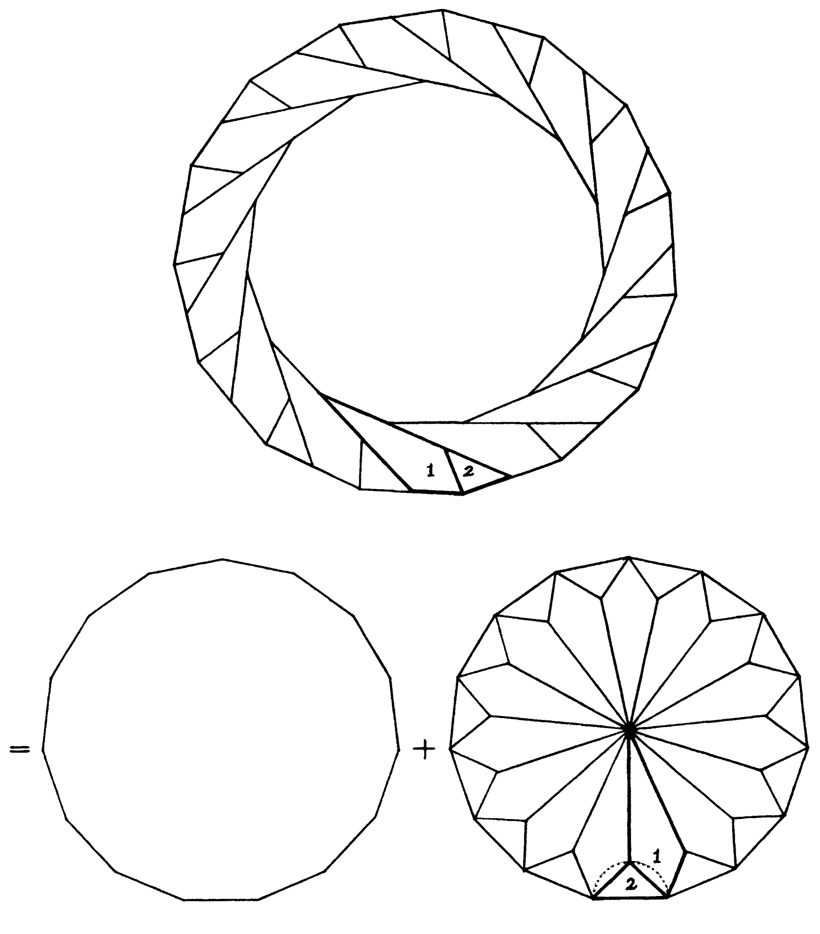
Vol. 79, No. 2, April 2006

MATHEMATICS MAGAZINE



- Reflecting Well: Dissections of Two Regular Polygons to One
- The Evolution of the Normal Distribution
- Pythagorean Triples and the Problem $A = mP$ for Triangles
- Euler's Ratio-Sum Theorem and Generalizations

An Official Publication of The MATHEMATICAL ASSOCIATION OF AMERICA

EDITORIAL POLICY

Mathematics Magazine aims to provide lively and appealing mathematical exposition. The *Magazine* is not a research journal, so the terse style appropriate for such a journal (lemma-theorem-proof-corollary) is not appropriate for the *Magazine*. Articles should include examples, applications, historical background, and illustrations, where appropriate. They should be attractive and accessible to undergraduates and would, ideally, be helpful in supplementing undergraduate courses or in stimulating student investigations. Manuscripts on history are especially welcome, as are those showing relationships among various branches of mathematics and between mathematics and other disciplines.

A more detailed statement of author guidelines appears in this *Magazine*, Vol. 74, pp. 75–76, and is available from the Editor or at www.maa.org/pubs/mathmag.html. Manuscripts to be submitted should not be concurrently submitted to, accepted for publication by, or published by another journal or publisher.

Submit new manuscripts to Allen Schwenk, Editor, Mathematics Magazine, Department of Mathematics, Western Michigan University, Kalamazoo, MI, 49008. Manuscripts should be laser printed, with wide line spacing, and prepared in a style consistent with the format of *Mathematics Magazine*. Authors should mail three copies and keep one copy. In addition, authors should supply the full five-symbol 2000 Mathematics Subject Classification number, as described in *Mathematical Reviews*.

Cover image: see page 130

AUTHORS

Greg Frederickson has been a professor of computer science at Purdue University for the last quarter century. His research area is the design and analysis of algorithms, in particular graph and network algorithms and data structures. He has authored two books, "Dissections: Plane & Fancy" and "Hinged Dissections: Swinging & Twisting." Originally trained in economics, he is now both a consumer and a producer: He enjoys reflecting on the gems of the past as well as creating the raw material for future reflection.

Saul Stahl received his B.A., M.A. and Ph.D. at Brooklyn College, The University of California at Berkeley, and Western Michigan University, respectively. His research interest is graph theory, and his mathematical avocation is the evolution of mathematical concepts. He has published six junior/senior level texts that incorporate the evolution of the subject matter into its pedagogy. His other hobbies include dance, and currently he is learning the Argentine tango.

Lubomir Markov is Assistant Professor of Mathematics at Barry University in Miami, Florida. His main area is analysis, with strong secondary interests in other fields, especially number theory and physics. Born in the sage city of Sofia, Bulgaria, he obtained his master's degree from the University of Sofia and then came to the USA where he obtained a Ph.D. from the University of South Florida with a dissertation in nonlinear functional analysis. Among his favorite nonmathematical activities are reading and biking, and lately he has been toying with the idea that some day he might try some cooking; whether that will happen before the proof of the transcendence of $e\pi$ remains an open question.

Branko Grünbaum was born in what used to be Yugoslavia, and received his Ph.D. from the Hebrew University in Jerusalem. After coming to the United States in 1958, he spent various lengths of time at the Institute for Advanced Study in Princeton, UCLA, Kansas University, University of Washington, Michigan State University, and the Hebrew University. He settled in 1966 at the University of Washington, where he became Emeritus in 2000. His interests range from combinatorics to many areas of discrete and combinatorial geometry.

Murray Seymour Klamkin was born in the United Stated. He received his undergraduate degree in Chemistry, served four years in the U.S. Army during WW2, and later earned an M.S. in Physics. In his career he worked both in industry and in universities, and was for several years the Chair of the Mathematics Department at the University of Alberta in Edmonton. He is well-known to readers of *Mathematics Magazine* (and many other journals) through his activities in Olympiads and in the problems sections. More details about the life and achievements of Klamkin can be found in the eulogies published in *Focus* (November 2004, page 32) and in the *Notes of the Canadian Mathematical Society* (November 2004, pages 19–20).

Grünbaum and Klamkin first met in 1967, at a CUPM conference in Santa Barbara. They remained in contact over the years, but the present paper is their first joint publication.

Vol. 79, No. 2, April 2006

MATHEMATICS MAGAZINE

MATHEMATICS MAGAZINE (ISSN 0025-570X) is published by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, D.C. 20036 and Montpelier, VT, bimonthly except July/August.
The annual subscription price for *MATHEMATICS MAGAZINE* to an individual member of the Association is $131. Student and unemployed members receive a 66% dues discount; emeritus members receive a 50% discount; and new members receive a 20% dues discount for the first two years of membership.)

Subscription correspondence and notice of change of address should be sent to the Membership/ Subscriptions Department, Mathematical Association of America, 1529 Eighteenth Street, N.W., Washington, D.C. 20036. Microfilmed issues may be obtained from University Microfilms International, Serials Bid Coordinator, 300 North Zeeb Road, Ann Arbor, MI 48106.

Advertising correspondence should be addressed to Frank Peterson (*FPetersonj@aol.com*), Advertising Manager, the Mathematical Association of America, 1529 Eighteenth Street, N.W., Washington, D.C. 20036.

Periodicals postage paid at Washington, D.C. and additional mailing offices.

Postmaster: Send address changes to Membership/ Subscriptions Department, Mathematical Association of America, 1529 Eighteenth Street, N.W., Washington, D.C. 20036-1385.

Printed in the United States of America

ARTICLES

Reflecting Well: Dissections of Two Regular Polygons to One

GREG N. FREDERICKSON
Department of Computer Science
Purdue University
West Lafayette, IN 47907
gnf@cs.purdue.edu

A geometric dissection is a cutting of one or more geometric figures into pieces that we can rearrange to form another figure [**5, 8**]. Some of the most beautiful geometric dissections are those of two congruent regular polygons to one larger regular polygon with twice the area. For many years now, we have known of two general methods, which for any $n \geq 5$ produce such dissections for regular n-sided polygons that use $2n+1$ pieces. Since the resulting dissections are lovely and symmetrical and the first method was discovered a century and a quarter ago, we might be tempted to assume that they cannot be improved upon. Yet this story could certainly not be complete without relating how the recent recovery of a "lost manuscript" catalyzed the discovery of improved methods, ones that will make you quite literally flip.

Let's start by taking a peek at FIGURE 1, a dissection of two regular pentakaidecagons (15 sides) to one. The diagram comes from the world's first book-length manuscript [**6**] on geometric dissections, which was lost for nearly fifty years and was recovered only a few years ago. Compiled by Ernest Irving Freese, an energetic and idiosyncratic architect from Los Angeles, and completed shortly before his death in 1957, it predates Harry Lindgren's volume [**8**] by seven years. The long-lost manuscript consists of 200 plates that contain over 300 dissections by Freese and others. Exquisitely laid out, with graphically stunning artwork and attractive hand lettering and numbering, it is jam-packed with irresistible diagrams like the one in FIGURE 1.

Freese's dissection is easy to understand and works for any regular n-sided polygon. For simplicity, let's denote a regular n-sided polygon by $\{n\}$. Freese left one of the two small $\{n\}$s whole and cut the other one into $2n$ pieces. Of these, n pieces are isosceles right triangles whose hypotenuses have length equal to the sidelength of the small $\{n\}$. The other n pieces are pointy quadrilaterals with a $360°/n$ angle coinciding with the center of the small $\{n\}$, an angle of 135° on either side of the first angle, and the remaining angle of $90°–360°/n$. When we assemble the pieces into the large $\{n\}$, we arrange the pointy quadrilaterals around the uncut $\{n\}$, taking advantage of the fact that the exterior angle made by two adjacent edges of the central $\{n\}$ is $360°/n$. We thus position each quadrilateral so as to fill the exterior angle and thus extend the line for each side further. Finally, we place each isosceles right triangle so that its 45° angles fit against the 135° angles of consecutive quadrilaterals. Note that the right angle of the isosceles right triangle fits up against the $90°–360°/n$ angle of a quadrilateral, creating the desired angle of the large $\{n\}$.

After enjoying the visual and mathematical appeal of this figure, we might address two questions that are not so easy to answer. First, since Freese's manuscript contains no references to prior work, how do we determine to what extent he relied on the ideas

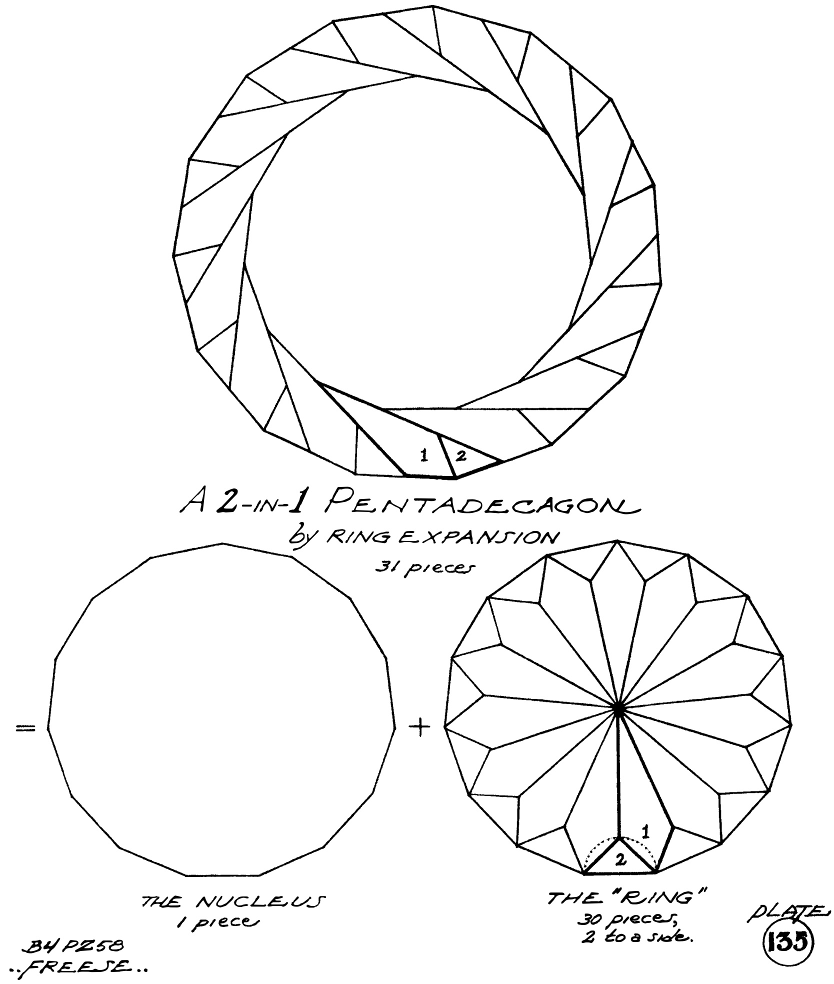


Figure 1 Plate 135 from Ernest Irving Freese's *Geometric Transformations*

of others? And second, in terms of minimizing the number of pieces, how close are his dissections to being optimal? We shall employ various sorts of reflection to lead us to answers for both of these questions. Reaching back into the late 19th century from the mid-20th century, we will identify two people who most likely played a role in the discovery of the method employed in FIGURE 1. And proceeding forward from Freese's time, we shall explore new improvements inspired by his technique.

Freese was obviously intrigued by two-to-one dissections: His manuscript presents many examples for congruent regular figures, namely triangles, squares, pentagons, hexagons, octagons, enneagons, decagons, dodecagons, pentakaidecagons, hexakaidecagons, icosagons, and icosikaitetragons. (Freese used the alternate names nonagon, pentadecagon, hexadecagon, and icositetragon.) Some of Freese's dissections mentioned above are specific to certain regular polygons, such as his 9-piece dissection of two hexagons to one (his Plate 68). Some are applicable to certain classes of regular polygons, such as his n-piece dissection of $\{n\}$s, where n is a multiple of 4. And some are quite general, such as the $(2n+1)$-piece dissection that we have already discussed, which turns out to be a variation of a method by the English geometer Harry Hart [**7**].

Relationship to earlier discoveries

Prior to Freese's work, Hart had described a general and elegant method for dissecting two similar polygons to one. It works for two different classes of polygons, as Hart explained, and uses $2n+1$ pieces for n-sided polygons (Note below that Euclid I. 47 is the Pythagorean theorem.):

> The truth of Euclid I. 47 has been shown in many ways by the dissection of two squares and transposition of the parts, and in Mr. Perigal's proof (*Messenger*, vol. II., pp. 103–106), the larger of the two squares is so dissected that the parts of it can be placed so as to form as it were the four corner portions of the square on the hypothenuse, leaving a vacant space in the middle into which the smaller square (which remains intact) may be exactly fitted.
>
> I have thus been led to examine whether two similar polygons might be divided and the parts so placed as to form a third polygon similar to either, the smaller polygon being left intact. In two cases I have that this can be simply effected.
>
> 1st. When the polygons admit of circles being inscribed in them. [...]
>
> 2ndly. When the polygons admit of circles being described about them. [...]

Although Hart's method is correct, the diagram that he provided to demonstrate his method is rather unattractive and confusing. Clear descriptions are available in [**9**] and [**5**, pp. 224–226] and will not be repeated here. Instead, let's examine Hart's method when applied to regular pentakaidecagons (FIGURE 2). Note that a regular polygon is simultaneously in both classes of polygons to which his methods apply, and we get identical dissections in such a case. In that setting, Hart would leave one small $\{n\}$ whole and cut the other into $2n$ pieces. He would next make a cut from the center of the small $\{n\}$ to the midpoint of each side, resulting in n wedge-shaped pieces, whose pointy vertex has an angle of $360°/n$. Hart would then cut an isosceles right triangle out of each wedge, with one vertex of the right triangle coincident with a vertex from the small $\{n\}$ and the right angle flush against a right angle of the wedge. The resulting n pieces are quadrilaterals.

