrt{n}}\right]$ which we call the 95% **confidence interval** for the mean μ of the given population. Again, it's important to realize that once the sample has been taken and $\overline{x}$ has been calculated, there's nothing random at all about the above confidence interval: it's not correct that it contains the true mean μ with probability 95%, it either does or it doesn't!

Of course, there's nothing really special about the confidence level 95%—it's just a traditionally used one. Other confidence levels frequently used are 90% and 99%, but, of course, any confidence level

could be used. To form a 90% confidence interval from a measured mean $\overline{x}$, we would replace the number 1.96 used above with the value of z for which random samples from a normal population with mean 0 and standard deviation 1 would lie between $\pm z$ 99% of the time. Here, it turns out that $z \approx 2.58$:



In general, the $(1-\alpha) \times 100\%$ confidence interval for the mean is obtained by determining the value $z_{\alpha/2}$ such that a normally-distributed random variable Z of mean 0 and standard deviation 1 will satisfy

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) \;=\; 1-\alpha.$$

Below are tabulated some of the more traditional values:

Confidence Level $(\mathbf{1-\alpha})$	$\boldsymbol{\alpha}$	**Relevant z-value** $\boldsymbol{z_{\alpha/2}}$
0.90	0.10	1.645
0.95	0.05	1.960
0.98	0.02	2.326
0.99	0.01	2.576

In summary, the $(1-\alpha) \times 100\%$ confidence interval for the mean is formed from the sample mean $\overline{x}$ by constructing

$$\left[\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\, \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

Furthermore, we expect that $(1-\alpha)\times 100$ percent of the intervals so constructed will contain the true population mean.

Note, finally, that as the confidence level rises, the width of the confidence interval also increases. This is obvious for the wider the interval, the more confident we should be that it will "capture" the true population mean!

EXERCISES

1. Suppose that we are going to sample the random variable $X = 4 \times \mathsf{rand}$. Is this a normal random variable? What is the mean and variance of X? Suppose that we instead sample $\overline{X}$, where $X = 4\times\mathsf{rand}$ and $\overline{X}$ is computed by taking 50 independent samples and forming the average. Is $\overline{X}$ close to being normally distributed? To help in answering this question, write the simple TI code into your calculator

   PROGRAM: NORMCHECK
   :{0} → L_1
   :For(I,1,100)
   :4*rand(50)→ L_2
   :mean(L_2) → L_1(I)
   :END

   A moment's thought reveals that this program collects 100 samples of $\overline{X}$, where each mean is computed from 50 samples each and putting the result into list variable L_1. Finally draw a histograms of these 100 samples of the mean; does it look normal? This little experiment is tantamount to sending out 100 statisticians and having each collecting 50 independent samples and computing the mean. The statisticians all return to combine their results into a single histogram.

2. Suppose that you go out and collect 50 samples of the random variable $4 \times$ rand and compute the mean $\overline{x}$. Compute the 95% confidence interval so obtained. Does it contain the true mean μ? (See Exercise 1, above.)

3. We can build on Exercise 2, as follows. The following simple TI code can be used to count how many out of 100 95% confidence intervals for the mean μ of the random variable 4*rand will actually contain the true mean (= 2):

```
PROGRAM: CONFINT
  :0 → C
  :For(I,1,100)
  :4*rand(50)→ L1
  :mean(L1) → M
  :M − .32 → L
  :M + .32 → U
  :C + (L ≤ 2)(2 ≤ U) → C
  :END
  :Disp C
  :Stop
```

   (a) What is the number .32?

   (b) What is C trying to compute?

   (c) Run this a few times and explain what's going on.

6.4.2 Confidence intervals for the mean; unknown variance

In this section we shall develop a method for finding confidence intervals for the mean μ of a population when we don't already know the variance σ^2 of the population. In the last section our method was based on the fact that the statistic $\dfrac{\overline{X} - \mu}{\sigma}$ was approximately normally distributed. In the present section, since we don't know σ, we shall replace σ^2 with

its **unbiased estimate** s_x^2, the **sample variance**. We recall from page 375 that s_x^2 is defined in terms of the sample by setting

$$s_x^2 \;=\; \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2.$$

Again, this is unbiased because the expected value of this statistics is the population variance σ^2 (see the footnote on page 375.

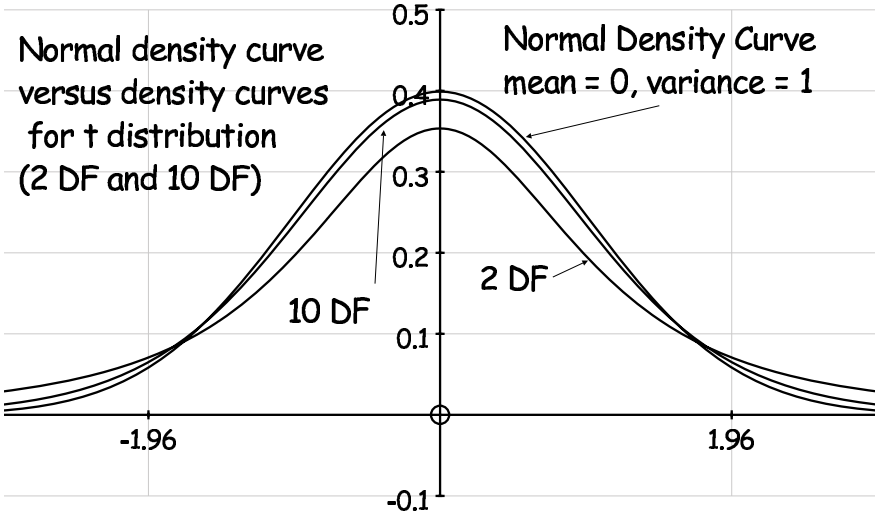
We now consider the statistic $T = \dfrac{\overline{X}-\mu}{S_x/\sqrt{n}}$ which takes on the value $t \;=\; \dfrac{\overline{x}-\mu}{s_x/\sqrt{n}}$ from a sample of size n. Of course, we don't know μ, but at least we can talk about the distribution of this statistic in two important situations, viz.,

- The sample size is small but the underlying population being sampled from is approximately normal; or
- The sample size is large ($n \geq 30$).

In either of the above two situations, T is called the ***t* statistic** and has what is called the ***t* distribution** with mean 0, variance 1 and having ***n* − 1 degrees of freedom**. If n is large, then T has close to a normal distribution with mean 0 and variance 1. However, even when n is large, one usually uses the t distribution.[25]

Below are the density functions for the t distribution with 2 and 10 degrees of freedom (DF). As the number of degrees of freedom tends to infinity, the density curve approaches the normal curve with mean 0 and variance 1.

[25]Before electronic calculators were as prevalent as they are today, using the t statistic was not altogether convenient as the t distribution changes slightly with each increased degree of freedom. Thus, when $n \geq 30$ one typically regarded T as normal and used the methods of the previous section to compute confidence intervals. However, the t distribution with any number of degrees of freedom is now readily available on such calculators as those in the TI series, making unnecessary using the normal approximation (and introducing additional error into the analyses).