When we assemble the $2n+1$ pieces into the large $\{n\}$, we place the quadrilaterals around the uncut $\{n\}$, taking advantage of the fact that the exterior angle made by

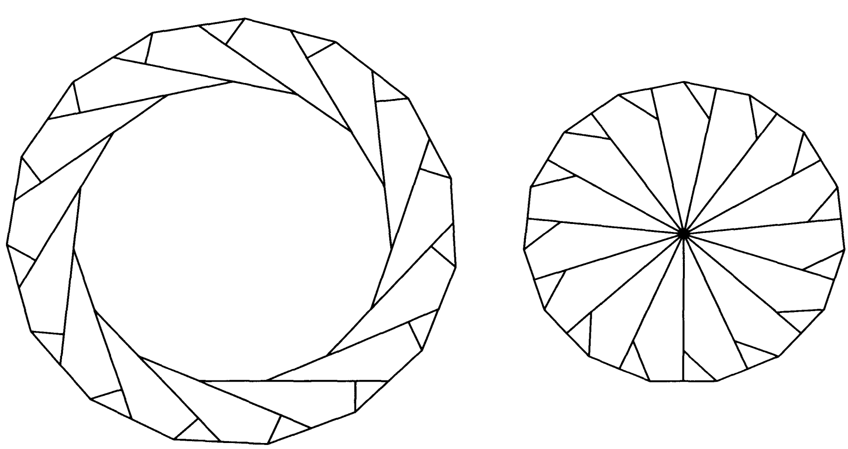


Figure 2 Hart's dissection applied to two {15}s to one

two adjacent edges of the central $\{n\}$ is $360°/n$. We thus place each quadrilateral so as to fill the exterior angle and thus extend the line for each side further. Finally, we place each isosceles right triangle so that its right angle fits against the right angle of a quadrilateral. Freese also illustrated in his manuscript examples of Hart's method when restricted to regular polygons.

Had Freese rediscovered this subcase of Hart's method? It is tantalizing to wonder whether Freese had also been inspired by Perigal's dissection and had also discovered the idea behind Hart's dissections. However, a loose piece of paper found in Freese's effects suggests a more likely scenario. Handwritten by Freese and entitled "(References in Rouse Ball) Geometrical Dissections," it neatly lists the author, topic, journal, year, volume, and pages for the articles of Henry Perigal [**13**], Henry Dudeney [**3**, p. 32], Henry Taylor [**14**], James Travers [**15**], William Macaulay [**10, 11, 12**], Brooks, Smith, Stone, and Tutte [**1**], and Albert Wheeler [**16**].

Three of these references are to articles in the *Messenger of Mathematics*, a complete collection of which was available at the library at Caltech. Located in Pasadena, California, the California Institute of Technology was a scant 5-mile drive from Freese's home in Highland Park. While looking up these references, Freese could easily have encountered Hart's article, which appeared only four years after Perigal's. He could have inspected the figures at the end of the volume or identified Hart's article by its title, "Geometric Dissections and Transpositions." Freese was aware of Perigal's dissection, because he illustrated it in his Plate 35. Reflecting on this situation, I believe it likely that Freese was aware of Hart's article.

Yet if Freese were aware of the article, why did he limit his exposition to regular figures, rather than the two classes of polygons that Hart identified? With its unsymmetrical figures and its inconsistent layout, Hart's diagram could have put off Freese and convinced him to stick with the regular polygons that appear in abundance throughout his manuscript.

Further reflection suggests how Freese could have come to produce his nifty variation for congruent regular polygons from Hart's dissection. Once you think to look for the variation, it is straightforward to derive the one dissection from the other. The solid lines in FIGURE 3 show a portion of Hart's dissection as applied to pentakaidecagons. In the complete dissection there would be 31 pieces: 15 copies of piece 1, plus 15 copies of piece 2, plus the uncut pentakaidecagon.

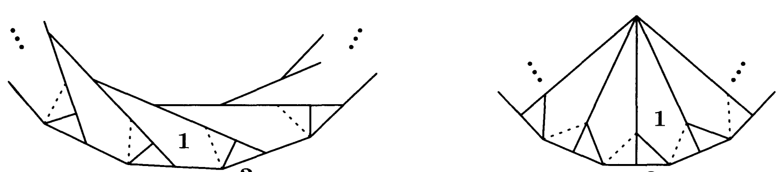


Figure 3 Deriving Freese's dissection from Hart's dissection

What Freese probably did was to take Hart's dissection and cut an isosceles right triangle off of each piece 1 and merge that triangle with a corresponding piece 2, producing an isosceles right triangle of twice the area. We see the new cuts as dotted lines in FIGURE 3 and note that the merging to get larger isosceles right triangles is consistent in the small pentakaidecagon and in the ring formed from its pieces. Unlike Hart's method, which applies to any two similar polygons in two broad classes, Freese's variation requires that the polygons be regular and congruent. Both Hart's dissection and Freese's variation use $2n+1$ pieces for n-sided polygons.

Improved dissections

So Freese appears to have retreated from the greater generality of Hart's dissection, only to achieve a bit more symmetry in the process. Yet as long as we are going to cut and merge isosceles right triangles, perhaps we should try to achieve a greater gain. As shown in a portion of the ring in FIGURE 4, we could first merge a small isosceles right triangle onto the second longest side of piece 1. We could then accommodate this expanded piece in the small pentakaidecagon that we cut, if we mate it with a copy of the original piece 1 that we have reflected and from which we also cut out a small isosceles right triangle.

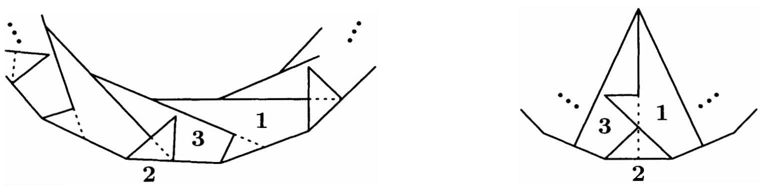


Figure 4 Deriving the new dissection from Hart's dissection

The resulting piece 3 has a cavity in the ring into which a merged-together pair of isosceles right triangles (piece 2) fits quite nicely. Thus we take advantage of the restriction that the polygons be regular. This approach produces $2n - \lfloor n/2 \rfloor + 1$ pieces, of which we must turn over $\lfloor n/2 \rfloor$ of them. (We can imagine such pieces as having an $*$ on one side and a $\star$ on the other.) The resulting 24-piece pentakaidecagon dissection is in FIGURE 5. Its major drawback is that the accommodating nontriangular pieces are so accommodating that they allow (and even require) themselves to be turned over!

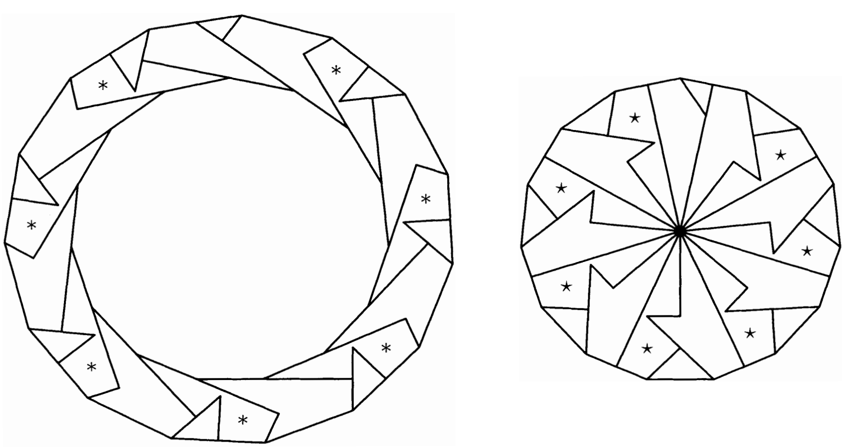


Figure 5 24-piece dissection of two {15}s to one, with pieces turned over

This adaptation will work whenever $n > 12$. As is clear from FIGURES 4 and 5, the method works when $n = 15$. As the number of sides decreases, the neck of piece 3 becomes thinner, until $n = 12$, in which case piece 3 is severed by piece 2 in the ring and by piece 1 in the small dodecagon. We can see this in FIGURE 6, where line

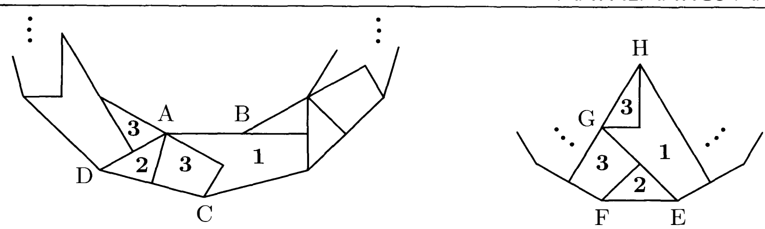


Figure 6 The severing of piece 3 by either piece 2 or piece 1 when $n = 12$

segments AB, AD, EF, and GH are all the same length, which is also equal to the distances from A to C, from B to C, and from F to G.

Should we be allowed to turn over pieces? In his manuscript, Freese gave no dissections that turned over pieces. Furthermore, he went out of his way in several dissections (in his Plates 88, 134, and 169) to avoid turning over pieces at the expense of using more of them. Specifically, in his dissection of a pentakaidecagon to a square (his Plate 134), he cut a right triangle into two isosceles triangles that he reassembled to form the mirror-image right triangle. Freese stated his opinion in a letter to Dorman Luke dated November 4, 1956:

> You are entirely correct in repeating my criticism of Cundy's "dissection" of the dodecagon (Cundy – pages 22+23) wherein he states that "six of the pieces must be turned over".

Thus if we want this inquiry to reflect well on our sense of fair play, then we had better find a dissection that has fewer than 31 pieces, but with none turned over. In response, I give a 28-piece dissection in FIGURE 8. The idea is to cut an $\lfloor n/2 \rfloor / n$ fraction of the large $\{n\}$ from FIGURE 5, and then flip it over to give (more or less) its mirror image.

In FIGURE 7, I have cut the large $\{n\}$ into sections of size 7/15 and 8/15 (with dotted lines) and then reflected the smaller section. One cut goes from the top vertex of the large $\{n\}$ to its center. A second cut goes from the center to the bottom left vertex of the large $\{n\}$. To make the figures clearer, I will not only specify angle measurements as a function of n but also instantiate them for $n = 15$. The two small right triangles (marked by an S) that are cut from the top of the small $\{n\}$, which have acute angles of $45°–180°/n = 33°$ and $45° + 180°/n = 57°$, fill in a gap at the bottom of the new ring. Also, two somewhat larger triangles (marked by an L) cut from pieces beneath the center of the small $\{n\}$, which have angles $360°/n = 24°$, $45° + 180°/n = 57°$, and $135°–540°/n = 99°$, fill in a space just below the top two pieces in the new ring.

A slice that goes from a vertex of the larger pentakaidecagon to its center will not go through a vertex of the smaller, inner pentakaidecagon. Thus we cut the inner pentakaidecagon and rearrange the pieces (forming an irregular $(n+2)$-sided polygon) rather than flipping a piece. Each cut on the small $\{n\}$ on the right is at the same distance from its nearest vertex, ensuring that the three pieces form the $(n+2)$-sided polygon.

When n is odd, note that before the cuts indicated by dotted lines in the small $\{n\}$ on the left, all but two pieces are grouped into groups of a piece 1, a piece 2, and a piece 3 that form a fat wedge, with the appropriate number of these fat wedges reflected. The very thin quadrilateral will have one angle of $180°–360°/n = 156°$, two angles of $135°–540°/n = 99°$, and the last angle of $1440°/n–90° = 6°$. Finally, note that we can save two pieces if we merge mirror-image pairs of small triangles created by the initial slices and also cut holes to accommodate these merged pieces.

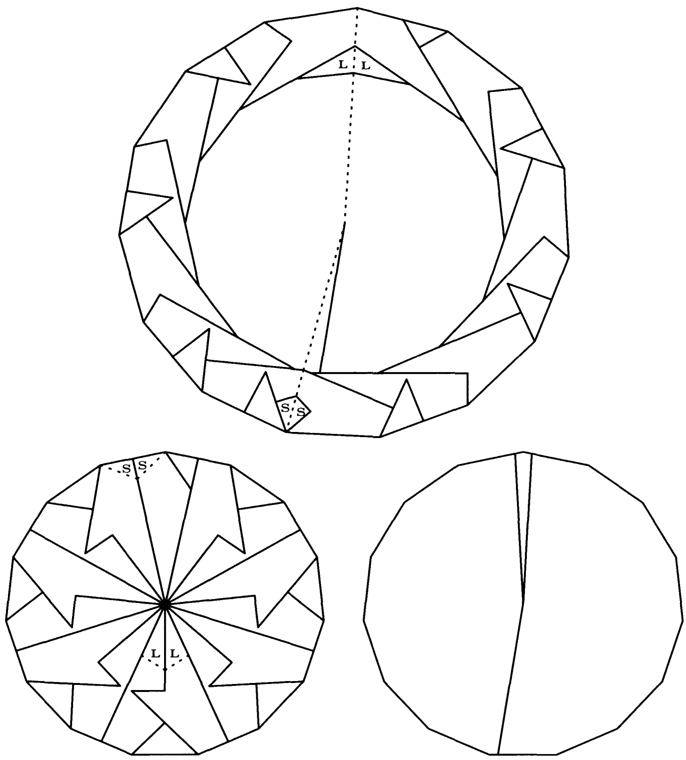


Figure 7 Deriving the 28-piece dissection of two {15}s to one in FIGURE 8

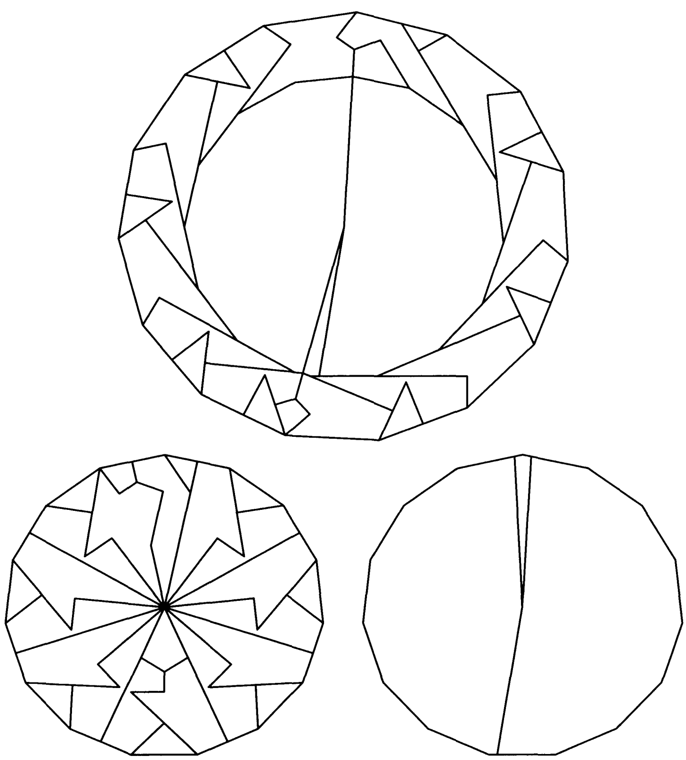


Figure 8 28-piece dissection of two {15}s to one, with no pieces turned over

It is not hard to verify that these ideas will produce a dissection of fewer than $2n+1$ pieces for any regular polygon with $n > 12$ sides. The total number of pieces will turn out to be at most $2n+1 - \lfloor n/2 \rfloor + 2\lceil n/8 \rceil$, derived as follows: We start with $2n - \lfloor n/2 \rfloor + 1$ pieces, as illustrated in FIGURE 5. I will shortly show that when we cut through one side of the ring (as shown in FIGURE 7), we cut through at most $\lceil n/8 \rceil$ pieces. Since we make two such cuts, each creating one additional piece for each piece we cut, we increase the number of pieces by at most $2\lceil n/8 \rceil$. Next we subtract 2, because we can always match up two pairs of mirror-image triangles as we did in FIGURE 8. Finally we add 2, because we split the other small $\{n\}$ into 3 pieces. This then leads directly to the claimed bound.