The philosophy behind the confidence intervals where σ is unknown is pretty much the same as in the previous section. We first choose a desired level of confidence $(1-\alpha)\times 100\%$ and then choose the appropriate level $t_{\alpha/2}$ which contains $\alpha \times 100\%$ of the population in the two tails of the distribution. Of course, which t distribution we choose is dependent on the size of the sample we take; as mentioned above, the degrees of freedom is equal to $n-1$, where n is the sample size. These levels are tabulated in any statistics book; as a sample we show how they are typically displayed (a more complete table is given at the end of this chapter):

Degrees of Freedom	$\boldsymbol{t_{.050}}$	$\boldsymbol{t_{.025}}$	$\boldsymbol{t_{.005}}$
⋮	⋮	⋮	⋮
10	1.812	2.228	3.169
11	1.796	2.201	3.106
12	1.782	2.179	3.055
13	1.771	2.160	3.012
⋮	⋮	⋮	⋮

Once we have collected the sample of size n and have computed the average $\overline{x}$ of the sample, the $(1-\alpha)\times 100\%$ confidence interval becomes

$$\left[\overline{x} - t_{\alpha/2}\frac{s_x}{\sqrt{n}},\ \overline{x} + t_{\alpha/2}\frac{s_x}{\sqrt{n}}\right].$$

EXERCISE

1. As we have already seen it's possible to use your TI calculators to generate examples (simulations) of your own, as follows. Try the following.

   (a) On your TI, invoke randNorm$(10.3, 2.4, 5) \to L_1$. What does this command do?[26]

   (b) Next, use your TI calculator to compute a 95% confidence interval for the mean. (Use TInterval and run the Data option)

   (c) Did this interval capture the true mean?

   (d) If you were to perform this experiment 100 times, how many times would you expect the computed confidence interval to capture the true mean?

   (e) Here's a code that will construct the above confidence interval and compute the number of times it captures the true mean. Run it and report on your findings. (The run time for this program on a TI-83 is about four minutes.)

[26] It generates a (small) sample of size 5 taken from a normal population with mean 10.3 and standard deviation 2.4.

PROGRAM: CONFINT1
:$0 \to C$
:Input "POP MEAN ", M
:Input "POP STD ", S
:Input "NO OF EXPER ", N
:$5 \to$ K
:For(I,1,100)
:randNorm(M,S,K)$\to L_1$
:mean(L_1) $\to X$
:(K/(K-1))(mean(L_1^2) $-$ X^2) $\to$ V
:$2.776\sqrt{\text{V}/\text{K}} \to$ Q
:X $-$ Q $\to$ L
:X + Q $\to$ U
:$C + (L \leq 2)(2 \leq U) \to C$
:END
:Disp C
:Stop

(f) In the above program, change the commands as follows

Input "POP MEAN ", M	to	$1/3 \to$ M
randNorm(M, S, K) $\to L_1$	to	rand(K)$^2 \to L_1$
Input "POP STD ", S	to	anything (it's now irrelevant)

Notice that this time we are taking **small** samples from a highly non-normal population (rand^2). Are we still capturing the true mean (= 1/3) roughly 95% of the time?

6.4.3 Confidence interval for a population proportion

Professional pollsters love to estimate proportions: in any political race and at virtually any time, they will take samples from the voting population to determine whether they prefer candidate A or candidate B. Of course, what the pollsters are trying to determine is the overall preference—as a proportion—of the entire population. I seem to remember reading sometime during the 2004 U.S. presidential campaign

that a Gallop Poll survey of 10,000 voters led to the prediction that 51% of the American voters preferred Kerry over Bush with a sampling error of ±3% and a confidence level of 95%. What this means, of course, is the essence of confidence intervals for proportions.

The methods of this section are based on the assumption that large enough samples from an even larger binomial population are taken so that the test statistic—the sample proportion—can assumed to be normally distributed. Thus, we are going to be sampling from a very large binomial population, i.e., one with exactly two types A and B. If the population size is N, then the **population proportion** p can be then defined to be fraction of those of type A to N. When sampling from this population, we need for the population size to be rather large compared with the sample size. In practice, the sampling is typically done without replacement which strictly speaking would lead to a hypergeometric distribution. However, if the population size is much larger than the sample size, then the samples can be regarded as independent of each other, whether or not the sampling is done without replacement. Once a sample of size n has been taken, the **sample proportion** $\hat{p}$ is the statistic measuring the ratio of type A selected to the sample size n.

Assume, then, that we have a large population where p is the proportion of type A members. Each time we randomly select a member of this population, we have sampled a **Bernoulli random variable** B whose mean is p and whose variance is $p(1-p)$. By the Central Limit Theorem, when n is large, the sum $B_1 + B_2 + \cdots + B_n$ of n independent Bernoulli random variables, each having mean p and variance $p(1-p)$ has approximately a normal distribution with mean np and variance $np(1-p)$. The random variable

$$\widehat{P} \;=\; \frac{B_1 + B_2 + \cdots + B_n}{n}$$

is therefore approxmately normally distributed (when n is large) and has mean p and variance $\frac{p(1-p)}{n}$.

If $b_1, b_2, \ldots, b_n$ are the observed outcomes, i.e.,

$$b_i = \begin{cases} 1 & \text{if type A is observed;} \\ 0 & \text{if type B is observed,} \end{cases}$$

then the relevant test statistic is

$$\hat{p} = \frac{b_1 + b_2 + \cdots + b_n}{n}.$$

Notice that since we don't know p (we're tying to estimate it), we know **neither** the mean nor the variance of the test statistic. With a large enough sample, $\widehat{P}$ will be approximately normally distributed with mean p and variance $p(1-p)$. Therefore $\dfrac{\widehat{P} - p}{\sqrt{p(1-p)/n}}$ will be approximately normal with mean 0 and variance 1. The problem with the above is all of the occurrences of the unknown p. The remedy is to approximate the variance $\dfrac{p(1-p)}{n}$ by the sample variance based on $\hat{p}$: $\dfrac{\hat{p}(1-\hat{p})}{n}$. Therefore, we may regard

$$Z = \frac{\widehat{P} - p}{\sqrt{\widehat{P}(1-\widehat{P})/n}}$$

as being approximately normally distributed with mean 0 and variance 1. Having this we now build our $(1-\alpha) \times 100\%$ confidence intervals based on the values $z_{\alpha/2}$ taken from normal distribution with mean 0 and variance 1. That is to say, the $(1-\alpha) \times 100\%$ confidence interval for the population proportion p is

$$\left[\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \overline{x} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right].$$

CAUTION: If we are trying to estimate a population parameter which we know to be either very close to 0 or very close to 1, the method above performs rather poorly unless the sample size is **very large**.

The reason is the highly skewed nature of a binomial population with parameter p very close to either 0 or 1, meaning that the Central Limit Theorem will need much larger samples before the distribution starts to become acceptably normal. A proposed modification[27] is to replace $\hat{p}$ in the above interval by the new statistic $p^* = \dfrac{x+2}{n+4}$, where x is the measured number of type A members in the sample and n is the sample size. Also, the sample standard deviation $\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$ is replaced by the expression $\sqrt{\dfrac{p^*(1-p^*)}{n+4}}$. The resulting confidence interval performs better for the the full range of possibilities for p, even when n is small!

RULE OF THUMB: Since the methods of this section rely on the test statistics $\dfrac{\widehat{P}-p}{\sqrt{\widehat{P}(1-\widehat{P})/n}}$ being approximately normally distributed, any sort of guidance which will help us assess this assumption will be helpful. One typically used one is that if the approximate assumption of normality is satisfied, then $\hat{p} \pm$ three sample standard deviations should both lie in the interval $(0, 1)$. Failure of this to happen indicates that the sample size is not yet large enough to counteract the skewness in the binomial distribution. That is to say, we may assume that the methods of this section are valid provided that

$$0 < \hat{p} \pm 3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < 1.$$

6.4.4 Sample size and margin of error

In the above discussions we have seen the our confidence intervals had the form

$$\text{Estimate } \pm \text{ Margin of Error}$$

at a given confidence level. We have also seen that decreasing the

[27] See Agresti, A., and Coull, B.A., *Approximate is better than 'exact' for interval estimation of binomial proportions*, THE AMERICAN STATISTICIAN, Vol. 52, N. 2, May 1998, pp. 119–126.

margin of error also decreases the confidence level. A natural question to ask is whether we can decrease the margin of error *without* at the same time sacrificing confidence? The answer is yes: by *increasing the sample size*. We flesh this out in the following example.

EXAMPLE. Suppose that we are interested in the average cost μ of a new house in the United States in 1966, and that a random selection of the cost of 50 homes revealed the 95% confidence interval

$$\$20,116 \ \leq \mu \ \leq \$30,614,$$

along with the sample mean $\overline{x} \approx \$25,365$, and estimate $\sigma \approx s_x = \$18,469$. If we use this as an estimate of the population standard deviation σ, then we see that a $(1-\alpha) \times 100\%$ confidence interval becomes

$$\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

We see also that the margin of error associated with the above estimate is one-half the width of the above interval, viz., \$5,249.

QUESTION: Suppose that we wish to take a new sample of new houses and obtain a confidence interval for μ with the same level of confidence (95%) but with a margin of error of at most $\$3,000$?

SOLUTION. This is easy, for we wish to choose n to make the margin of error no more than $\$3,000$:

$$z_{.025}\frac{\sigma}{\sqrt{n}} \ \leq \ \$3,000.$$

Using $z_{.025} = 1.96$ and $\sigma \approx \$18,469$ we quickly arrive at

$$n \ \geq \ \left(\frac{1.96 \times 18,469}{3,000}\right)^2 \approx 146.$$

That is to say, if we take a sample of at least 146 data, then we will have narrowed to margin of error to no more than $\$3,000$ without sacrificing any confidence.

We can similarly determine sample sizes needed to a given bound on the margin of error in the case of confidence intervals for proportions, as follows. In this case the margin of error for a confidence interval with confidence $(1-\alpha)\times 100\%$ is $z_{\alpha/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$, where, as usual, $\hat{p}$ is the sampled population proportion. A very useful approximation is obtained by noting that since $0 \leq \hat{p} \leq 1$, then $0 \leq \hat{p}(1-\hat{p}) \leq \frac{1}{4}$. Therefore, if we wish for the margin of error to be less than a given bound B, all we need is a sample size of at least

$$n \geq \left(\frac{z_{\alpha/2}}{2B}\right)^2,$$

because regardless of the sampled value $\hat{p}$ we see that

$$\frac{z_{\alpha/2}}{2B} \;\geq\; \frac{z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})}}{B}.$$

EXERCISES

1. Assume that we need to estimate the mean diameter of a very critical bolt being manufactured at a given plant. Previous studies show that the machining process results in a standard deviation of approximately 0.012 mm. Estimate the sample size necessary to compute a 99% confidence interval for the mean bolt diameter with a margin of error of no more than 0.003 mm.

2. Assume that a polling agency wishes to survey the voting public to estimate the percentage of voters which prefer candidate A. What they seek is a sampling error of no more than .02% at a confidence level of 98%. Find a minimum sample size which will guarantee this level of confidence and precision.

6.5 Hypothesis Testing of Means and Proportions

Suppose we encounter the claim by a manufacturer that the precision bolts of Exercise 1 above have a mean of 8.1 mm and that we are

to test the accuracy of this claim. This claim can be regarded as a **hypothesis** and it is up to us as statisticians to decide whether or not to reject this hypothesis. The above hypothesis is usually called the **null hypothesis** and is an assertion about the mean μ about the population of manufactured bolts. This is often written

$$H_0: \quad \mu = 8.1.$$

We have no a priori reason to believe otherwise, unless, of course, we can find a **significant** reason to reject this hypothesis. In hypothesis testing, one typically doesn't **accept** a null hypothesis, one usually **rejects** (or doesn't reject) it on the basis of statistical evidence.

We can see there are four different outcomes regarding the hypothesis and its rejection. A **type I error** occurs when a true null hypothesis is rejected, and a **type II error** occurs when we fail to reject a false null hypothesis. These possibilities are outlined in the table below.

	$\boldsymbol{H_0}$ **is true**	$\boldsymbol{H_0}$ **is false**
Reject $\boldsymbol{H_0}$	Type I error	Correct decision
Do not reject $\boldsymbol{H_0}$	Correct decision	Type II error

Perhaps a useful comparison can be made with the U.S. system of criminal justice. In a court of law a defendent is **presumed innocent** (the null hypothesis), unless proved guilty ("beyond a shadow of doubt"). Convicting an innocent person is then tantamount to making a type I error. Failing to convict an guilty person is a type II error. Furthermore, the language used is strikingly similar to that used in statistics: the defendent is never found "innocent," rather, he is merely found "not guilty."

It is typical to define the following conditional probabilities:

$$\begin{aligned} \alpha &= P(\text{rejecting } H_0 \mid H_0 \text{ is true}), \\ \beta &= P(\text{not rejecting } H_0 \mid H_0 \text{ is false}). \end{aligned}$$

Notice that as α becomes smaller, β becomes larger, and vice versa.

Again, in the U.S. judicial justice system, it is assumed (or at least hoped) that α is **very small**, which means that β can be large (too large for many people's comfort).

Let's move now to a simple, but relatively concrete example. Assume that a sample of 60 bolts was gathered from the manufacturing plant whose claim was that the bolts they produce have a mean diameter of 8.1 mm. Suppose that you knew that the standard deviation of the bolts was $\sigma = 0.04$ mm. (As usual, it's unreasonable to assume that you would know this in advance!) The result of the sample of 60 bolts is that $\overline{x} = 8.117$. This doesn't look so bad; what should you do?

We proceed by checking how significantly this number is away from the mean, as following. First, notice that the **test statistic** (a random variable!)

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{60}},$$

where μ represents the hypothesized mean, will be approximately normally distributed with mean 0 and variance 1. The observed value of this test statistic is then

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{60}} \approx 3.29.$$

Whoa! Look at this number; it's over three standard deviations away from the mean of Z and hence is way out in the right-hand tail of the normal distribution.[28] The probability for us to have gotten such a large number under the correct assumption that $H_0 : \mu = 8.1$ were true is very small (roughly .1%). This suggests strongly that we reject this null hypothesis!