To prove that we cut at most $\lceil n/8 \rceil$ pieces when we cut through one side of the ring, consider the large $\{n\}$ in FIGURE 9, copied from FIGURE 5. I have added a dashed edge that goes from the center of the large $\{n\}$ to the midpoint of a side of the small $\{n\}$. A second dashed edge starts at that midpoint and follows an edge of one of the pieces in the ring out to the midpoint of a side of the large $\{n\}$. A third dashed edge goes from the midpoint of this side back to the center of the large $\{n\}$, thus completing a triangle described with dashed edges. This triangle is a right triangle, because the first dashed edge is perpendicular to the second. Since the large $\{n\}$ has twice the area of the small $\{n\}$, the length of the third dashed edge is $\sqrt{2}$ times the length of the first dashed edge. By the Pythagorean theorem, the second dashed edge will be of the same length as the first. Thus we have an isosceles right triangle, with an angle of 45° at the center of the large $\{n\}$.

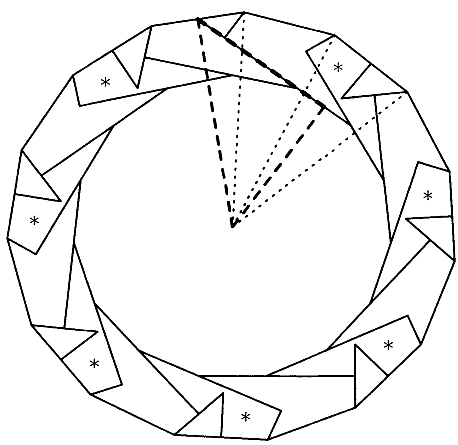


Figure 9 Bounding the number of pieces cut when cutting the ring

Now let's consider dotted line segments from the center of the large $\{n\}$ to its vertices. (I draw a few of the dotted line segments in FIGURE 9, namely those near the dashed triangle.) Because the dotted line segments come every $360°/n$ around the ring and the nontriangular pieces also come every $360°/n$, the number of pieces that one such dotted line segment crosses equals the number of dotted line segments that cross one piece. (Note, crucially, that the nontriangular pieces are arranged in the ring such that no dotted line segment goes through the interior of the piece, exits the piece, then re-enters the interior of the same piece.) The number of dotted line segments crossing such a piece will equal the number of dotted line segments crossing the outer leg of the dashed triangle. This number is simply bounded above by the acute angle of the dashed triangle divided by the angle between two consecutive dotted line segments, or

$\lceil 45°/(360°/n) \rceil$, which equals $\lceil n/8 \rceil$. Thus each cut of one side of the ring cuts through at most $\lceil n/8 \rceil$ pieces, as claimed.

When $n > 16$, each of the two cuts through the ring can create more than two triangles, so that there will be more than two pairs of triangles. I suspect that the technique of matching up pairs of mirror-image triangles can be extended to handle all such pairs. Thus I conjecture that dissections with $2n+3 - \lfloor n/2 \rfloor + \lceil n/8 \rceil$ pieces are possible, although it may be tedious to work out the details.

Extensions of the ideas behind FIGURES 5 and 8 lead to other dissections that are beyond the scope of this article. Those dissections, and the ones in this article, are discussed in a book **[4]** that I am writing.

Conclusion

We started with a lovely dissection of two regular 15-sided polygons to one in a 50-year-old manuscript and examined questions prompted by the dissection. We concluded that the author of this manuscript, Ernest Irving Freese, was probably aware of the previous work of Harry Hart, and showed how Freese might have derived his own variation from Hart's method. We then went on to extend the method and see what advantage we could create by trading around sections of pieces. In the end, we reduced the number of pieces for all but a finite number of cases, using reflection in several ways. We can only hope that these successive improvements, from the 19th to the 20th to the 21st century, reflect well on us all.

Acknowledgment. I am especially grateful to Vanessa Kibbe for providing me with access to Ernest Irving Freese's manuscript and records. It is my pleasure to acknowledge Dana Roth, reference librarian at the Millikan Library at Caltech, for determining the acquisition history of relevant volumes of the *Messenger of Mathematics*. Finally, I would like to thank the referees for helpful suggestions.

REFERENCES

1. R. L. Brooks, C. A. B. Smith, A. H. Stone, and W. T. Tutte, The dissection of rectangles into squares, *Duke Math. J.* **7** (1940), 312–340.
2. H. Martyn Cundy and A. P. Rollett, *Mathematical Models*, Oxford, 1952.
3. Henry Ernest Dudeney, *Amusements in Mathematics*, Thomas Nelson and Sons, London, 1917.
4. Greg N. Frederickson, Ernest Irving Freese's *Geometric Transformations:* the Lovely Log of LA's Lone Wolf, Manuscript plus commentary.
5. Greg N. Frederickson, *Dissections Plane & Fancy*, Cambridge Univ. Pr., New York, 1997.
6. Ernest Irving Freese, Geometric transformations, A graphic record of explorations and discoveries in the diversional domain of Dissective Geometry. Comprising 200 plates of expository examples. Unpublished manuscript, 1957.
7. Harry Hart, Geometrical dissections and transpositions, *Messenger of Mathematics* **6** (1877), 150–151.
8. Harry Lindgren, *Geometric Dissections*, D. Van Nostrand Company, Princeton, New Jersey, 1964.
9. Harry Lindgren, Assembling two polygons into one, *J. Recreational Math.* **2** (1969), 178–180.
10. W. H. Macaulay, The dissection of rectilineal figures, *Math. Gaz.* **7** (1914), 381–388.
11. W. H. Macaulay, The dissection of rectilineal figures, *Math. Gaz.* **8** (1915), 72–76 and 109–115.
12. W. H. Macaulay, The dissection of rectilineal figures, *Messenger of Mathematics* **48** (1919), 159–165.
13. Henry Perigal, On geometric dissections and transformations, *Messenger of Mathematics* **2** (1873), 103–105.
14. H. M. Taylor, On some geometrical dissections, *Messenger of Mathematics* **35** (1905), 81–101.
15. J. Travers, Puzzles and problems, *The Education Outlook* (1933), 145.
16. A. H. Wheeler, Problem E4, *Amer. Math. Monthly* **42** (1935), 509–510.

The Evolution of the Normal Distribution

SAUL STAHL
Department of Mathematics
University of Kansas
Lawrence, KS 66045, USA
stahl@math.ku.edu

Statistics is the most widely applied of all mathematical disciplines and at the center of statistics lies the normal distribution, known to millions of people as the bell curve, or the bell-shaped curve. This is actually a two-parameter family of curves that are graphs of the equation

$$y = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{1}$$

Several of these curves appear in FIGURE 1. Not only is the bell curve familiar to these millions, but they also know of its main use: to describe the general, or idealized, shape of graphs of data. It has, of course, many other uses and plays as significant a role in the social sciences as differentiation does in the natural sciences. As is the case with many important mathematical concepts, the rise of this curve to its current prominence makes for a tale that is both instructive and amusing. Actually, there are two tales here: the invention of the curve as a tool for computing probabilities and the recognition of its utility in describing data sets.

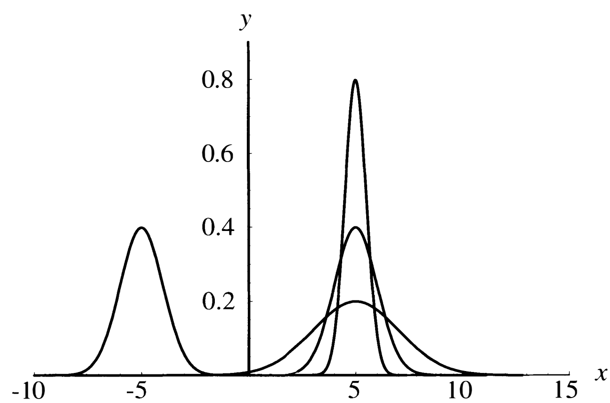


Figure 1 Bell-shaped curves

An approximation tool

The origins of the mathematical theory of probability are justly attributed to the famous correspondence between Fermat and Pascal, which was instigated in 1654 by the queries of the gambling Chevalier de Méré [**6**]. Among the various types of problems they considered were binomial distributions, which today would be described by such sums as

$$\sum_{k=i}^{j} \binom{n}{k} p^k (1-p)^{n-k}. \tag{2}$$

This sum denotes the likelihood of between i and j successes in n trials with success probability p. Such a trial—now called a *Bernoulli trial*—is the most elementary of all random experiments. It has two outcomes, usually termed *success* and *failure*. The kth term in (2) is the probability that k of the n trials are successful.

As the binomial examples Fermat and Pascal worked out involved only small values of n, they were not concerned with the computational challenge presented by the evaluation of general sums of this type. However, more complicated computations were not long in coming.

For example, in 1712 the Dutch mathematician 'sGravesande tested the hypothesis that male and female births are equally likely against the actual births in London over the 82 years 1629–1710 [**14, 16**]. He noted that the relative number of male births varies from a low of $7765/15,448 = 0.5027$ in 1703 to a high of $4748/8855 = 0.5362$ in 1661. 'sGravesande multiplied these ratios by 11,429, the average number of births over this 82 year span. These gave him nominal bounds of 5745 and 6128 on the number of male births in each year. Consequently, the probability that the observed excess of male births is due to randomness alone is the 82nd power of

$$\Pr\left[5745 \le x \le 6128 \mid p = \frac{1}{2}\right] = \sum_{x=5745}^{6128} \binom{11,429}{x} \left(\frac{1}{2}\right)^{11,429}$$
$$\approx \frac{3,849,150}{13,196,800} \approx 0.292$$

(Hald explains the details of this rational approximation [**16**].) 'sGravesande did make use of the recursion

$$\binom{n}{x+1} = \binom{n}{x} \frac{n-x}{x+1}$$

suggested by Newton for similar purposes, but even so this is clearly an onerous task. Since the probability of this difference in birth rates recurring 82 years in a row is the extremely small number 0.292^{82}, 'sGravesande drew the conclusion that the higher male birth rates were due to divine intervention.

A few years earlier Jacob Bernoulli had found estimates for binomial sums of the type of (2). These estimates, however, did not involve the exponential function e^x.

De Moivre began his search for such approximations in 1721. In 1733, he proved [**16, 25**] that

$$\binom{n}{\frac{n}{2}+d} \left(\frac{1}{2}\right)^n \approx \frac{2}{\sqrt{2\pi n}} e^{-2d^2/n} \tag{3}$$

and

$$\sum_{|x-n/2|\le d} \binom{n}{x} \left(\frac{1}{2}\right)^n \approx \frac{4}{\sqrt{2\pi}} \int_0^{d/\sqrt{n}} e^{-2y^2} dy. \tag{4}$$

De Moivre also asserted that (4) could be generalized to a similar asymmetrical context, with x varying from $n/2$ to $d + n/2$. This is easily done, with the precision of the approximation clarified by De Moivre's proof.

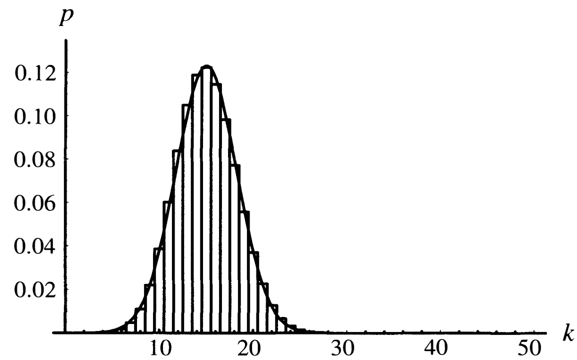


Figure 2 An approximation of binomial probabilities

FIGURE 2 demonstrates how the binomial probabilities associated with 50 independent repetitions of a Bernoulli trial with probability $p = 0.3$ of success are approximated by such a exponential curve. De Moivre's discovery is standard fare in all introductory statistics courses where it is called the normal approximation to the binomial and rephrased as

$$\sum_{i}^{j} \binom{p}{k} p^k (1-p)^{n-k} \approx \mathrm{N}\left(\frac{j-np}{\sqrt{np(1-p)}}\right) - \mathrm{N}\left(\frac{i-np}{\sqrt{np(1-p)}}\right)$$

where

$$\mathrm{N}(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2}\, dx$$

Since this integral is easily evaluated by numerical methods and quite economically described by tables, it does indeed provide a very practical approximation for cumulative binomial probabilities.

The search for an error curve

Astronomy was the first science to call for accurate measurements. Consequently, it was also the first science to be troubled by measurement errors and to face the question of how to proceed in the presence of several distinct observations of the same quantity. In the 2nd century BC, Hipparchus seems to have favored the midrange. Ptolemy, in the 2nd century AD, when faced with several discrepant estimates of the length of a year, may have decided to work with the observation that fit his theory best [**29**]. Towards the end of the 16th century, Tycho Brahe incorporated the repetition of measurements into the methodology of astronomy. Curiously, he failed to specify how these repeated observations should be converted into a single number. Consequently, astronomers devised their own, often ad hoc, methods for extracting a mean, or *data representative*, out of their observations. Sometimes they averaged, sometimes they used the median, sometimes they grouped their data and resorted to both averages and medians. Sometimes they explained their procedures, but often they did not. Consider, for example, the following excerpt, which comes from Kepler [**19**] and reports observations made, in fact, by Brahe himself:

On 1600 January 13/23 at $11^h\ 50^m$ the right ascension of Mars was:

	°	′	″
using the bright foot of Gemini	134	23	39
using Cor Leonis	134	27	37
using Pollux	134	23	18
at $12^h\ 17^m$, using the third in the wing of Virgo	134	29	48
The mean, treating the observations impartially:	134	24	33

Kepler's choice of data representative is baffling. Note that

Average: 134° 26′ 5.5″
Median: 134° 25′ 38″

and it is difficult to believe that an astronomer who recorded angles to the nearest second could fail to notice a discrepancy larger than a minute. The consensus is that the chosen mean could not have been the result of an error but must have been derived by some calculations. The literature contains at least two attempts to reconstruct these calculations [**7**, p. 356], [**35**] but this author finds neither convincing, since both explanations are ad hoc, there is no evidence of either ever having been used elsewhere, and both result in estimates that differ from Kepler's by at least five seconds.

To the extent that they recorded their computation of data representatives, the astronomers of the time seem to be using improvised procedures that had both averages and medians as their components [**7, 29, 30**]. The median versus average controversy lasted for several centuries and now appears to have been resolved in favor of the latter, particularly in scientific circles. As will be seen from the excerpts below, this decision had a strong bearing on the evolution of the normal distribution.

The first scientist to note in print that measurement errors are deserving of a systematic and scientific treatment was Galileo in his famous *Dialogue Concerning the Two Chief Systems of the World—Ptolemaic and Copernican* [**9**], published in 1632. His informal analysis of the properties of random errors inherent in the observations of celestial phenomena is summarized by Stigler [**16**], in five points:

1. There is only one number which gives the distance of the star from the center of the earth, the true distance.
2. All observations are encumbered with errors, due to the observer, the instruments, and the other observational conditions.
3. The observations are distributed symmetrically about the true value; that is the errors are distributed symmetrically about zero.
4. Small errors occur more frequently than large errors.
5. The calculated distance is a function of the direct angular observations such that small adjustments of the observations may result in a large adjustment of the distance.

Unfortunately, Galileo did not address the question of how the true distance should be estimated. He did, however, assert that: " . . . it is plausible that the observers are more likely to have erred little than much . . . " [**9**, p. 308]. It is therefore not unreasonable to attribute to him the belief that the most likely true value is that which minimizes the sum of its deviations from the observed values. (That Galileo believed in the straightforward addition of deviations is supported by his calculations on pp. 307–308 of the *Dialogue* [**9**].) In other words, faced with the observed values $x_1, x_2, \ldots, x_n$, Galileo would probably have agreed that the most likely true value is the x that minimizes the function

$$f(x) = \sum_{n=1}^{n} |x - x_i| \tag{5}$$

As it happens, this minimum is well known to be the median of $x_1, x_2, \ldots, x_n$ and not their average, a fact that Galileo was likely to have found quite interesting.

This is easily demonstrated by an inductive argument [**20**], which is based on the observation that if these values are reindexed so that $x_1 < x_2 < \cdots < x_n$, then

$$\sum_{i=1}^{n} |x - x_i| = \sum_{i=2}^{n-1} |x - x_i| + (x_n - x_1) \quad \text{if } x \in [x_1, x_n],$$

whereas

$$\sum_{i=1}^{n} |x - x_i| > \sum_{i=2}^{n-1} |x - x_i| + (x_n - x_1) \quad \text{if } x \notin [x_1, x_n].$$

It took hundreds of years for the average to assume the near universality that it now possesses and its slow evolution is quite interesting. Circa 1660, we find Robert Boyle, later president of the Royal Society, arguing eloquently against the whole idea of repeated experiments:

> ... experiments ought to be estimated by their value, not their number; ... a single experiment ... may as well deserve an entire treatise. ... As one of those large and orient pearls ... may outvalue a very great number of those little ... pearls, that are to be bought by the ounce ...