[28]The probability $P(|Z| \geq 3.29)$ of measuring a value this far from the mean is often called the P-value of the outcome of our measurement, and the smaller the P-value, the more significance we attribute to the result. In this particular case, $P(|Z| \geq 3.29) \approx 0.001$, which means that before taking our sample and measuring the sample mean, the probability that we would have gotten something this far from the true mean is roughly one-tenth of one percent!

Continuing the above example, assume more realistically that we didn't know in advance the variance of the population of bolts, but that in the sample of 60 bolts we measure a sample standard deviation of $s_x = .043$. In this case sample statistic

$$T = \frac{\overline{X} - \mu}{S_x/\sqrt{60}}$$

has the t distribution with 59 degrees of freedom (hence is very approximately normal). The observed value of this sample statistic is

$$t = \frac{\overline{x} - \mu}{s_x/\sqrt{60}} \approx 3.06.$$

As above, obtaining this result would be extremely unlikely if the hypothesis $H_0 : \ \mu = 8.1$ were true.

Having treated the above two examples informally, we shall, in the subsequent sections give a slightly more formal treatment. As we did with confidence intervals, we divide the treatment into the cases of known and unknown variances, taking them up individually in the next two sections.

EXERCISES

1. In leaving for school on an overcast April morning you make a judgement on the null hypothesis: The weather will remain dry. The following choices itemize the results of making type I and type II errors. Exactly one is true; which one?

   (A) Type I error: get drenched
   Type II error: needlessly carry around an umbrella

   (B) Type I error: needlessly carry around an umbrella
   Type II error: get drenched

   (C) Type I error: carry an umbrella, and it rains
   Type II error: carry no umbrella, but weather remains dry

   (D) Type I error: get drenched
   Type II error: carry no umbrella, but weather remains dry

(E) Type I error: get drenched
Type II error: carry an umbrella, and it rains

2. Mr. Surowski's grading policies have come under attack by the Central Administration as well as by the Board of Directors of SAS. To analyze the situation, a null hypothesis together with an alternative hypothesis have been formulated:

   H_0 : Mr. Surowski's grading policies are fair
   H_a: Mr. Surowski plays favorites in awarding grades.

   The Board of Directors finds no irregularities, and therefore takes no actions against him, but the rumors among the students is that it is advantageous for Mr. Surowski's students to regularly give him chocolate-covered expresso coffee beans. If the rumors are true, has an error been made? If so, which type of error?

3. An assembly-line machine produces precision bolts designed to have a mean diameter of 8.1 mm. Each morning the first 50 bearings produced are pulled and measured. If their mean diameter is under 7.8 mm or over 8.4 mm, the machinery is stopped and the foreman is called on to make adjustments before production is resumed. The quality control procedure may be viewed as a hypothesis test with the null hypothesis $H_0 : \mu = 8.1$. The engineer is asked to make adjustments when the null hypothesis is rejected. In test terminology, what would be the result of a Type II error (choose one)?

   (A) A warranted halt in production to adjust the machinery

   (B) An unnecessary stoppage of the production process

   (C) Continued production of wrong size bolts

   (D) Continued production of proper size bolts

   (E) Continued production of bolts that randomly are the right or wrong size

6.5.1 Hypothesis testing of the mean; known variance

Throughout this and the next section, the null hypothesis will have the form $H_0 : \ \mu = \mu_0$. However, in the course of rejecting this hypothesis, we shall consider **one-** and **two-sided alternative hypotheses**. The one-sided alternatives have the form $H_a : \ \mu < \mu_0$ or $H_a : \ \mu > \mu_0$. The two-sided alternatives have the form $H_a : \ \mu \neq \mu_0$.

A one-sided alternative is appropriate in cases where the null hypothesis is $H_0 : \ \mu = \mu_0$ but that anything $\leq \mu_0$ is acceptable (or that anything $\geq \mu_0$ is acceptable). This leads to two possible sets of hypotheses:

$$H_0 : \ \mu = \mu_0, \quad H_a : \ \mu < \mu_0,$$

or

$$H_0 : \ \mu = \mu_0, \quad H_a : \ \mu > \mu_0.$$

EXAMPLE 1. Suppose that a manufacturer of a mosquito repellant claims that the product remains effective for (at least) six hours. In this case, anything greater than or equal to six hours is acceptable and so the appropriate hypotheses are

$$H_0 : \ \mu = 6, \quad H_a : \ \mu < 6,$$

Therefore, a one-sided alternative is being used here.

EXAMPLE 2. In the example of precision bolts discussed above, large deviations on either side of the mean are unacceptable. Therefore, a two-sided alternative is appropriate:

$$H_0 : \ \mu = \mu_0, \quad H_a : \ \mu \neq \mu_0,$$

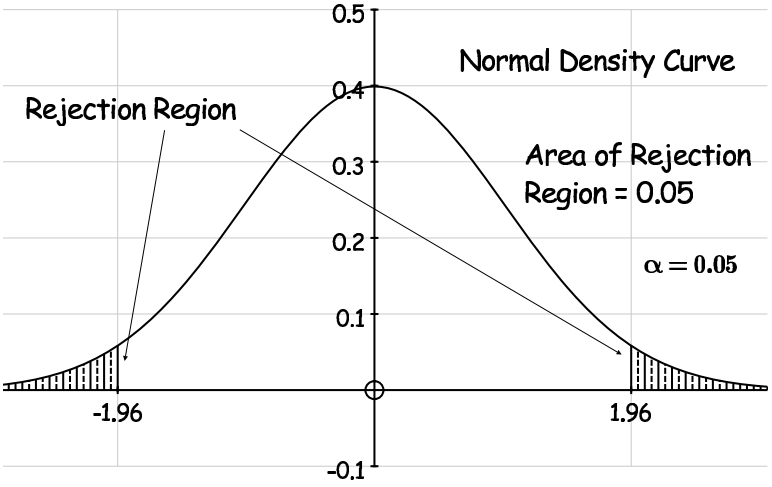
Next, one decides on a criterion by which H_0 is to be rejected; that is to say, on decides on the probability α of making a Type I error.

(Remember, the smaller this error becomes, the larger the probability β becomes of making a Type II error.) The most typical rejection level is $\alpha = 5\%$. As mentioned above, the test statistic becomes

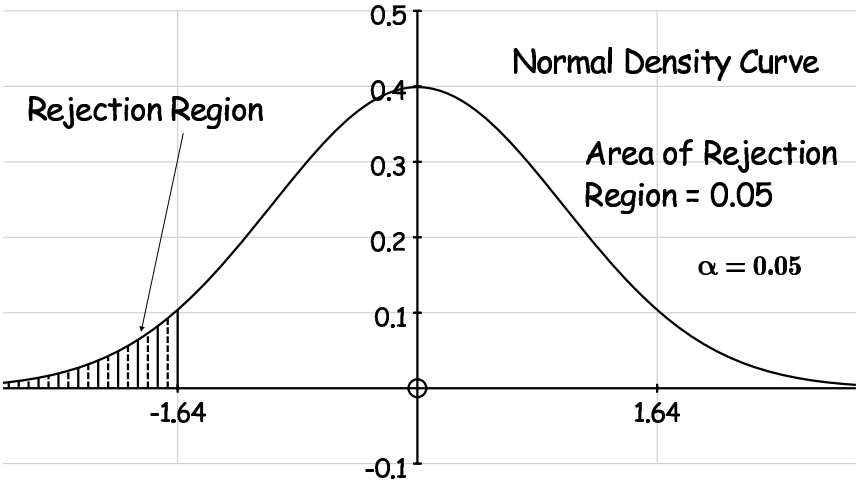
$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}.$$

This is normally distributed with mean 0 and variance 1. The 5% **rejection region** is dependent upon the alternative hypothesis. It's easiest just to draw these:

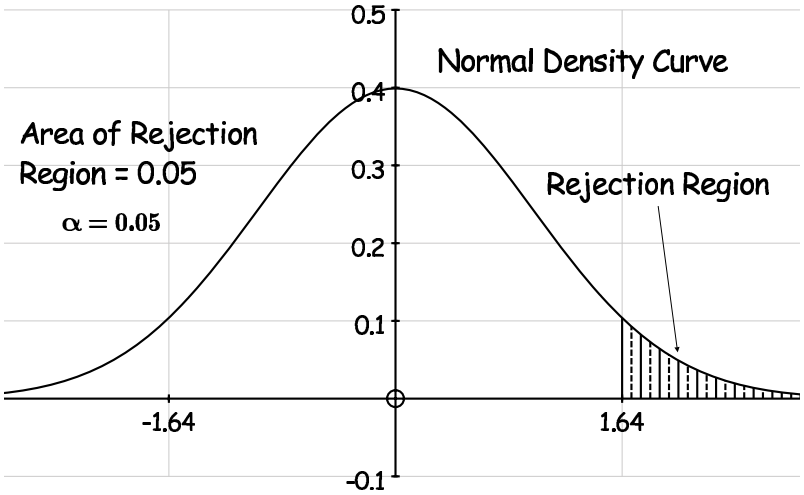
$\boldsymbol{H_0: \ \mu = \mu_0, \ \ H_a: \ \mu \neq \mu_0.}$



$\boldsymbol{H_0: \ \mu = \mu_0, \ \ H_a: \ \mu < \mu_0.}$



$\boldsymbol{H_0: \ \mu = \mu_0, \ \ H_a: \ \mu > \mu_0.}$



6.5.2 Hypothesis testing of the mean; unknown variance

In this setting, the formulation of the null and alternative hypotheses don't change. What changes is the test statistic:

$$T = \frac{\overline{X} - \mu}{S_x \sqrt{n}}.$$

This has the t-distribution with $n - 1$ degrees of freedom in either of the two cases itemized on page 386, namely either when we're sampling from an approximately normal population or when the sample size is reasonably large. As in the previous section, the rejection regions at the α level of significance are determined on the basis of the alternative hypothesis. Furthermore, unless one implements the test automatically (as on a TI calculator), in finding the boundary of the rejection region one needs to consider the number of degrees of freedom of the t statistic.

6.5.3 Hypothesis testing of a proportion

If we encounter the claim that at least 55% percent of the American voting public prefer candidate A over candidate B, then a reasonable set of hypotheses to test is

$$H_0 : \ p = .55, \quad H_a : p < .55.$$

The test statistic would then be

$$Z \ = \ \frac{\widehat{P} - p}{\sqrt{\widehat{P}(1-\widehat{P})/n}},$$

which is n is large enough is approximately normally distributed. Therefore testing the hypothesis at the $(1-\alpha)\%$ level of significance can be handled in the usual fashion.

6.5.4 Matched pairs

One of the most frequent uses of statistics comes in evaluating the effect of one or more **treatments** on a set of subjects. For example, people often consider the effects of listening to Mozart while performing an intellectual task (such as a test). In the same vein, one may wish to compare the effects of two insect repellants.

To effect such comparisons, there are two basic—but rather different—experimental designs which could be employed. The first would be to divide a group of subjects into two distinct groups and apply the different "treatments" to the two groups. For example, we may divide the group of students taking a given test into groups A and B, where group A listens to Mozart while taking the test, whereas those in group B do not. Another approach to comparing treatments is to successively apply the treatments to the same members of a given group; this design is often called a **matched-pairs design**. In comparing the effects of listening of Mozart, we could take the same group of students and allow them to listen to Mozart while taking one test and then at another time have them take a similar test without listening to Mozart.

In such situations, we would find ourselves comparing μ_1 versus μ_2, where μ_1, μ_2 represent means associated with treatments 1 and 2. The sensible null hypothesis would be expressed as $H_0 : \ \mu_1 = \mu_2$ and the alternative will be either one or two sided, depending on the situation.

Without delving into the pros and cons of the above two designs, suffice it to say that the statistics used are different. The first design

represents taking two independent samples; we won't go into the appropriate statistic for evaluating H_0. On the other hand, the matched-pairs design is easier; all one needs to do is to compute the difference of the two effects. If X is the random variable representing the difference, and if μ_X is the mean of X, then we are simply evaluating $H_0 : \ \mu_X = 0$ against an appropriately-chosen alternative. The methods of the above sections apply immediately.

EXERCISES

1. The TI calculator command $\mathsf{randNorm}(0, 1)$ generates a random number from a normal distribution with mean 0 and variance 1. Do you believe this? The command

$$\mathsf{randNorm}(0, 1, 100) \longrightarrow L_1$$

will put 100 independent samples of this distribution into the list variable L_1. Test the hypothesis $\mu = 0$ at the 99% significance level.

2. [29] Sarah cycles to work and she believes that the mean time taken to complete her journey is 30 minutes. To test her belief, she records the times (in minutes) taken to complete her journey over a 10-day period as follows:

30.1	32.3	33.6	29.8	28.9	30.6	31.1	30.2	32.1	29.4

Test Sarah's belief, at the 5% significance level.

3. Suppose it is claimed that 80% of all SAS graduating seniors go on to attend American Universities. Set up null and alternative hypotheses for testing this claim.

4. Suppose instead it is claimed that at least 80% of all SAS graduating seniors go on to attend American Universities. Set up null and alternative hypotheses for testing this claim.

[29] From IB Mathematics HL Examination, May 2006, Paper 3 (Statistics and Probability), #3.

5. Suppose that a coin is tossed 320 times, with the total number of "heads" being 140. At the 5% level, should the null hypothesis that the coin is fair be rejected?

6. A candidate claims that he has the support of at least 54% of the voting public. A random survey of 1000 voters reveals that among those sampled, this candidate only had the support of 51%. How would you report these results?

7. Ten healthy subjects had their diastolic blood pressures measured before and after a certain treatment. Evaluate the null hypothesis that there was no change against the alternative that the blood pressure was lowered as a result of the treatment. Use a 95% significance level.

Systolic Blood Pressure

Before Treatment	83	89	86	91	84	91	88	90	86	90
After Treatment	77	83	85	92	85	86	91	88	88	83

8. A growing number of employers are trying to hold down the costs that they pay for medical insurance for their employees. As part of this effort, many medical insurance companies are now requiring clients to use generic-brand medicines when filling prescriptions. An independent consumer advocacy group wanted to determine if there was a difference, in milligrams, in the amount of active ingredient between a certain "name" brand drug and its generic counterpart. Pharmacies may store drugs under different conditions. Therefore, the consumer group randomly selected ten different pharmacies in a large city and filled two prescriptions at each of these pharmacies, one of the "name" brand and the other for the generic brand of the same drug. The consumer group's laboratory then tested a randomly selected pill from each prescription to determine the amount of active ingredient in the pill. The results are given in the table below.