In an article that was published posthumously in 1722 [**5**], Roger Cotes made the following suggestion:

> Let p be the place of some object defined by observation, q, r, s, the places of the same object from subsequent observations. Let there also be weights P, Q, R, S reciprocally proportional to the displacements which may arise from the errors in the single observations, and which are given from the given limits of error; and the weights P, Q, R, S are conceived as being placed at p, q, r, s, and their center of gravity Z is found: I say the point Z is the most probable place of the object, and may be safely had for its true place.

Cotes apparently visualized the observations as tokens $x_1, x_2, \ldots, x_n (= p, q, r, s, \ldots)$ with respective physical weights $w_1, w_2, \ldots, w_n$ $(= P, Q, R, S, \ldots)$ lined up on a horizontal axis. FIGURE 3 displays a case where $n = 4$ and all the tokens have equal weight. When Cotes did this, it was natural for him to suggest that the center of gravity Z of this system should be designated to represent the observations. After all, in physics too, a body's entire mass is assumed to be concentrated in its center

Figure 3 A well balanced explanation of the average

of gravity and so it could be said that the totality of the body's points are represented by that single point. That Cotes's proposed center of gravity agrees with the weighted average can be argued as follows. By the definition of the center of gravity, if the axis is pivoted at Z it will balance and hence, by Archimedes's law of the lever,

$$\sum_{i=1}^{n} w_i(Z - x_i) = 0$$

or

$$Z = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \tag{6}$$

Of course, when the weights w_i are all equal, Z becomes the classical average

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

It has been suggested that this is an early appearance of the method of least squares [**26**]. In this context, the method proposes that we represent the data $x_1, x_2, \ldots, x_n$ by the x that minimizes the function

$$g(x) = \sum_{i=1}^{n} w_i(x - x_i)^2 \tag{7}$$

Differentiation with respect to x makes it clear that it is the Z of (6) that provides this minimum. Note that the median minimizes the function $f(x)$ of (5) whereas the (weighted) average minimizes the function $g(x)$ of (7). It is curious that each of the two foremost data representatives can be identified as the minimizer of a nonobvious, though fairly natural, function. It is also frustrating that so little is known about the history of this observation.

Thomas Simpson's paper of 1756 [**36**] is of interest here for two reasons. First comes his opening paragraph:

> It is well known to your Lordship, that the method practiced by astronomers, in order to diminish the errors arising from the imperfections of instruments, and of the organs of sense, by taking the Mean of several observations, has not been generally received, but that some persons, of considerable note, have been of opinion, and even publickly maintained, that one single observation, taken with due care, was as much to be relied on as the Mean of a great number.

Thus, even as late as the mid-18th century doubts persisted about the value of repetition of experiments. More important, however, was Simpson's experimentation with specific *error curves*—probability densities that model the distribution of random errors. In the two propositions, Simpson [**36**] computed the probability that the error in the mean of several observations does not exceed a given bound when the individual errors take on the values

$$-v, \ldots, -3, -2, -1, 0, 1, 2, 3, \ldots, v$$

with probabilities that are proportional to either

$$r^{-v}, \ldots, r^{-3}, r^{-2}, r^{-1}, r^{0}, r^{1}, r^{2}, r^{3}, \ldots, r^{v}$$

or

$$r^{-v}, 2r^{1-v}, 3r^{2-v} \ldots, (v+1)r^0 \ldots, 3r^{v-2}, 2r^{v-1}, r^v$$

Simpson's choice of error curves may seem strange, but they were in all likelihood dictated by the state of the art of probability at that time. For $r = 1$ (the simplest case), these two distributions yield the two top graphs of FIGURE 4. One year later, Simpson, while effectively inventing the notion of a continuous error distribution, dealt with similar problems in the context of the error curves described in the bottom of FIGURE 4 [**37**].

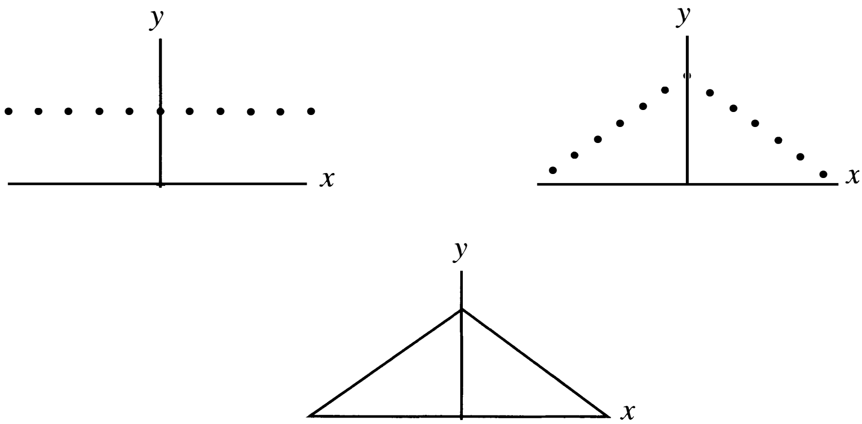


Figure 4 Simpson's error curves

In 1774, Laplace proposed the first of his error curves [**21**]. Denoting this function by $\phi(x)$, he stipulated that it must be symmetric in x and monotone decreasing for $x > 0$. Furthermore, he proposed that

> ... as we have no reason to suppose a different law for the ordinates than for their differences, it follows that we must, subject to the rules of probabilities, suppose the ratio of two infinitely small consecutive differences to be equal to that of the corresponding ordinates. We thus will have
>
> $$\frac{d\phi(x+dx)}{d\phi(x)} = \frac{\phi(x+dx)}{\phi(x)}$$
>
> Therefore
>
> $$\frac{d\phi(x)}{dx} = -m\phi(x).$$
>
> ... Therefore
>
> $$\phi(x) = \frac{m}{2}e^{-m|x|}.$$

Laplace's argument can be paraphrased as follows. Aside from their being symmetrical and descending (for $x > 0$), we know nothing about either $\phi(x)$ or $\phi'(x)$. Hence, presumably by Occam's razor, it must be assumed that they are proportional (the simpler assumption of equality leads to $\phi(x) = Ce^{|x|}$, which is impossible). The resulting differential equation is easily solved and the extracted error curve is displayed in FIGURE 5. There is no indication that Laplace was in any way disturbed by this curve's

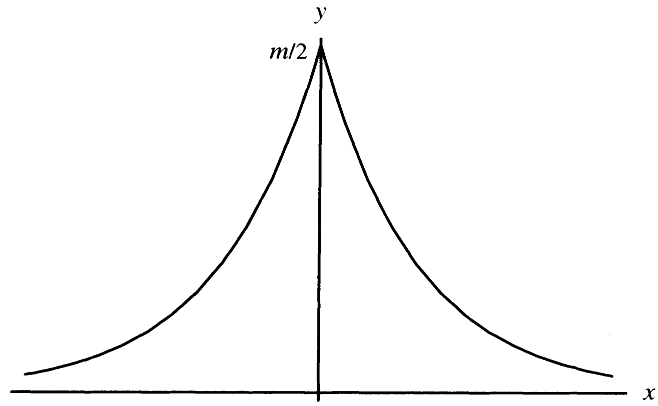


Figure 5 Laplace's first error curve

nondifferentiability at $x = 0$. We are about to see that he was perfectly willing to entertain even more drastic singularities.

Laplace must have been aware of the shortcomings of his rationale, for three short years later he proposed an alternative curve [**23**]. Let a be the supremum of all the possible errors (in the context of a specific experiment) and let n be a positive integer. Choose n points at random within the unit interval, thereby dividing it into $n + 1$ spacings. Order the spacings as:

$$d_1 > d_2 > \cdots > d_{n+1}, \qquad d_1 + d_2 + \cdots + d_{n+1} = 1.$$

Let $\bar{d}_i$ be the expected value of d_i. Draw the points $(i/n, \bar{d}_i)$, $i = 1, 2, \ldots, n + 1$ and let n become infinitely large. The limit configuration is a curve that is proportional to $\ln(a/x)$ on $(0, a]$. Symmetry and the requirement that the total probability must be 1 then yield Laplace's second candidate for the error curve (FIGURE 6):

$$y = \frac{1}{2a} \ln\left(\frac{a}{|x|}\right) \qquad -a \le x \le a.$$

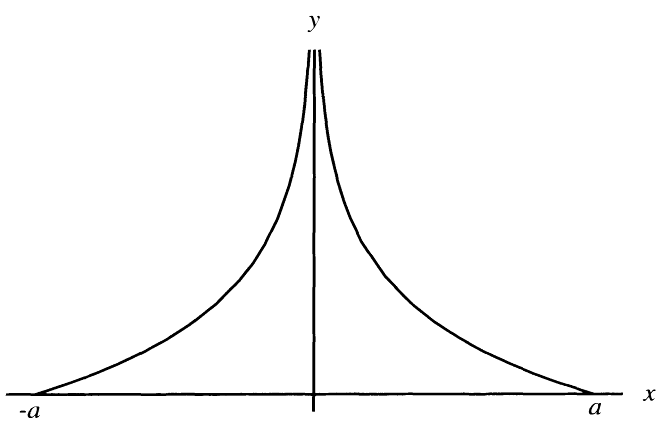


Figure 6 Laplace's second error curve

This curve, with its infinite singularity at 0 and finite domain (a reversal of the properties of the error curve of FIGURE 5 and the bell-shaped curve) constitutes a step backwards in the evolutionary process and one suspects that Laplace was seduced by the considerable mathematical intricacies of the curve's derivation. So much so that he seemed compelled to comment on the curve's excessive complexity and to suggest that error analyses using this curve should be carried out only in "very delicate" investigations, such as the transit of Venus across the sun.

Shortly thereafter, in 1777, Daniel Bernoulli wrote [**2**]:

> Astronomers as a class are men of the most scrupulous sagacity; it is to them therefore that I choose to propound these doubts that I have sometimes entertained about the universally accepted rule for handling several slightly discrepant observations of the same event. By these rules the observations are added together and the sum divided by the number of observations; the quotient is then accepted as the true value of the required quantity, until better and more certain information is obtained. In this way, if the several observations can be considered as having, as it were, the same weight, the center of gravity is accepted as the true position of the object under investigation. This rule agrees with that used in the theory of probability when all errors of observation are considered equally likely.
>
> But is it right to hold that the several observations are of the same weight or moment or equally prone to any and every error? Are errors of some degrees as easy to make as others of as many minutes? Is there everywhere the same probability? Such an assertion would be quite absurd, which is undoubtedly the reason why astronomers prefer to reject completely observations which they judge to be too wide of the truth, while retaining the rest and, indeed, assigning to them the same reliability.

It is interesting to note that Bernoulli acknowledged averaging to be universally accepted. As for the elusive error curve, he took it for granted that it should have a finite domain and he was explicit about the tangent being horizontal at the maximum point and almost vertical near the boundaries of the domain. He suggested the semi-ellipse as such a curve, which, following a scaling argument, he then replaced with a semicircle.

The next important development had its roots in a celestial event that occurred on January 1, 1801. On that day the Italian astronomer Giuseppe Piazzi sighted a heavenly body that he strongly suspected to be a new planet. He announced his discovery and named it Ceres. Unfortunately, six weeks later, before enough observations had been taken to make possible an accurate determination of its orbit, so as to ascertain that it was indeed a planet, Ceres disappeared behind the sun and was not expected to reemerge for nearly a year. Interest in this possibly new planet was widespread and astronomers throughout Europe prepared themselves by compu-guessing the location where Ceres was most likely to reappear. The young Gauss, who had already made a name for himself as an extraordinary mathematician, proposed that an area of the sky be searched that was quite different from those suggested by the other astronomers and he turned out to be right. An article in the 1999 MAGAZINE [**42**] tells the story in detail.

Gauss explained that he used the least squares criterion to locate the orbit that best fit the observations [**12**]. This criterion was justified by a theory of errors that was based on the following three assumptions:

1. Small errors are more likely than large errors.
2. For any real number ϵ the likelihood of errors of magnitudes ϵ and $-\epsilon$ are equal.
3. In the presence of several measurements of the same quantity, the most likely value of the quantity being measured is their average.

On the basis of these assumptions he concluded that the probability density for the error (that is, the error curve) is

$$\phi(x) = \frac{h}{\sqrt{\pi}} e^{-h^2x^2}$$

where h is a positive constant that Gauss thought of as the "precision of the measurement process". We recognize this as the bell curve determined by $\mu = 0$ and $\sigma = 1/\sqrt{2}h$.

Gauss's ingenious derivation of this error curve made use of only some basic probabilistic arguments and standard calculus facts. As it falls within the grasp of undergraduate mathematics majors with a course in calculus based statistics, his proof is presented here with only minor modifications.

The proof

Let p be the true (but unknown) value of the measured quantity, let n independent observations yield the estimates $M_1, M_2, \ldots, M_n$, and let $\phi(x)$ be the probability density function of the random error. Gauss took it for granted that this function is differentiable. Assumption 1 above implies that $\phi(x)$ has a maximum at $x = 0$ whereas Assumption 2 means that $\phi(-x) = \phi(x)$. If we define

$$f(x) = \frac{\phi'(x)}{\phi(x)}$$

then

$$f(-x) = -f(x)$$

Note that $M_i - p$ denotes the error of the ith measurement and consequently, since these measurements (and errors) are assumed to be stochastically independent, it follows that

$$\Omega = \phi(M_1 - p)\phi(M_2 - p)\ldots\phi(M_n - p)$$

is the joint density function for the n errors. Gauss interpreted Assumption 3 as saying, in modern terminology, that

$$\bar{M} = \frac{M_1 + M_2 + \cdots + M_n}{n}$$

is the maximum likelihood estimate of p. In other words, given the measurements $M_1, M_2, \ldots, M_n$, the choice $p = \bar{M}$ maximizes the value of Ω. Hence,

$$\begin{aligned} 0 = \frac{\partial\Omega}{\partial p}_{|p=\bar{M}} &= -\phi'(M_1 - \bar{M})\phi(M_2 - \bar{M})\cdots\phi(M_n - \bar{M}) \\ &\quad - \phi(M_1 - \bar{M})\phi'(M_2 - \bar{M})\cdots\phi(M_n - \bar{M}) - \cdots \\ &\quad - \phi(M_1 - \bar{M})\phi(M_2 - \bar{M})\cdots\phi'(M_n - \bar{M}) \\ &= -\left(\frac{\phi'(M_1 - \bar{M})}{\phi(M_1 - \bar{M})} + \frac{\phi'(M_2 - \bar{M})}{\phi(M_2 - \bar{M})} + \cdots + \frac{\phi'(M_n - \bar{M})}{\phi(M_n - \bar{M})}\right)\Omega. \end{aligned}$$

It follows that

$$f(M_1 - \bar{M}) + f(M_2 - \bar{M}) + \cdots + f(M_n - \bar{M}) = 0. \qquad (8)$$

Recall that the measurements M_i can assume arbitrary values and in particular, if M and N are arbitrary real numbers we may use

$$M_1 = M, \qquad M_2 = M_3 = \cdots = M_n = M - nN$$

for which set of measurements $\bar{M} = M - (n-1)N$.

Substitution into (8) yields

$$f((n-1)N) + (n-1)f(-N) = 0 \quad \text{or} \quad f((n-1)N) = (n-1)f(N).$$

It is a well-known exercise that this homogeneity condition, when combined with the continuity of f, implies that $f(x) = kx$ for some real number k. This yields the differential equation

$$\frac{\phi'(x)}{\phi(x)} = kx.$$

Integration with respect to x produces

$$\ln \phi(x) = \frac{k}{2}x^2 + c \quad \text{or} \quad \phi(x) = Ae^{kx^2/2}.$$

In order for $\phi(x)$ to assume a maximum at $x = 0$, k must be negative and so we may set $k/2 = -h^2$. Finally, since

$$\int_{-\infty}^{\infty} e^{-h^2x^2}\,dx = \frac{\sqrt{\pi}}{h}$$

it follows that

$$\phi(x) = \frac{h}{\sqrt{\pi}} e^{-h^2x^2},$$

which completes the proof. ■

FIGURE 7 displays a histogram of some measurements of the right ascension of Mars [**32**] together with an approximating exponential curve. The fit is certainly striking.