Active Ingredient (in milligrams)

Pharmacy	1	2	3	4	5	6	7	8	9	10
Name brand	245	244	240	250	243	246	246	246	247	250
Generic brand	246	240	235	237	243	239	241	238	238	234

Based on the above data, what should be the consumer group's laboratory report about the difference in the active ingredient in the two brands of pills? Give appropriate statistical evidence to support your response.

6.6 χ^2 and Goodness of Fit

Perhaps an oversimplification, the χ^2 statistic gives us a means for measuring the discrepancy between how we feel something ought to behave versus how it actually appears to behave. In order to flesh out this very cryptic characterization, suppose we have a die which we believe to be fair, and roll it 200 times, with the outcomes as follows:

Outcome	1	2	3	4	5	6
No. of occurrences	33	40	39	28	36	24

Does this appear to be the way a fair die should behave? Is there a statistic appropriate for us to measure whether its descrepancy from "fairness" is significant?

Notice that the underlying null hypothesis would be that the die is fair, expressed in the form

$$H_0: \ p_1 = \frac{1}{6}, \ p_2 = \frac{1}{6}, \ p_3 = \frac{1}{6}, \ p_4 = \frac{1}{6}, \ p_5 = \frac{1}{6}, \ p_6 = \frac{1}{6},$$

(where the probabilities have the obvious definitions) versus the alternative

$$H_a: \quad \text{at least one of the proportions exceeds } \frac{1}{6}.$$

The appropriate test statistic, sometimes called the **χ^2 statistic**, is

given by

$$\chi^2 \;=\; \sum \frac{(n_i - E(n_i))^2}{E(n_i)},$$

where the sum is over each of the k possible outcomes (in this case, $k = 6$), where n_i is the number of times we observe outcome i, and where $E(n_i)$ is the expected number of times we would observe the outcome under the null hypothesis. Thus, in the present case, we would have $E(n_i) = 200/6,\ i = 1, 2, \ldots, 6$, and $n_1 = 33$, $n_2 = 40$, $\ldots$, $n_6 = 24$. Of course, just because we have denoted this sum by χ^2 doesn't already guarantee that it has a χ^2 distribution. Checking that it really does would again take us into much deeper waters. However, if we consider the simplest of all cases, namely when there are only two categories, then we can argue that the distribution of the above statistic really is approximately χ^2 (and with one degree of freedom).

When there are only two categories, then of course we're really doing a **binomial experiment**. Assume, then, that we make n measurements (or "trials") and that the probability of observing a outcome falling into category 1 is p. This would imply that if n_1 is the number of observations in category 1, then $E(n_1) = np$. Likewise, the if n_2 is the number of observations in category 2, then $n_2 = n - n_1$ and $E(n_2) = n(1 - p)$.

In this case, our sum takes on the appearance

$$\begin{aligned}
\chi^2 = \frac{(n_1 - E(n_1))^2}{E(n_1)} + \frac{(n_2 - E(n_2))^2}{E(n_2)} &= \frac{(n_1 - np)^2}{np} + \frac{(n_2 - n(1-p))^2}{n(1-p)} \\
&= \frac{(n_1 - np)^2}{np} + \frac{(n - n_1 - n(1-p))^2}{n(1-p)} \\
&= \frac{(n_1 - np)^2}{np} + \frac{(n_1 - np)^2}{n(1-p)} \\
&= \frac{(n_1 - np)^2}{np(1-p)}
\end{aligned}$$

However, we may regard n_1 as the observed value of a binomial random variable N_1 with mean np and standard deviation $\sqrt{np\,(1-p)}$; furthermore, if n is large enough, then N_1 is approximately normal. Therefore

$$Z = \frac{N_1 - np}{\sqrt{np\,(1-p)}}$$

is approximately normally distributed with mean 0 and standard deviation 1. This means that

$$Z^2 = \frac{(N_1 - np)^2}{np\,(1-p)}$$

has approximately the χ^2 distribution with one degree of freedom, completing the argument in this very special case.

EXAMPLE 1. Let's flesh out the above in a very simple hypothesis-testing context. That is, suppose that someone hands you a coin and tells you that it is a fair coin. This leads you to test the hypotheses

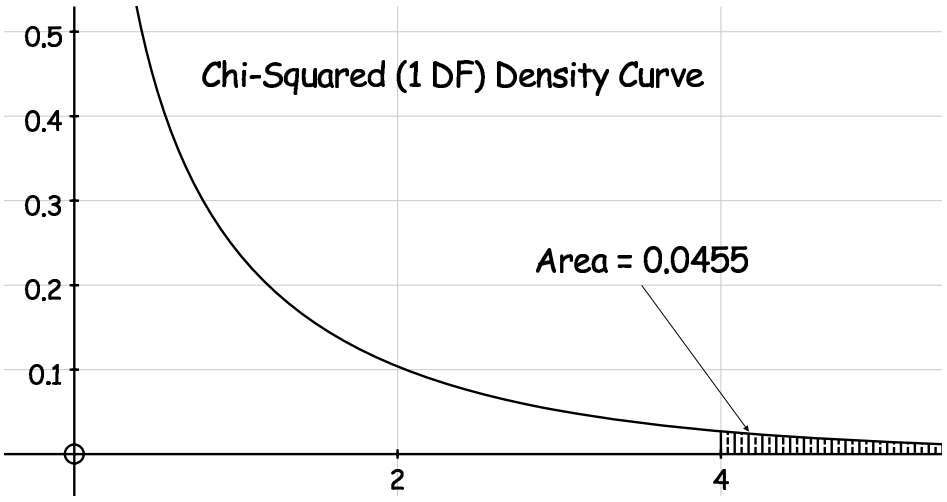
$$H_0 : p = 1/2 \quad \text{against the alternative} \quad H_a : p \neq 1/2,$$

where p is the probability that the coins lands on heads on any given toss.

To test this you might then toss the coin 100 times. Under the null hypothesis, we would have $E(n_H) = 50$, where n_H is the number of heads (the random variable in this situation) observed in 100 tosses. Assume that as a result of these 100 tosses, we get $n_H = 60$, and so $n_T = 40$, where, obviously, n_T is the number of tails. We plug into the above χ^2 statistic, obtaining

$$\chi^2 = \frac{(60-50)^2}{50} + \frac{(40-50)^2}{50} = 2 + 2 = 4.$$

So what is the P-value of this result? As usual, this is the probability $P(\chi^2 \geq 4)$ which is the area under the χ^2-density curve for $x \geq 4$:



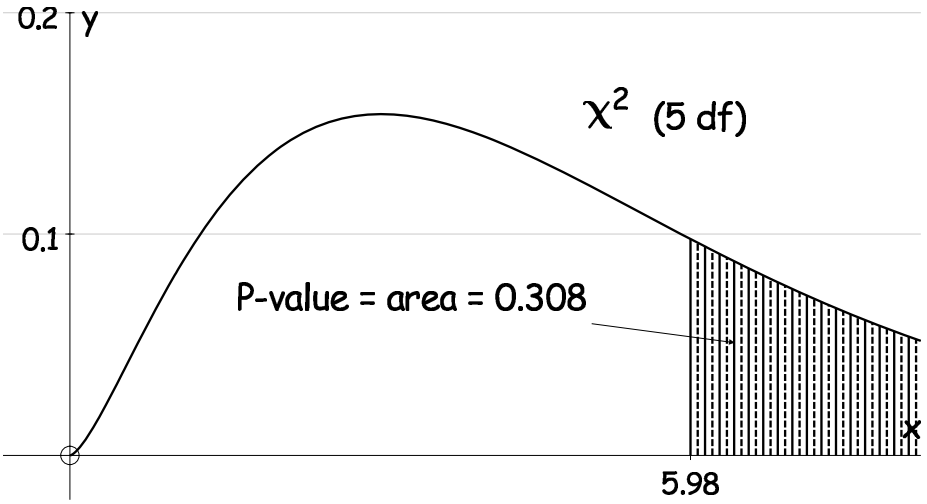
(The above calculation can be done on your TI-83, using $1-\chi^2\text{cdf}(0, 4, 1)$. The third argument, 1, is just the number of degrees of freedom.)

Since the P-value is .0455, one sees that there is a fair amount of significance that can be attached to this outcome. We would—at the 5% level of significance—reject H_0 and say that the coin is not fair.

EXAMPLE 2. Let's return to the case of the allegedly fair die and the results of the 200 tosses. The χ^2 test results in the value:

$$\begin{aligned}\chi^2 &= \sum \frac{(n_i - E(n_i))^2}{E(n_i)} \\ &= \frac{(33-\frac{200}{6})^2}{200/6} + \frac{(40-\frac{200}{6})^2}{200/6} + \frac{(39-\frac{200}{6})^2}{200/6} \\ &+ \frac{(28-\frac{200}{6})^2}{200/6} + \frac{(36-\frac{200}{6})^2}{200/6} + \frac{(24-\frac{200}{6})^2}{200/6} \approx 5.98.\end{aligned}$$

The P-value corresponding to this measured value is $P(\chi^2_5 \geq 5.98) \approx 0.308$. (We sometimes write χ^2_n for the χ^2 random variable with n degrees of freedom.) This is really not small enough (not "significant enough) for us to reject the null hypothesis of the fairness of the die.



In general, experiments of the above type are called **multinomial experiments** , which generalize in a natural way the familiar binomial experiments. A multinomial experiment results in a number of occurrences in each of possibly many categories; the binomial experiment results in a number of occurrences in each of only two categories. The result of a multinomial experiment is typically summarized in a **one-way table**, the table on page 405 being a good example. The χ^2 test used to test the null hypothesis regarding the individual category probabilities is often referred to as a **test for homogeneity**.

The TI-83 calculators are quite adept at a variety of tests of hypotheses; however, they don't have a built-in code to test homogeneity hypotheses.[30] (They do, however, have a built-in code to test for independence, which we'll consider below.) At any rate, here's a simple code which will test for homogeneity. Preparatory to running this program, one puts into list variable L_1 the observed counts and into list variable L_2 the expected counts. For the problem of the putative fair die, the entries 33, 40, 39, 28, 36, 24 are placed in L_1 and the entry $200/6 = 33.333$ is placed into the first six entries of L_2.

The following simple code will finish the job:

[30]This was remedied on the TI-84s.

PROGRAM: CHISQ
: Input " DF ", N
: $0 \to S$
: For$(I, 1, N+1)$
: $S + (L_1(I) - L_2(I))^2/(L_2(I)) \to S$
: End
: $1 - \chi^2$cdf$(0, S, N) \to P$
: Disp "CHISQ:", S
: Disp "P-VALUE", P

Running the above program (using the fact that there are 5 degrees of freedom) results in the output:

CHISQ : 5.98
P − VALUE : .308

Example 3.[31] Suppose that before a documentary was aired on public television, it had been determined that 7% of the viewing public favored legalization of marijuana, 18% favored decriminalization (but not legalization), 65% favored the existing laws, and 10% had no opinion. After the documentary was aired, a random sample of 500 viewers revealed the following opinions, summarized in the following one-way table:

Distribution of Opinions About Marijuana Possession

Legalization	Decriminalization	Existing Laws	No Opinion
39	99	336	26

Running the above TI code yielded the following output:

CHISQ: 13.24945005
P-VALUE: .0041270649

This tells us that there is a significant departure from the pre-existing proportions, suggesting that the documentary had a significant effect

[31]This example comes from STATISTICS, Ninth edition, James T. McClave and Terry Sinich, Prentice Hall, 2003, page 710. (This is the text we use for our AP Statistics course.)

on the viewers!

6.6.1 χ^2 tests of independence; two-way tables

Students who have attended my classes will probably have heard me make a number of rather cavalier—sometimes even reckless—statements. One that I've often made, despite having only anecdotal evidence, is that among students having been exposed to both algebra and geometry, girls prefer algebra and boys prefer algebra. Now suppose that we go out and put this to a test, taking a survey of 300 students which results in the following **two-way contingency table**[32]:

		Gender		
		Male	Female	Totals
Subject Preference	Prefers Algebra	69	86	155
	Prefers Geometry	78	67	145
	Totals	147	153	300

Inherent in the above table are two categorical random variables X=gender and Y= subject preference. We're trying to assess the independence of the two variables, which would form our null hypothesis, versus the alternative that there is a gender dependency on the subject preference.

In order to make the above more precise, assume, for the sake of argument that we knew the exact distributions of X and Y, say that

$$P(X = \text{male}) = p, \quad \text{and} \quad P(Y \text{ prefers algebra}) = q.$$

If X and Y are really independent, then we have equations such as

$$\begin{aligned} P(X = \text{male and } Y \text{ prefers algebra}) &= P(X = \text{male}) \cdot P(Y \text{ prefers algebra}) \\ &= pq. \end{aligned}$$

Given this, we would expect that among the 300 students sampled, roughly $300pq$ would be males and prefer algebra. Given that the actual

[32]These numbers are hypothetical—I just made them up!

number in this category was found to be 69, then the contribution to the χ^2 statistic would be

$$\frac{(69 - 300pq)^2}{300pq}.$$

Likewise, there would be three other contributions to the χ^2 statistic, one for each "cell" in the above table.