It was noted above that the average is in fact a least squares estimator of the data. This means that Gauss used a particular least squares estimation to justify his theory of errors which in turn was used to justify the general least squares criterion. There is an element of boot-strapping in this reasoning that has left later statisticians dissatisfied and may have had a similar effect on Gauss himself. He returned to the subject twice, twelve and thirty years later, to explain his error curve by means of different chains of reasoning.

Actually, a highly plausible explanation is implicit in the Central Limit Theorem published by Laplace in 1810 [**22**]. Laplace's translated and slightly paraphrased statement is:

> ... if it is assumed that for each observation the positive and negative errors are equally likely, the probability that the mean error of n observations will be

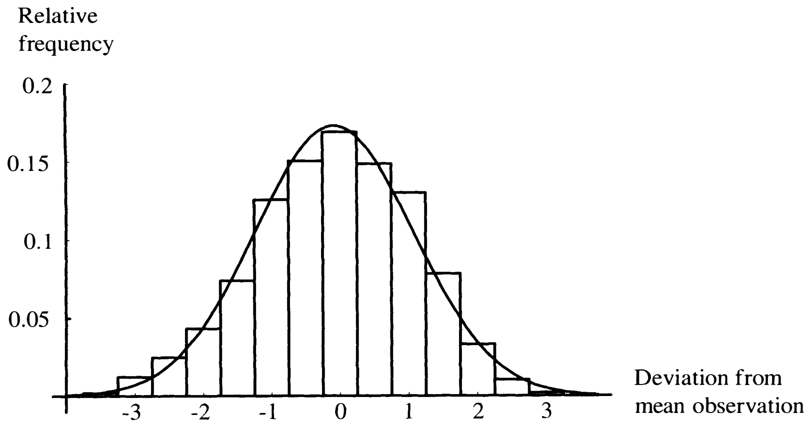


Figure 7 Normally distributed measurements

contained within the bounds $\pm rh/n$, equals

$$\frac{2}{\sqrt{\pi}} \cdot \sqrt{\frac{k}{2k'}} \cdot \int dr \cdot e^{-\frac{k}{2k'} \cdot r^2}$$

where h is the interval within which the errors of each observation can fall. If the probability of error $\pm x$ is designated by $\phi(x/h)$, then k is the integral $\int dx \cdot \phi(x/h)$ evaluated from $x = -\frac{1}{2}h$ to $x = \frac{1}{2}h$, and k' is the integral $\int \frac{x^2}{h^2} \cdot dx \cdot \phi(x/h)$ evaluated in the same interval.

Loosely speaking, Laplace's theorem states that if the error curve of a single observation is symmetric, then the error curve of the sum of several observations is indeed approximated by one of the Gaussian curves of (1). Hence if we take the further step of imagining that the error involved in an individual observation is the aggregate of a large number of "elementary" or "atomic" errors, then this theorem predicts that the random error that occurs in that individual observation is indeed controlled by De Moivre and Gauss's curve (1).

This assumption, promulgated by Hagen [**15**] and Bessel [**4**], became known as the *hypothesis of elementary errors*. A supporting study had already been carried out by Daniel Bernoulli in 1780 [**3**], albeit one of much narrower scope. Assuming a fixed error $\pm\alpha$ for each oscillation of a pendulum clock, Bernoulli concluded that the accumulated error over, say, a day, would be, in modern terminology, approximately normally distributed.

This might be the time to recapitulate the average's rise to the prominence it now enjoys as the estimator of choice. Kepler's treatment of his observations shows that around 1600 there still was no standard procedure for summarizing multiple observations. Around 1660 Boyle still objected to the idea of *combining* several measurements into a single one. Half a century later, Cotes proposed the average as the best estimator. Simpson's article of 1756 indicates that the opponents of the process of averaging, while apparently a minority, had still not given up. Bernoulli's article of 1777 admitted that the custom of averaging had become universal. Finally, some time in the first decade of the 19th century, Gauss assumed the optimality of the average as an axiom for the purpose of determining the distribution of measurement errors.

Beyond errors

The first mathematician to extend the provenance of the normal distribution beyond the distribution of measurement errors was Adolphe Quetelet (1796–1874). He began his career as an astronomer but then moved on to the social sciences. Consequently, he possessed an unusual combination of qualifications that placed him in just the right position for him to be able to make one of the most influential scientific observations of all times.

TABLE 1: Chest measurements of Scottish soldiers

Girth	Frequency
33	3
34	18
35	81
36	185
37	420
38	749
39	1,073
40	1,079
41	934
42	658
43	370
44	92
45	50
46	21
47	4
48	1
	5,738

In his 1846 book *Letters addressed to H. R. H. the grand duke of Saxe Coburg and Gotha, on the Theory of Probabilities as Applied to the Moral and Political Sciences* [**32**, p. 400], Quetelet extracted the contents of Table 1 from the *Edinburgh Medical and Surgical Journal* (1817) and contended that the pattern followed by the variety of its chest measurements was identical with that formed by the type of repeated measurements that are so common in astronomy. In modern terminology, Quetelet claimed that the chest measurements of Table 1 were normally distributed. Readers are left to draw their own conclusions regarding the closeness of the fit attempted in FIGURE 8. The more formal χ^2 normality test yields a χ^2_{Test} value of 47.1, which is much larger than the cutoff value of $\chi^2_{10,.05} = 18.3$, meaning that by modern standards these data cannot be viewed as being normally distributed. (The number of bins was reduced from 16 to 10 because six of them are too small.) This discrepancy indicates that Quetelet's justification of his claim of the normality the chest measurements merits a substantial dose of skepticism. It appears here in translation [**32**, Letter XX]:

> I now ask if it would be exaggerating, to make an even wager, that a person little practiced in measuring the human body would make a mistake of an inch in measuring a chest of more than 40 inches in circumference? Well, admitting this probable error, 5,738 measurements made on one individual would certainly

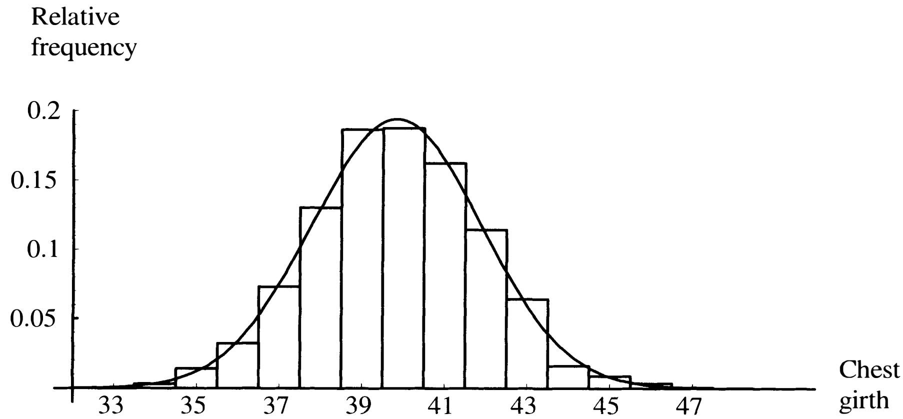


Figure 8 Is this data normally distributed?

not group themselves with more regularity, as to the order of magnitude, than the 5,738 measurements made on the scotch [sic] soldiers; and if the two series were given to us without their being particularly designated, we should be much embarrassed to state which series was taken from 5,738 different soldiers, and which was obtained from one individual with less skill and ruder means of appreciation.

This argument, too, is unconvincing. It would have to be a strange person indeed who could produce results that diverge by 15″ while measuring a chest of girth 40″. Any idiosyncracy or unusual conditions (fatigue, for example), that would produce such unreasonable girths is more than likely to skew the entire measurement process to the point that the data would fail to be normal.

It is interesting to note that Quetelet was the man who coined the phrase *the average man*. In fact, he went so far as to view this mythical being as an ideal form whose various corporeal manifestations were to be construed as measurements that are beset with errors [**34**, p. 99]:

If the average man were completely determined, we might ... consider him as the type of perfection; and everything differing from his proportions or condition, would constitute deformity and disease; everything found dissimilar, not only as regarded proportion or form, but as exceeding the observed limits, would constitute a monstrosity.

Quetelet was quite explicit about the application of this, now discredited, principle to the Scottish soldiers. He takes the liberty of viewing the measurements of the soldiers' chests as a repeated estimation of the chest of the average soldier:

I will astonish you in saying that the experiment has been done. Yes, truly, more than a thousand copies of a statue have been measured, and though I will not assert it to be that of the Gladiator, it differs, in any event, only slightly from it: these copies were even living ones, so that the measures were taken with all possible chances of error, I will add, moreover, that the copies were subject to deformity by a host of accidental causes. One may expect to find here a considerable probable error [**32**, p. 136].

Finally, it should be noted that TABLE 1 contains substantial errors. The original data was split amongst tables for eleven local militias, cross-classified by height and chest girth, with no marginal totals, and Quetelet made numerous mistakes in extracting his data. The actual counts are displayed in TABLE 2 where they are compared to Quetelet's counts.

TABLE 2: Chest measurements of Scottish soldiers

Girth	Actual frequency	Quetelet's frequency
33	3	3
34	19	18
35	81	81
36	189	185
37	409	420
38	753	749
39	1,062	1,073
40	1,082	1,079
41	935	934
42	646	658
43	313	370
44	168	92
45	50	50
46	18	21
47	3	4
48	1	1
	5,738	5,732

Quetelet's book was very favorably reviewed in 1850 by the eminent and eclectic British scientist John F. W. Herschel [**18**]. This extensive review contained the outline of a different derivation of Gauss's error curve, which begins with the following three assumptions:

1. ... the probability of the concurrence of two or more independent simple events, is the product of the probabilities of its constituents considered singly;
2. ... the greater the error the less its probability ...
3. ... errors are equally probable if equal in numerical amount ...

Herschel's third postulate is much stronger than the superficially similar symmetry assumption of Galileo and Gauss. The latter is one-dimensional and is formalized as $\phi(\epsilon) = \phi(-\epsilon)$ whereas the former is multi-dimensional and is formalized as asserting the existence of a function ψ such that

$$\phi(x)\phi(y)\cdots\phi(t) = \psi(x^2 + y^2 + \cdots + t^2).$$

Essentially the same derivation had already been published by the American R. Adrain in 1808 [**1**], prior to the publication of Gauss's paper [**12**] but subsequent to the location of Ceres. In his 1860 paper on the kinetic theory of gases [**24**], the renowned British physicist J. C. Maxwell repeated the same argument and used it, in his words:

> To find the average number of particles whose velocities lie between given limits, after a great number of collisions among a great number of equal particles.

The social sciences were not slow to realize the value of Quetelet's discovery to their respective fields. The American Benjamin A. Gould, the Italians M. L. Bodio and Luigi Perozzo, the Englishman Samuel Brown, and the German Wilhelm Lexis all endorsed it [**31**, p. 109]. Most notable amongst its proponents was the English gentleman and scholar Sir Francis Galton who continued to advocate it over the span of several decades. This aspect of his career began with his 1869 book *Hereditary Genius* [**10**, pp. 22–32] in which he sought to prove that genius runs in families. As he was aware that exceptions to this rule abound, it had to be verified as a statistical, rather than absolute, truth. What was needed was an efficient quantitative tool for describing populations and that was provided by Quetelet whose claims of the wide ranging applicability of Gauss's error curve Galton had encountered and adopted in 1863.

As the description of the precise use that Galton made of the normal curve would take us too far afield, we shall only discuss his explanation for the ubiquity of the normal distribution. In his words [**11**, p. 38]:

> Considering the importance of the results which admit of being derived whenever the law of frequency of error can be shown to apply, I will give some reasons why its applicability is more general than might have been expected from the highly artificial hypotheses upon which the law is based. It will be remembered that these are to the effect that individual errors of observation, or individual differences in objects belonging to the same generic group, are entirely due to the aggregate action of variable influences in different combinations, and that these influences must be
>
> (1) all independent in their effects,
>
> (2) all equal,
>
> (3) all admitting of being treated as simple alternatives "above average" or "below average;"
>
> (4) the usual Tables are calculated on the further supposition that the variable influences are infinitely numerous.

This is, of course, an informal restatement of Laplace's Central Limit Theorem. The same argument had been advanced by Herschel [**18**, p. 30]. Galton was fully aware that conditions (1-4) never actually occur in nature and tried to show that they were unnecessary. His argument, however, was vague and inadequate. Over the past two centuries the Central Limit Theorem has been greatly generalized and a newer version exists, known as Lindeberg's Theorem [**8**], which makes it possible to dispense with requirement (2). Thus, De Moivre's curve (1) emerges as the limiting, observable distribution even when the aggregated "atoms" possess a variety of nonidentical distributions. In general it seems to be commonplace for statisticians to attribute the great success of the normal distribution to these generalized versions of the Central Limit Theorem. Quetelet's belief that all deviations from the mean were to be regarded as errors in a process that seeks to replicate a certain ideal has been relegated to the same dustbin that contains the phlogiston and aether theories.

Why *normal*?

A word must be said about the origin of the term *normal*. Its aptness is attested by the fact that three scientists independently initiated its use to describe the error curve

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

These were the American C. S. Peirce in 1873, the Englishman Sir Francis Galton in 1879, and the German Wilhelm Lexis, also in 1879 [**41**, pp. 407–415]. Its widespread use is probably due to the influence of the great statistician Karl E. Pearson, who had this to say in 1920 [**27**, p. 185]:

> Many years ago [in 1893] I called the Laplace-Gaussian curve the *normal* curve, which name, while it avoids the international question of priority, has the disadvantage of leading people to believe that all other distributions of frequency are in one sense or another *abnormal.*

At first it was customary to refer to Gauss's error curve as the *Gaussian curve* and Pearson, unaware of De Moivre's work, was trying to reserve some of the credit of discovery to Laplace. By the time he realized his error the *normal curve* had become one of the most widely applied of all mathematical tools and nothing could have changed its name. It is quite appropriate that the name of the error curve should be based on a misconception.

Acknowledgment. The author thanks his colleagues James Church and Ben Cobb for their helpfulness, Martha Siegel for her encouragement, and the anonymous referees for their constructive critiques. Much of the information contained in this offering came from the excellent histories [**16, 17, 39, 41**].

REFERENCES

1. R. Adrain, Research concerning the probabilities of the errors which happen in making observations, *Analyst* **1** (1808), 93–109. Reprinted in Stigler, 1980.
2. D. Bernoulli, Dijudicatio maxime probabilis plurium observationam discrepantium atque versimillima inductio inde formanda. *Acta Acad. Sci. Imp. Petrop.*, **1** (1777) 3–23. Reprinted in *Werke* **2** (1982), 361–375. Translated into English by C. G. Allen as "The most probable choice between several discrepant observations and the formation therefrom of the most likely induction," *Biometrica*, **48** (1961) 1–18; reprinted in Pearson and Kendall (1970).
3. ———, Specimen philosophicum de compensationibus horologicis, et veriorimensura temporis. *Acta Acad. Sci. Imp. Petrop.* **2** (1777), 109–128. Reprinted in *Werke* **2** (1982), 376–390.
4. F. W. Bessel, Untersuchungen über die Wahrscheinlichkeit der Beobachtungsfehler, *Astron. Nachr.* **15** (1838), 369–404. Reprinted in *Abhandlungen* **2**.
5. R. Cotes, Aestimatio Errorum in Mixta Mathesi, per Variationes Partium Trianguli Plani et Sphaerici. In *Opera Miscellanea* Cambridge, 1722.
6. F. N. David, *Games, Gods, and Gambling*, New York, Hafner Pub. Co., 1962.
7. C. Eisenhart, The background and evolution of the method of least squares. Unpublished. Distributed to participants of the ISI meeting, 1963. Revised 1974.
8. W. Feller, *An Introduction to Probability Theory and its Applications*, 2 vols, John Wiley & Sons, New York, 1968.
9. G. Galilei, *Dialogue Concerning the Two Chief World Systems—Ptolemaic & Copernican* (S. Drake translator), 2nd ed., Berkeley, Univ. California Press, 1967.
10. F. Galton, *Hereditary Genius: An Inquiry into its Laws and Consequences*. London, Macmillan, first ed. 1863, 2nd ed. 1892.
11. ———, Statistics by intercomparison, with remarks on the law of frequency of error. *Philosophical Magazine*, 4th series, **49** (1875), 33–46.
12. C. F. Gauss, *Theoria Motus Corporum Celestium*. Hamburg, Perthes et Besser, 1809. Translated as *Theory of Motion of the Heavenly Bodies Moving about the Sun in Conic Sections* (trans. C. H. Davis), Boston, Little, Brown 1857. Reprinted: New York, Dover 1963.
13. C. C. Gillispie, Mémoires inédits ou anonymes de Laplace sur la théorie des erreurs, les polynômes de Legendre, et la philosophie des probabilités. *Révue d'histoire des sciences* **32** (1979), 223–279.
14. G. J. 'sGravesande, Démonstration mathématique de la direction de la providence divine. *Oeuvres* **2** (1774), 221–236.