However, it's unlikely that we know the parameters of either X or Y, so we use the data in the table to estimate these quantities. Clearly, the most reasonable estimate of p is $\hat{p} = \frac{147}{300}$ and the most reasonable estimate for q is $\hat{q} = \frac{155}{300}$. This says that the estimated expected count of those in the Male/Algebra category becomes $E(n_{11}) = 300 \times \frac{147}{300} \times \frac{155}{300} = \frac{147 \cdot 155}{300}$. This makes the corresponding to the χ^2 statistic

$$\frac{(n_{11} - E(n_{11}))^2}{E(n_{11})} = \frac{(69 - \frac{147 \cdot 155}{300})^2}{\left(\frac{147 \cdot 155}{300}\right)}.$$

The full χ^2 statistic in this example is

$$\begin{aligned}
\chi^2 &= \frac{(n_{11} - E(n_{11}))^2}{E(n_{11})} + \frac{(n_{12} - E(n_{12}))^2}{E(n_{12})} + \frac{(n_{21} - E(n_{21}))^2}{E(n_{21})} + \frac{(n_{22} - E(n_{22}))^2}{E(n_{22})} \\
&= \frac{(69 - \frac{147 \cdot 155}{300})^2}{\left(\frac{147 \cdot 155}{300}\right)} + \frac{(86 - \frac{153 \cdot 155}{300})^2}{\left(\frac{153 \cdot 155}{300}\right)} + \frac{(78 - \frac{147 \cdot 145}{300})^2}{\left(\frac{147 \cdot 145}{300}\right)} + \frac{(67 - \frac{153 \cdot 145}{300})^2}{\left(\frac{153 \cdot 145}{300}\right)} \\
&\approx 2.58.
\end{aligned}$$

We mention finally, that the above χ^2 has only 1 degree of freedom: this is the number of rows minus 1 times the number of columns minus 1. The P-value associated with the above result is $P(\chi_1^2 \geq 2.58) = 0.108$. Note this this result puts us in somewhat murky waters, it's small (significant) but perhaps not small enough to reject the null hypothesis of independence. Maybe another survey is called for!

In general, given a two-way contingency table, we wish to assess whether the random variables defined by the rows and the columns

are independent. If the table has r rows and c columns, then we shall denote the entries of the table by n_{ij}, where $1 \leq i \leq r$ and $1 \leq j \leq c$. The entries n_{ij} are often referred to as the **cell counts**. The sum of all the cell counts is the total number n in the sample. We denote by $C_1, C_2, \ldots, C_c$ the column sums and by $R_1, R_2, \ldots, R_r$ the row sums. Then in analogy with the above example, the contribution to the χ^2 statistic from the (i, j) table cell is $(n_{ij} - \frac{R_i C_j}{n})^2 / \frac{R_i C_j}{n}$, as under the null hypothesis of independence of the random variables defined by the rows and the columns, the fraction $\frac{R_i C_j}{n}$ represents the **expected** cell count. The complete χ^2 statistic is given by the sum of the above contributions:

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - \frac{R_i C_j}{n})^2}{\left(\frac{R_i C_j}{n}\right)},$$

and has $(r-1)(c-1)$ degrees of freedom.

EXAMPLE 3. It is often contended that one's physical health is dependent upon one's material wealth, which we'll simply equate with one's salary. So suppose that a survey of 895 male adults resulted in the following contingency table:

	Salary (in thousands U.S.$)				
Health	15–29	30–39	40–59	≥ 60	Totals
Fair	52	35	76	63	226
Good	89	83	78	82	332
Excellent	88	83	85	81	337
Totals	229	201	239	226	895

One computes $\chi^2_6 = 13.840$. Since $P(\chi^2_6 \geq 13.840) = 0.031$, one infers a significant deviation from what one would expect if the variables really were independent. Therefore, we reject the independence assumption. Of course, we still can't say any more about the "nature" of the dependency of the salary variable and the health variable. More detailed analyses would require further samples and further studies!

We mention finally that the above can be handled relatively easily by the TI calculator χ^2 test. This test requires a single matrix input, A,

where, in this case, A would be the cell counts in the above contingency table. The TI-calculator will automatically generate from the matrix A a secondary matrix B consisting of the expected counts. Invoking the χ^2 test using the matrix

$$A = \begin{bmatrix} 52 & 35 & 76 & 63 \\ 89 & 83 & 78 & 82 \\ 88 & 83 & 85 & 81 \end{bmatrix}$$

results in the output

χ^2-Test

$\chi^2 = 13.83966079$

P=.0314794347

df=6.

EXERCISES

1. The TI command *randInt*(0,9) will randomly generate an integer (a "digit") between 0 and 9. Having nothing better to do, we invoke this command 200 times, resulting in the table:

digit	0	1	2	3	4	5	6	7	8	9
frequency	17	21	15	19	25	27	19	23	18	17

   We suspect that the command randInt ought to generate random digits *uniformly*, leading to the null hypothesis

$$H_0 : \ p_i = \frac{1}{10}, \ i = 0, 1, 2, \ldots, 9,$$

   where p_i is the probability of generating digit i, $i = 0, 1, 2, \ldots, 9$. Test this hypothesis against its negation at the 5% significance level.

2. [33] Eggs at a farm are sold in boxes of six. Each egg is either brown or white. The owner believes that the number of brown eggs in a

[33] Adapted from IB Mathematics HL Examination, Nov 2003, Paper 2 (Statistics), #6 (iv).

box can be modeled by a binomial distribution. He examines 100 boxes an obtains the following data:

Number of brown eggs in a box	Frequency
0	10
1	29
2	31
3	18
4	8
5	3
6	1

(a) Estimate the percentage p of brown eggs in the population of all eggs.

(b) How well does the binomial distribution with parameter p model the above data? Test at the 5% level.

3. Suppose you take six coins and toss them simultaneously 100, leading to the data below:

Number of heads obtained	Frequency	Expected under H_0
0	0	
1	4	
2	13	
3	34	
4	30	
5	15	
6	4	

Suppose that I tell you that of these six coins, five are fair and one has two heads. Test this as a null hypothesis at the 5% level. (Start by filling in the expected counts under the appropriate null hypothesis.)

4. Here's a more extended exercise. In Exercise 18 on page 344 it was suggested that the histogram representing the number of trials

needed for each of 200 people to obtain all of five different prizes bears a resemblance with the Poisson distribution. Use the TI code given in part (c) to generate your own data, and then use a χ^2 test to compare the goodness of a Poisson fit. (Note that the mean waiting time for five prizes is $\mu = \frac{137}{12}$.)

5. People often contend that divorce rates are, in some sense, related to one's religious affiliation. Suppose that a survey resulted in the following data, exhibited in the following two-way contingency table:

		Religious Affiliation				
		A	B	C	None	Totals
Marital History	Divorced	21	32	15	32	100
	Never Divorced	78	90	34	90	292
	Totals	99	122	49	122	392

Formulate an appropriate null hypothesis and test this at the 5% level.

6. (Here's a cute one!)[34] The two-way contingency table below compares the level of education of a sample of Kansas pig farmers with the sizes of their farms, measured in number of pigs. Formulate and test an appropriate null hypothesis at the 5% level.

		Education Level		
		No College	College	Totals
Farm Size	<1,000 pigs	42	53	95
	1,000–2,000 pigs	27	42	69
	2,001–5,000 pigs	22	20	42
	>5,000 pigs	27	29	56
	Totals	118	144	262

[34] Adapted from STATISTICS, Ninth edition, James T. McClave and Terry Sinich, Prentice Hall, 2003, page 726, problem #13.26.

Table entry for p and C is the critical value t^* with probability p lying to its right and probability C lying between $-t^*$ and t^*.

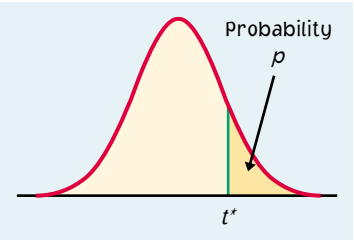


TABLE C t distribution critical values

	Upper tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
z^*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

Index