15. G. Hagen, *Grundzüge der Wahrscheinlichskeit-Rechnung*, Dümmler, Berlin, 1837.
16. A. Hald, *A History of Probability and Statistics and Their Applications before 1750.* New York, John Wiley & Sons, 1990.
17. ———, *A History of Mathematical Statistics From 1750 to 1930.* New York, John Wiley & Sons, 1998.
18. J. F. W. Herschel, Quetelet on probabilities, *Edinburgh Rev.* **92**(1850) 1–57.
19. J. Kepler, *New Astronomy* (W. H. Donahue, translator) Cambrdige, Cambridge University Press, 1992.
20. J. Lancaster, private communication, 2004.
21. P. S. Laplace, Mémoire sur la probabilité des causes par les évènements. *Mémoires de l'Academie royale des sciences presentés par divers savan* **6** (1774), 621–656. reprinted in Laplace, 1878-1912, Vol. 8, pp 27–65. Translated in [38].
22. ———, Mémoire sur les approximations des formules qui sont fonctions de très grand nombres et sur leur applications aux probabilités. *Mémoires de l'Academie des sciences de Paris*, 1809, pp. 353–415, 559–565. Reprinted in *Oeuvres complètes de Laplace*, Paris: Gauthier-Villars, vol. 12, pp. 301–353.
23. ———, Recherches sur le milieu qu'il faut choisir entre les résultat de plusieurs obervations. In [13] pp. 228–256.
24. J. C. Maxwell, Illustrations of the dynamical theory of gases. *Phil. Mag.* **19** (1860), 19–32. Reprinted in *The Scientific Papers of James Clerk Maxwell.* Cambridge UK, Cambridge University Press, 1890, and New York, Dover, 1952.
25. A. de Moivre, *Approximatio ad Summam Terminorum Binomii* $(a+b)^n$ *in Seriem Expansi.* Printed for private circulation, 1733.
26. A. de Morgan, Least Squares, *Penny Cyclopaedia.* London, Charles Knight, 1833–1843.
27. E. S. Pearson, and M. Kendall (eds.), *Studies in the history of statistics and probability.* Vol. 1, London, Griffin, 1970.
28. K. Pearson, Notes on the History of Correlation. *Biometrika* **13** (1920) 25–45. Reprinted in [27].
29. R. L. Placket, The principle of the arithmetic mean, *Biometrika* **45** (1958), 130–135. Reprinted in [27].
30. ———, Data Analysis Before 1750, *International Statistical Review*, **56** (1988), No. 2, 181–195.
31. T. M. Porter, *The Rise of statistical Thinking 1820–1900*, Princeton, Princeton University Press, 1986.
32. A. Quetelet, *Lettres à S. A. R. Le Duc Régnant de Saxe Cobourg et Gotha, sur la théorie des probabilités, appliquée aux sciences morales et politique.* Brussels, Hayez, 1846.
33. ———, *Letters addressed to H. R. H. the grand duke of Saxe Coburg and Gotha, on the Theory of Probabilities as Applied to the Moral and Political Sciences*, trans. O. G. Downes. London, Layton, 1849. Translation of [32].
34. ———, *A Treatise on Man and the Development of his Faculties.*, New York, Burt Franklin, 1968.
35. O. B. Sheynin, *The history of the theory of errors.* Deutsche Hochschulschriften [German University Publications], 1118. Hänsel-Hohenhausen, Egelsbach, 1996, 180 pp.
36. T. Simpson, A Letter to the Right Honourable George Macclesfield, President of the Royal Society, on the Advantage of taking the Mean, of a Number of Observations, in practical Astronomy, *Phil. Trans.* **49** (1756), 82–93.
37. ———, *Miscellaneous Tracts on Some Curious, and Very Interesting Subjects in Mechanics, Physical-Astronomy, and Speculative Mathematics.* London, Nourse, 1757.
38. S. M. Stigler, Memoir on the probability of the causes of events. *Statistical Science* **1** (1986) 364–378.
39. ———, *The History of Statistics: The Measurement of Uncertainty before 1900.* Cambridge, The Belknap Press of the Harvard University Press, 1986.
40. ———, *American Contributions to Mathematical Statistics in the Nineteenth Century.* 2 vols., New York, Arno Press, 1980
41. ———, *Statistics on the Table*, Cambridge, Harvard University Press, 1999.
42. D. Teets and K. Whitehead, The discovery of Ceres: how Gauss became famous, this MAGAZINE **72**:2 (1999), 83–93.

Pythagorean Triples and the Problem $A = mP$ for Triangles

LUBOMIR P. MARKOV
Barry University
Miami Shores, FL 33161
lmarkov@mail.barry.edu

Triangles with integer sides and integer area, which today are commonly called *Heronian*, have fascinated mathematicians for centuries. Let us consider the problem of finding all Heronian triangles for which the area, A, is an integer multiple of the perimeter, P:

$$A = mP \quad \text{for } m \in \mathbb{N}. \tag{1}$$

This problem and its variations have a long and distinguished history. For example, as early as 1904, Whitworth and Biddle [**5**, p. 199] proved that there are only five triangles with the property $A = P$, namely the two right triangles (6, 8, 10) and (5, 12, 13), and the three obtuse triangles (6, 25, 29), (7, 15, 20), and (9, 10, 17).

Another interesting relative is the problem $P = \lambda A (\lambda > 0)$ investigated by Subbarao [**9**], who proved that there are no integer-sided triangles such that $A = (1/\lambda)P$ for $\lambda \geq 3$.

In 1985, Goehl [**6**] introduced (1) as stated above and solved it in the special case of right triangles. To reproduce his solution, suppose that a and b are the legs of a right triangle with hypotenuse $c = \sqrt{a^2+b^2}$. Setting the area equal to a multiple of the perimeter gives $ab/2 = m(a + b + \sqrt{a^2+b^2})$, which is readily reduced to the factorization

$$8m^2 = (a - 4m)(b - 4m). \tag{2}$$

Now (2) allows us to determine the sides of the solution triangles according to the following procedure: Find all divisors of $8m^2$; for each divisor δ, equate $a - 4m$ with $\delta_1 = \delta$ and $b - 4m$ with $\delta_2 = 8m^2/\delta$; solve for a and b (to keep $a < b$ and avoid repetitions, take only the divisors that do not exceed $\sqrt{8m^2} = 2\sqrt{2}m$). Clearly, this simple algorithm produces all right triangles solving (1) for a fixed m. The next logical undertaking would be to investigate (1) in general: How can we find *all* Heronian triangles whose area is an integer multiple of the perimeter?

This problem turned out to be significantly more difficult, and to the best of our knowledge has remained open until now; it is the purpose of this paper to solve it. The essence of our argument is the fact that it is possible to reduce (1) to a problem about Pythagorean triples. We will show how the classical area formula of Heron can be rearranged into quantities forming a Pythagorean triple, thereby allowing us to attack the problem via known parametrizations of such triples. As it turns out, solving (1) is equivalent to solving a system of three equations in six unknowns, whose solution (a challenge, by all accounts!) leads to an algorithm for the listing of all Heronian triangles with the property $A = mP$ for any m. It is interesting that a key moment in the proof involves a factorization identity that may be considered a broad generalization of (2).

Pythagorean form of Heron's formula

Recall that (x, y, z) is a *Pythagorean triple* (PT) if its components are nonnegative integers satisfying the relation $x^2 + y^2 = z^2$. Let us agree that the component z will always represent the largest number, whereas the components x and y need not appear in any specific order (for obvious reasons, x and y are often referred to as the *legs* of the PT). The PT (x, y, z) is *primitive* if x, y, and z have no common factor greater than 1; it is obvious that all PTs are obtained from the primitive ones by integer scalar multiplication. The following parametrization of the primitive PTs will be central to our investigation. Beauregard and Suryanarayan [**2**] give a self-contained proof.

PROPOSITION 1. *Depending on whether the first leg x is odd or even, every primitive PT (x, y, z) is uniquely expressed as $(u^2 - v^2, 2uv, u^2 + v^2)$ where u and v are relatively prime of opposite parity, or*

$$\left(\frac{u^2 - v^2}{2}, uv, \frac{u^2 + v^2}{2}\right)$$

where u and v are relatively prime and odd.

It is important to point out that this statement does not say which leg is larger; observe, for instance, how one obtains the primitive PT (5, 12, 13) when $u = 3$ and $v = 2$, and (12, 5, 13) when $u = 5$ and $v = 1$.

The key formula Throughout, denote the sides of a triangle by a, b, c and assume that c is always the largest side. Our key observation here is that Heron's formula

$$4A = \sqrt{(a + b + c)(a + b - c)(a + c - b)(b + c - a)}$$

can be put in the form

$$[c^2 - (a^2 + b^2)]^2 + (4A)^2 = (2ab)^2. \tag{3}$$

This is easily proved upon rewriting (3) as $(4A)^2 = (2ab)^2 - [c^2 - (a^2 + b^2)]^2$ and completely factoring the right-hand side. Simple as its proof may be, we note that (3) is an interesting formula in its own right, as it connects in a Pythagorean relation the area of a right triangle with legs a and b, the area of a general triangle with sides a, b, c, and a quarter of the *Pythagorean difference* $\pm[c^2 - (a^2 + b^2)]$.

The relation (3) allows us to reduce our problem to a problem about Pythagorean triples, in the sense that a triangle with sides a, b, c will be Heronian only if $(\pm[c^2 - (a^2 + b^2)], 4A, 2ab)$ is a PT of even components (for if two components of a PT are even, then the third one must also be even). Proposition 1 then implies that all PTs that solve our problem either have the form $(2n(u^2 - v^2), 4nuv, 2n(u^2 + v^2))$, for $n \in \mathbb{N}$ (with u, v relatively prime of opposite parity), or the form

$$\left(2n\frac{u^2 - v^2}{2}, 2nuv, 2n\frac{u^2 + v^2}{2}\right),$$

for $n \in \mathbb{N}$ (with u, v relatively prime and odd). Combining the two, we conclude that in order to solve our problem, we must consider all PTs of the form $(k(u^2 - v^2), 2kuv, k(u^2 + v^2))$, for $k \in \mathbb{N}$, where $u > v$ are relatively prime positive integers, and equate their components to the respective expressions in (3). Moreover, we may assume without loss of generality that the first leg corresponds to the first quantity in (3); it must be kept in mind, of course, that this quantity could be $c^2 - (a^2 + b^2) > 0$ (the case of an obtuse triangle) or $c^2 - (a^2 + b^2) < 0$ (the case of an acute triangle).

Remembering that the area must equal a multiple of the perimeter, we arrive at our next result.

PROPOSITION 2. *For a fixed m, solving the problem $A = mP$ is equivalent to determining all integers a, b, and c that satisfy the equation*

$$[c^2 - (a^2 + b^2)]^2 + [4m(a + b + c)]^2 = (2ab)^2. \tag{4}$$

This amounts to finding all PTs whose components correspond to the respective quantities in (4); that is, to solving in positive integers the following system of three equations in six unknowns:

$$\begin{cases} \pm[c^2 - (a^2 + b^2)] = k(u^2 - v^2); \\ 4m(a + b + c) = 2kuv; \\ 2ab = k(u^2 + v^2). \end{cases} \tag{5}$$

It turns out that determining all solutions to our problem depends heavily on the quantity $a + b - c$ (cf. Bates [**1**, Thm. 1]), especially on the fact that it is bounded for a fixed m. Readers may recognize this as the *excess* of a Pythagorean triple, defined by McCullough in a recent article [**8**]. The following proposition furnishes the important properties of that quantity:

PROPOSITION 3. *Assume that the triple (a, b, c) solves (1). Then:*

(A) *$a + b - c$ is an even integer;*

(B) *$a + b - c < 4m\sqrt{3}$; and*

(C) *The resulting triangle is obtuse, right, or acute if and only if $a + b - c$ is less than, equal to, or greater than $4m$.*

Proof. To prove (A), we note that the quantities $a + b + c$, $a + b - c$, $a + c - b$, $b + c - a$ are either all even or all odd. Since the product $(a + b + c)(a + b - c)(a + c - b)(b + c - a)$ is even, it follows that all factors must be even. Next, let r be the inradius of the triangle and γ be the largest angle (which means it is opposite the side c). The following formulas are derived by easy trigonometry [**7**]:

$$2A = r(a + b + c), \tan\frac{\gamma}{2} = \frac{2r}{a + b - c}.$$

The first of these implies that $r = 2m$. Since $60° < \gamma < 180°$ and the tangent function increases, we obtain

$$a + b - c = \frac{2r}{\tan(\gamma/2)} < \frac{4m}{\tan 30°} = 4m\sqrt{3}.$$

Finally, the relation $a + b - c = 4m/\tan(\gamma/2)$ also shows that (C) must be true. ■

The main result

We proceed to examine the system (5), beginning with a brief outline of the method. The problem has to be partitioned into the cases of obtuse, acute, and right triangles; we solve the obtuse case first, and then show how the acute case reduces to it. (The subsequent incorporation of the right-triangle case in the solution presents no difficulty.)

One should keep in mind that m is fixed in (5), and the unknowns are u, v, k (from the parametric representation of PTs) and a, b, c (sides of triangles to be determined). We will show that first of all, the us must be subjected to a necessary restriction depending only on m, and that each particular u necessarily implies corresponding values for the vs; only finitely many values are possible for either variable. For each fixed u and v we derive a factorization identity of the form

$$\left(\frac{16m^2}{u}\right)^2 (u^2+v^2) = E_1(k)\cdot E_2(k),$$

where the left-hand side is an integer and the right-hand side is a product of two expressions involving the unknown k. We then prove that to any divisor d of the quantity $(16m^2/u)^2(u^2+v^2)$ there corresponds a value for k, and thus each triple (u, v, d) gives rise to a triangle with sides (a, b, c). Fundamentally, the method aims at identifying all admissible triples (u, v, d).

So, let us first assume that $c^2 > a^2+b^2$ (the case of an obtuse triangle); thus, (5) becomes $c^2-(a^2+b^2)=k(u^2-v^2)$, $4m(a+b+c)=2kuv$, $2ab=k(u^2+v^2)$. Subtracting the third equation from the first and simplifying, we get $(a+b)^2-c^2=2kv^2$, and after factoring the left-hand side and using the second equation we get $a+b-c=4mv/u$. This implies that u must divide $2m$ because $a+b-c$ is even. Combining the last relation with $a+b+c=kuv/(2m)$ and solving the resulting system yields

$$b+a=\frac{ku^2v+8m^2v}{4mu},\qquad c=\frac{ku^2v-8m^2v}{4mu}.$$

On the other hand, upon adding the first and third equations and rearranging terms we obtain $(a-b)^2=c^2-2ku^2$. Let us assume for the time being that $b\ge a$; then we have $b-a=\sqrt{c^2-2ku^2}$, and it is clear that the radicand must be a square.

Rather than give a name to this square, we do some reorganizing. Put $Q=2m/u$ and substitute it in the expressions for c, $b+a$ and $b-a$. After simplification, we get

$$c=\frac{kv-2Q^2v}{2Q},\qquad b+a=\frac{kv+2Q^2v}{2Q},$$
$$b-a=\frac{1}{2Q}\sqrt{(kv-2Q^2v)^2-32km^2},\tag{6}$$

where the quantity under the radical must be a square. Put $(kv-2Q^2v)^2-32km^2=X^2$ and consider this as an equation in the variables X and k. Expand the square and rearrange, obtaining

$$k^2v^2-4k(v^2Q^2+8m^2)+4Q^4v^2=X^2.\tag{7}$$

The last equation is a Diophantine equation which can be solved by factoring. To solve (7), subtract the quantity $(kv-2(v^2Q^2+8m^2)/v)^2$ from both sides, simplify and rearrange terms. The result is

$$[2(v^2Q^2+8m^2)]^2-(2v^2Q^2)^2=(v^2k-2v^2Q^2-16m^2)^2-(Xv)^2.\tag{8}$$

Factor the left-hand side, substitute $Q=2m/u$ and simplify. Factor also the right-hand side and obtain

$$\left(\frac{16m^2}{u}\right)^2(u^2+v^2)=[v^2(k-2Q^2)-16m^2-Xv][v^2(k-2Q^2)-16m^2+Xv].\tag{9}$$

If m is fixed, u is a divisor of $2m$, and $v < u$ is relatively prime to u, we would know the left-hand side of (9). Finding all factors of that side will allow us to find k and X in each case, and hence to determine the values of a, b, c from (6). Thus,

$$a = \frac{kv + 2Q^2v - X}{4Q}, \quad b = \frac{kv + 2Q^2v + X}{4Q}, \quad c = \frac{kv - 2Q^2v}{2Q}. \tag{10}$$

We examine all factors of the left-hand side of (9) matching them with the expressions on the right-hand side as follows. Suppose

$$d_1 = v^2(k - 2Q^2) - 16m^2 - Xv, d_2 = v^2(k - 2Q^2) - 16m^2 + Xv.$$

Solving the last system for k and X yields

$$k = \frac{1}{v^2}\left[\frac{d_1 + d_2}{2} + 16m^2\right] + 2Q^2, X = \frac{d_2 - d_1}{2v}. \tag{11}$$

In (11), we express all variables in terms of $d = d_1$, u, v and obtain

$$k = \frac{(16m^2 + d)[(16m^2 + d)u^2 + 16m^2v^2]}{2du^2v^2},$$
$$X = \frac{(16m^2 + d)(16m^2 - d)u^2 + (16m^2)^2v^2}{2du^2v}.$$

Substituting these in (10) gives us the formulas for the sides of the triangle:

$$\begin{cases} a = \dfrac{(16m^2 + d)u^2 + 16m^2v^2}{8muv}, \\ b = \dfrac{2m(16m^2 + d)(u^2 + v^2)}{duv}, \\ c = \dfrac{(16m^2 + d)^2u^2 + (16m^2)^2v^2}{8mduv}. \end{cases} \tag{12}$$

It is a straightforward calculation to show that in these formulas, we always have $b \le c$ and $a \le c$. Now, let us ensure that $a \le b$. Solving the inequality

$$\frac{(16m^2 + d)u^2 + 16m^2v^2}{8muv} \le \frac{2m(16m^2 + d)(u^2 + v^2)}{duv}$$

yields

$$d \le \frac{16m^2}{u}\sqrt{u^2 + v^2}. \tag{13}$$

Restricting our attention to these divisors d, we will always have $a \le b \le c$.

Next, let us investigate the acute case; this means that the first equation in (5) is now $a^2 + b^2 - c^2 = k(u^2 - v^2)$. Proceeding as in the previous case, we first find that $a + b - c = 4um/v$ (and hence v must divide $2m$), and then obtain the formulas

$$a + b = \frac{kuv^2 + 8m^2u}{4mv}, \quad c = \frac{kuv^2 - 8m^2u}{4mv}.$$

Obviously, these are exactly the formulas from the obtuse case, with u and v interchanged; the rest of the derivation of the formulas for the sides will also hold with u and v interchanged. We conclude that to determine all acute triangles, we must choose d to be the divisor of $(16m^2/v)^2(u^2+v^2)$, and the sides will be given by formulas obtained from (12) by interchanging u and v. Now, interchanging u and v in the acute case is equivalent to allowing $v > u$ in the obtuse case; this can be done because we have $a+b-c = 4vm/u$ and $a+b-c < 4\sqrt{3}m$ which imply the bound $u < v < \sqrt{3}u$. Therefore, the acute case may be examined by taking u to be a divisor of $2m$, choosing v relatively prime to u, $u < v < \sqrt{3}u$, finding divisors d of $(16m^2/u)^2(u^2+v^2)$ and determining the sides from formulas (12).

This time, however, we do not have $b \le c$ for free, so we must ensure it: solving the inequality

$$\frac{2m(16m^2+d)(u^2+v^2)}{duv} \le \frac{(16m^2+d)^2u^2+(16m^2)^2v^2}{8mduv}$$

yields

$$d \ge \frac{16m^2}{u^2}(v^2-u^2), \tag{14}$$

which of course must still be combined with the previous bound (13). These restrictions on d entail no loss of generality.

Finally, all right triangles may be obtained from (12) by formally putting $u = v = 1$; in this case d has to be a divisor of $2(16m^2)^2$ such that $d \le 16m^2\sqrt{2}$, and the equations (12) reduce to

$$\begin{cases} a = \dfrac{32m^2+d}{8m}, \\ b = \dfrac{4m(16m^2+d)}{d}, \\ c = \dfrac{(16m^2+d)^2+(16m^2)^2}{8md}. \end{cases} \tag{15}$$

It is important to mention that formulas (12) produce rational values which may not be integers; discarding the noninteger solutions (this can always be done after finitely many steps) will leave us with the solution to the original problem (1). We summarize what was proved so far as our main result (all bounds have been stated using the greatest integer function $\lfloor\cdot\rfloor$):

THEOREM. *The following algorithm solves the problem $A = mP$:*

1. *For a fixed m, find all divisors u of $2m$.*
2. *For each u, find all v relatively prime to u and such that $1 \le v \le \lfloor\sqrt{3}u\rfloor$.*
3. *To find the obtuse-triangle solutions: Select $v < u$; for every pair u, v find all divisors d of $(16m^2/u)^2(u^2+v^2)$ such that $d \le \lfloor 16m^2/u\sqrt{u^2+v^2}\rfloor$; for every u, v, d, determine the sides a, b, c from the formulas (12).*
4. *To find the acute-triangle solutions: Select $u < v \le \lfloor\sqrt{3}u\rfloor$; for every pair u, v find all divisors d of $(16m^2/u)^2(u^2+v^2)$ such that*

$$\frac{16m^2}{u^2}(v^2-u^2) \le d \le \left\lfloor \frac{16m^2}{u}\sqrt{u^2+v^2} \right\rfloor;$$

for every u, v, d, determine the sides a, b, c from the formulas (12).

5. *To find the right-triangle solutions: Put* $u = v = 1$*; find all divisors* d *of* $2(16m^2)^2$ *such that*

$$d \leq \lfloor 16m^2\sqrt{2} \rfloor;$$

for every d *determine the sides* a, b, c *from the formulas (15).*

6. *Discard the noninteger solutions.*

Example

We shall illustrate our solution with the examination of the case $m = 2$. While we primarily chose it for its moderately-sized output of 18 triangles, this particular value is also quite interesting in that it is the smallest m for which an acute triangle appears, as well as the smallest m that produces an isosceles one. By the way, acute and isosceles triangles seem to be rare among all triangles satisfying $A = mP$; for instance, out of the 80 triangles with the property $A = 7P$, only one is acute and none are isosceles.

So, let $m = 2$ in the algorithm above; then $2m = 4$ and thus u could be 4, 2, or 1. For each u, determine the corresponding vs:

(A) $u = 4 \Rightarrow v = 1, 3;\ 5$

(B) $u = 2 \Rightarrow v = 1;\ 3$

(C) $u = 1 \Rightarrow v = 1$.

The resulting triangles are shown in the table below. The case $u = 4$, $v = 5$ does not produce integer solutions and has been omitted.

TABLE 1: Heronian triangles with the property $A = 2P$

u	v	type of △	d-range	$(16m^2/u)^2(u^2+v^2)$	d	(a, b, c)
4	1	obtuse	$d \leq 65$	$2^8 \cdot 17$	2^2	(18, 289, 305)
					2^3	(19, 153, 170)
					2^4	(21, 85, 104)
					2^5	(25, 51, 74)
					2^6	(33, 34, 65)
4	3	obtuse	$d \leq 80$	$2^8 \cdot 5^2$	2^3	(9, 75, 78)
					2^5	(11, 25, 30)
					$2^2 \cdot 5$	(10, 35, 39)
					$2^4 \cdot 5$	(15, 15, 24)
2	1	obtuse	$d \leq 71$	$2^{10} \cdot 5$	2^3	(11, 90, 97)
					2^4	(12, 50, 58)
					2^5	(14, 30, 40)
					2^6	(18, 20, 34)
					$2^3 \cdot 5$	(15, 26, 37)
2	3	acute	$80 \leq d \leq 115$	$2^{10} \cdot 13$	$2^3 \cdot 13$	(13, 14, 15)
1	1	right	$d \leq 90$	2^{13}	2^4	(9, 40, 41)
					2^5	(10, 24, 26)
					2^6	(12, 16, 20)

Final remarks

It is quite clear that (1) must have solutions for each m (why?). What is not obvious at all before the proof of our main result is that the problem must have finitely many solutions for each m. Therefore, any possible suggestion that a computer can easily find all triangles by inspection should be dismissed; instead, those eager to put the computer to work are invited to follow our algorithm and write a program (for instance, using *Mathematica* or *Maple*) which produces the solution triangles for a fixed m. The equation-loving reader might enjoy verifying the fact that, in the right-triangle case, the factorization (9) reduces to the factorization (2) (showing the equivalence of our solution to Goehl's [**6**]). It will be a very interesting problem to determine all (at least cyclic) quadrilaterals such that their area is an integer multiple of the perimeter. A recent paper in the MAGAZINE [**3**] shows that the problem $A = (1/m)P$ for cyclic quadrilaterals has no solution when $m \geq 5$.

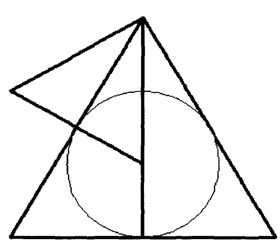
Acknowledgment. The author wishes to thank the referees for their careful reviews of the paper. Special thanks are due to the referee whose suggestions contributed to the paper's significant simplification. The author also wishes to thank Dr. John Goehl who introduced him to the problem, provided him with a database of triangles satisfying $A = mP$ for various m, and reviewed several versions of the paper as it was being prepared for submission.

REFERENCES

1. M. Bates, Serendipity on the area of a triangle, *Math. Teach.* **72** (1979), 273–275.
2. R. Beauregard and E. Suryanarayan, Pythagorean triples: the hyperbolic view, *College Math. J.* **27** (1996), 170–181.
3. R. Beauregard and K. Zelator, Perfect cyclic quadrilaterals, this MAGAZINE **75** (2002), 138–143.
4. H. Blichfeldt, On triangles with rational sides and having rational areas, *Ann. of Math.* **11** (1896–7), 57–60.
5. L. Dickson, *History of the Theory of Numbers*, Vol. II, Chelsea Publishing Co., New York, 1992 (reprint from the 1923 edition).
6. J. Goehl, Area = k(perimeter), *Math. Teach.* **76** (1985), 330–332.
7. L. Kells, W. Kern, and J. Bland, *Plane and Spherical Trigonometry*, 2nd ed., McGraw-Hill Book Co., New York & London, 1940.
8. D. McCullough, Height and excess of Pythagorean triples, this MAGAZINE, **78**:1 (2005), 26–44.
9. M. Subbarao, Perfect Triangles, *Amer. Math. Monthly* **72** (1971), 384–385.

Proof Without Words: Inradius of an Equilateral Triangle

The inradius of an equilateral triangle is one-third the height of the triangle.



—Participants of the Summer Institute Series 2004 Geometry Course
School of Education, Northeastern University
Boston, MA 02115

Euler's Ratio-Sum Theorem and Generalizations

BRANKO GRÜNBAUM
University of Washington
Seattle, WA 98195-4350
grunbaum@math.washington.edu

MURRAY S. KLAMKIN*
University of Alberta
Edmonton, Alberta, T6G 2G1
Canada

Over the centuries, many papers have been written about relations among different parts of a triangle, by well-known mathematicians as well as others. The main aim of this note is to show how asking the right questions can lead to new facts and to far-reaching generalizations that retain an elementary nature and would have been understandable to mathematicians of ages past.

Our starting point is one of the results of Euler's paper [5], which shows, in the notation of FIGURE 1, that

$$QB_1/A_1B_1 + QB_2/A_2B_2 + QB_3/A_3B_3 = 1, \qquad (*)$$

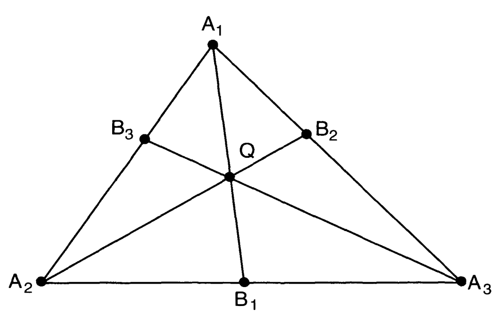


Figure 1 An example illustrating the notation used in Euler's theorem and in Theorem 1

where Q is an arbitrary point in the plane of the arbitrary triangle $A_1A_2A_3$ and B_i is the intersection point of the cevian line A_iQ with the side opposite A_i. Here and throughout, the only restriction is that all the points are well-defined and all the lengths appearing in the denominators are not zero. The lengths are understood as signed lengths; since only ratios of collinear segments are considered, the positive direction on the lines carrying the segments is irrelevant. Euler gives several proofs that use various elementary geometric or trigonometric arguments. We shall provide a simple proof, and show how this result can be generalized in a variety of ways: to analogs of triangles in

*Professor Klamkin passed away in the summer of 2004. As a friend and a mathematician he will be sorely missed by many of us. Professor Klamkin was still able to see the referees' comments on our paper and approve the proposed final version of it. A variety of unfortunate circumstances delayed the sending of that version to the Editor. But this had the silver lining contained in part (vii) of the last section, added September 15, 2005. BG three and higher dimensions, to polygons with more than three sides and their higher-dimensional analogs, and to other ratio-sums.

We shall first discuss the version of Euler's result that holds for d-dimensional simplices, that is, the simplest polytopes of dimension d—the analogs of the triangles in the plane and tetrahedra in 3-space. We shall tie this with a presentation of similar results for the five other ratio-sums that can be defined using cevians; so far, these seem to have received scant attention. In contrast to Euler's result, formulas with the five other ratio-sums involve the dimension d; in some cases, the formula takes the form of an inequality, with the case of equality precisely identified. The generalizations of these results to more general polygons, polyhedra, and polytopes (higher dimensional relatives of polygons and polyhedra) will then be presented, followed by historical and other comments.

Ratio-sums for simplices For $d \geq 2$, let T^d denote the d-dimensional simplex in Euclidean d-space E^d, with vertices A_i, $0 \leq i \leq d$. Thus T^d can be interpreted as the convex hull of $d+1$ points of the Euclidean d-space E^d that are not all contained in a hyperplane of smaller dimension. A good model to illustrate the concept and the following arguments is given by the d-simplex with vertices at the origin and at the unit points of a standard basis of E^d. For notational convenience we shall occasionally use $A_{d+1} = A_0$.

We start by describing the setting of the results. Let Q be a point of E^d, and F_i the facet (that is, the $(d-1)$-dimensional face) of T^d that is opposite A_i. Let B_i be the point of intersection of the line (the *cevian*) through A_i and Q with the hyperplane H_i that contains F_i. For the defintions, and some of the results, the point Q need not be in the interior of T^d; the only overall restriction on Q is that all points B_i must be well defined, and that the denominators in the various fractions be nonzero. This condition will be assumed throughout, and will not be repeated in the reformulation of our results. We shall be interested in various ratios involving the lengths $a_i = \|A_i - Q\|$, $b_i = \|Q - B_i\|$, $q_i = \|A_i - B_i\|$ of the segments A_iQ, QB_i, A_iB_i. As already mentioned, the lengths in question are to be taken as signed lengths; since we shall always consider ratios of collinear segments, the scale of measurement and the direction chosen as positive on each line are irrelevant. For $d = 2$, one illustration of the possibilities is indicated in FIGURE 1, and another in FIGURE 2.

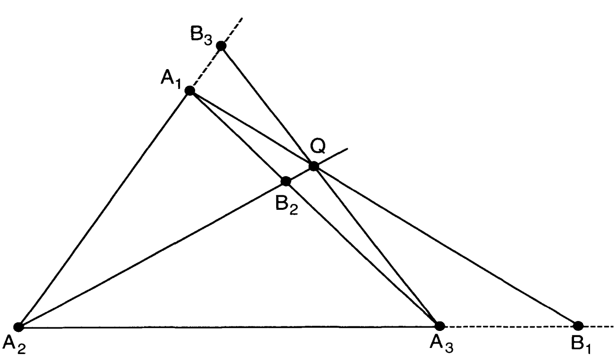


Figure 2 Another illustration of parts (i), (ii), (iv) and (vi) of Theorem 1. Parts (iii) and (v) are not applicable to this example since $q_2/b_2 < 0$.

To begin with, we are interested in the six ratio sums defined as follows:

$$\rho(b,q) = \Sigma_i b_i/q_i, \qquad \rho(a,q) = \Sigma_i a_i/q_i,$$
$$\rho(q,b) = \Sigma_i q_i/b_i, \qquad \rho(q,a) = \Sigma_i q_i/a_i,$$
$$\rho(a,b) = \Sigma_i a_i/b_i, \qquad \text{and } \rho(b,a) = \Sigma_i b_i/a_i,$$

where each sum is over all i, $0 \le i \le d$. We shall prove the following results.

THEOREM 1. *With the above notation the following statements are valid for all Q:*

(i) $\rho(b,q) = 1$.
(ii) $\rho(a,q) = d$.
(iii) *If $q_i/b_i > 0$ for all i, then $\rho(q,b) \ge (d+1)^2$; equality holds if and only if Q is the centroid of T^d.*
(iv) *If $q_i/a_i > 0$ for all i, then $\rho(q,a) \ge (d+1)^2/d$; equality holds if and only if Q is the centroid of T^d.*
(v) *If $q_i/b_i > 0$ for all i, then $\rho(a,b) \ge d(d+1)$; equality holds if and only if Q is the centroid of T^d.*
(vi) *If $q_i/a_i > 0$ for all i, then $\rho(b,a) \ge (d+1)/d$; equality holds if and only if Q is the centroid of T^d.*

Before we turn to the proofs, we recall two very useful lemmas.

Let T^d be a d-simplex with vertices A_i, $0 \le i \le d$, and let S^d denote the simplex with vertices $Q, A_1, A_2, \ldots, A_d$. Then, denoting by $V(T)$ the signed volume of the simplex T, we have:

LEMMA 1. $\|Q - B_0\|/\|A_0 - B_0\| = b_0/q_0 = V(S^d)/V(T^d)$.

This self-evident fact, which was called the "volume principle" in [**11**], has been used without a special name by other authors (see, for example, [**3**, p. 131] for $d = 3$). In case $d = 2$, it has been called the "area principle" in [**10**] and other publications, and it has been used starting at least two centuries ago.

A second well-known tool is the elementary

LEMMA 2. *For all $x > 0$, $x + 1/x \ge 2$, with equality if and only if $x = 1$.*

We shall frequently apply this lemma in the form $1/x \ge 2 - x$.

Proof of Theorem 1. We shall give here proofs for only the first three parts of Theorem 1, to serve as warm-up for the generalizations presented in Theorem 2.

The result of part (i) follows at once from Lemma 1, upon noticing that the signed volumes of the simplices with common apex Q that are spanned by the $d+1$ facets of T^d add up precisely to the signed volume of T^d. For part (ii) it is enough to note that

$$\rho(a,q) = \sum_i a_i/q_i = \sum_i (1 - b_i/q_i)$$
$$= (d+1) - \sum_i b_i/q_i = d + 1 - \rho(b,q) = d + 1 - 1 = d.$$

For part (iii), using Lemma 2, we have

$$\rho(q,b)/(d+1) = \sum_i q_i/((d+1)b_i) \ge 2(d+1) - \sum_i ((d+1)b_i)/q_i$$
$$= 2(d+1) - (d+1)\rho(b,q) = d+1,$$

which is equivalent to the inequality of (iii). Equality holds if and only if $((d+1)b_i)/q_i = 1$ for every i, which is a characterization of Q as the centroid of T^d. ∎

Further generalizations The role that the simplex T^d plays in the above theorems will become clearer as we move to generalize Theorem 1. It is convenient to introduce appropriate notation.

Let P denote a fixed polytope of dimension d in Euclidean d-space E^d. It is simplest to think of P as a convex polytope, that is, as a generalization of convex polygons in the plane or convex solids in 3-space, which one could call polytopes of dimension $d = 2$ or 3—the reader is welcome to use these to picture the generalizations. However, the restriction to convex polytopes is in no way necessary. For $d = 2$ and $d = 3$, we can admit polygons and polyhedra in the generality described in [**8**] and [**9**], that is, self-intersecting polygons, and self-intersecting polyhedra with possibly self-intersecting faces. For $d \geq 4$ we admit the obvious generalizations of these kinds of polygons and polyhedra. We shall use the term *polytope* for all dimensions $d \geq 2$.

We impose the following restrictions on the polytopes considered here: The polytopes must be orientable, and the d-polytopes and all their facets must have nonzero content (volume in dimension d or $d-1$, respectively). The content of a d-polytope P will be denoted by $V(P)$. The d-pyramid determined by a $(d-1)$-polytope F and point X will be denoted $F(X)$.

Polytopes satisfying these conditions shall be called *star-like*. The traditional Kepler-Poinsot polyhedra—that is, the nonconvex analogs of the Platonic regular solids—are star-like both in our sense and visually. So are many (but not all) of the uniform polyhedra presented in [**4**], and beautifully illustrated by photos of models in [**19**]. Many other examples appear in [**8**] and [**9**].

Let P be a star-like d-polytope. The f facets of P are labeled $F_1, F_2, \dots, F_f$ in an arbitrary order. Let A_j, $1 \leq j \leq f$, be a collection of points of E^d such that for suitable points B_j, with B_j in the hyperplane determined by F_j, the line $L_j = A_jB_j$ is well defined, intersects F_j only in B_j, and all lines L_j pass through a common point Q.

In analogy to the notation for Theorem 1, we put:

$$\rho(b, q; W) = \Sigma_j w_j b_j / q_j, \qquad \rho(q, b; W) = \Sigma_j q_j/(w_j b_j),$$

$$\rho(a, q; W) = \Sigma_j w_j a_j / q_j, \qquad \rho(q, a; W) = \Sigma_j q_j/(w_j a_j),$$

$$\rho(a, b; W) = \Sigma_j w_j a_j/(w_j b_j), \qquad \text{and } \rho(b, a; W) = \Sigma_j b_j/(w_j a_j),$$

where $W = (w_1, w_2, \dots, w_f)$ is an ordered f-tuple of suitable weights specified below; these weights depend on P and the points A_j, but are independent of Q. All summations are for $j = 1, 2, \dots, f$. In all parts of Theorem 2 it is understood that P is a star-like d-polytope with f facets, the points Q and A_j satisfy the above condition, and the weights W are given by $w_j = V(F_j(A_j))/V(P)$. We abbreviate $w = \Sigma_j w_j$ and $w^* = \Sigma_j 1/w_j$. FIGURE 3 illustrates the notation.

THEOREM 2. *With the above notation the following statements are valid:*

(i) $\rho(b, q; W) = 1$.

(ii) $\rho(a, q; W) = w - 1$.

(iii) *If* $q_j/b_j > 0$ *and* $w_j > 0$ *for all* j, *then* $\rho(q, b; W) \geq w(2f - w)$. *Equality holds if and only if* $q_j/(w_j b_j) = w$ *for all* $j = 1, 2, \dots, f$.

(iv) *If* $q_j/a_j > 0$ *and* $w_j > 0$ *for all* j, *then* $\rho(q, a; W) \geq f^2/(w-1)$, *with equality if and only if* $q_j/(w_j a_j) = f/(w-1)$ *for all* $j = 1, 2, \dots, f$.

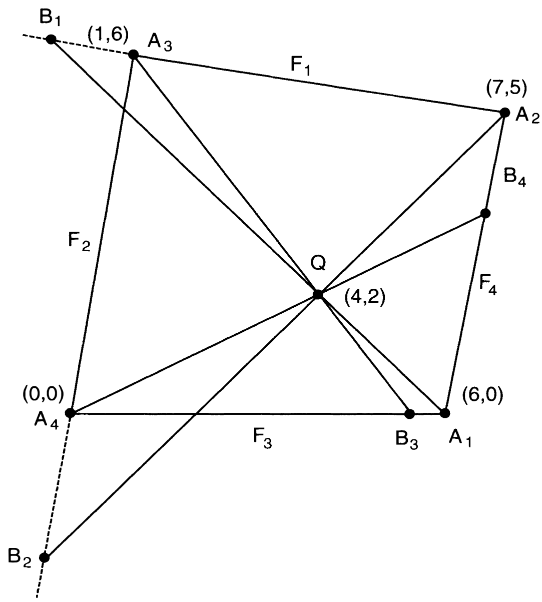


Figure 3 *An example of the situation covered by Theorem 2. As is easily verified, in this example the weights are given by* $W = (31/67, 37/67, 36/67, 30/67)$, *hence* $w = 2$

(v) *If* $q_j/b_j > 0$ *and* $w_j > 0$ *for all* j, *then* $\rho(a, b; W) \geq w(2f - w) - w^*$, *with equality if and only if* $q_j/(w_j b_j) = w$ *for all* $j = 1, 2, \dots, f$.

(vi) *If* $q_j/a_j > 0$ *and* $w_j > 0$ *for all* j, *then* $\rho(b, a; W) \geq f^2/(w - 1) - w^*$, *with equality if and only if* $q_j/(w_j a_j) = f/(w - 1)$ *for all* $j = 1, 2, \dots, f$.

Proof of Theorem 2. For part (i), we note that an easy generalization of what we called the "volume principle" shows that $b_j/q_j = V(F_j(Q))/V(F_j(A_j))$. Hence we have:

$$\begin{aligned}\rho(b, q; W) &= \Sigma_j w_j b_j/q_j = \Sigma_j w_j V(F_j(Q))/V(F_j(A_j)) \\ &= \Sigma_j V(F_j(Q))/V(P) = 1,\end{aligned}$$

since the sum of the volumes of the pyramids with apex Q equals the volume of P.

For part (ii), in analogy to the above and using Theorem 2(i), we have

$$\begin{aligned}\rho(A, q; W) &= \sum_j w_j a_j/q_j = \sum_j w_j(1 - b_j/q_j) \\ &= \sum_j w_j - \sum_j w_j V(F_j(Q))/V(F_j(A_j)) \\ &= \left(\sum_j w_j\right) - 1 = w - 1.\end{aligned}$$

For part (iii), in analogy to the proof of Theorem 1(iii), we have

$$\rho(q, b; W)/w = \sum_j q_j/(w w_j b_j) \geq \sum_j (2 - w w_j b_j/q_j) = 2f - w.$$

The equality criterion follows from Lemma 2.

For part (iv) we have

$$\begin{aligned}(w-1)\rho(q,a;W)/f &= \sum_j (w-1)q_j/(fw_ja_j)\\ &\geq \sum_j (2 - fw_ja_j/((w-1)q_j))\\ &= \sum_j 2 - (f\sum_j w_ja_j/q_j)/(w-1)\\ &= 2f - f(w-1)/(w-1) = f,\end{aligned}$$

which is equivalent to the claim. The equality condition is again a consequence of Lemma 2.

For part (v), in analogy to the above, and using Lemma 2 and part (i) of the theorem, we have

$$\begin{aligned}\rho(a,b;W) &= \sum_j a_j/(w_jb_j) = \sum_j q_j/(w_jb_j) - \sum_j 1/w_j\\ &= w\sum_j q_j/(ww_jb_j) - \sum_j 1/w_j\\ &\geq w\sum_j (2 - ww_jb_j/q_j) - w^* = 2wf - w^2\sum_j w_jb_j/q_j - w^*\\ &= 2wf - w^2 - w^*.\end{aligned}$$

Equality holds if and only if it holds in part (iii) of the theorem.

For part (vi), in analogy to the above, and using part (iv) of the theorem:

$$\begin{aligned}\rho(b,a;W) &= \sum_j b_j/(w_ja_j) = \sum_j q_j/(w_ja_j) - \sum_j 1/w_j\\ &= (f/(w-1))\sum_j (w-1)q_j/(fw_ja_j) - w^*\\ &\geq (f/(w-1))\left(\sum_j 2 - \sum_j fw_ja_j/((w-1)q_j)\right) - w^*\\ &= 2f^2/(w-1) - (f^2/(w-1)^2)\sum_j w_ja_j/q_j - w^*\\ &= 2f^2/(w-1) - (f^2/(w-1)^2)(w-1) - w^* = f^2/(w-1) - w^*,\end{aligned}$$

with equality if and only if equality holds in part (iv).

This completes the proof of all parts of Theorem 2. ■

Historical and other comments

(i) For $d = 2$, Theorem 1(i) contains Euler's result. Euler's theorem has been rediscovered by several authors; first among them is Gergonne [**6**]. Very few of these mention Euler—even the authoritative work of Zacharias [**21**] mentions only Gergonne. Surprisingly, the detailed survey of pre-20th century geometry by Simon [**18**] (which has references to well over 2000 authors!) does not mention the result at all. Without any attribution, Euler's result appears in [**1**, p. 162]. The extension to higher dimensions is also not new. For $d = 3$ the earliest mention we are aware of is by Gergonne [**6**]. Parts (i) and (ii) of Theorems 1 appear in texts [**2**, page 115] and [**3**, page 131]. For general d, our Theorem 1(i) appears in [**13**] and probably in several other places; it was also mentioned in a letter from Prof. H. Gülicher in 1998.

(ii) We are not aware of any mention of parts (iii) to (vi) of Theorem 1 in the literature. The fact that equality holds in these cases for Q at the centroid is obvious. In each of them, the characterization of Q as the centroid in case of equality is due to Klamkin [**14**].

(iii) It is easy to verify that the results of Theorem 2 reduce to those of Theorem 1 in the special case that P is the d-simplex T^d and the points A_i are the vertices of T^d. However, even if P is T^d the results of Theorem 2 are more general since they do not restrict the points A_i to be vertices of T^d.

(iv) Parts (i) and (ii) of Theorems 1 and 2 are somewhat analogous to the classical theorems of Ceva and Menelaus, and the new results on self-transversality (see [**11**]). These earlier results deal with products of ratios, while here we are concerned with sums of ratios. However, our other results seem not to have any multiplicative analogs.

(v) The ratios a_j/b_j for cevians of a triangle appear in Euler's paper [**5**], in the following result, formulated in the notation of Figure 1:

$$\begin{aligned} &A_1Q/QB_1 + A_2Q/QB_2 + A_3Q/QB_3 + 2 \\ &\quad = (A_1Q/QB_1) \cdot (A_2Q/QB_2) \cdot (A_2Q/QB_2). \end{aligned} \qquad (**)$$

This nonlinear relation seems to have been largely forgotten. it has been established in a simple way and its validity extended in the recent paper [**17**]. An analog of this result is due to Euler [**5**]; it deals with ratios of lengths of the segments into which each side of a triangle is partitioned by parallels to the other sides. It was independently found by Gülicher [**12**].

(vi) The idea to use weights attached to the ratios originated with Shephard [**16**]; he kindly sent a preprint of this paper to one of us. For polygons in the plane Shephard establishes in [**16**] a restricted version of part (i) of our Theorem 2. Weights were also assigned to ratios in [**7**], for ratio-sums of a slightly different kind.

(vii) Part (iii) of Theorem 1, as well as some results found in the literature, can be generalized so that, instead of cevians, we use arbitrary segments, which need not have a common point. Let T^d, A_i, Q, F_i, and H_i have the same meaning as in the discussion leading to Theorem 1. Let B_i be an arbitrary point of H_i, and let C_i denote the point of H_i such that the segment QC_i is parallel to the segment A_iB_i. FIGURE 4 illustrates a case with $d = 2$. Let q_i and c_i denote the signed lengths of the segments A_iB_i and

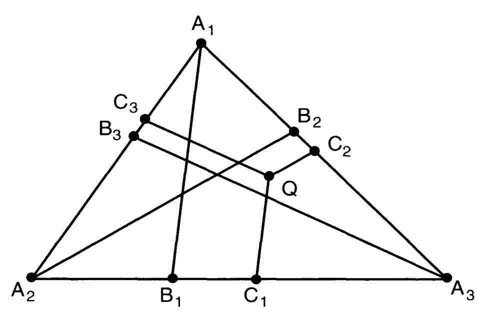


Figure 4 An illustration of the content and notation of comment (vii)

QC_i, let f and f_i denote the d-dimensional volumes of T^d and of the simplices with basis F_i, and let u and v be nonnegative reals. Using the obvious generalization of the volume principle we have

$$\sum_i q_i/(uc_i + vq_i) = \sum_i 1/(u(c_i/q_i) + v) = \sum_i 1/(u(f_i/f) + v)$$
$$\geq (d+1)^2 / \sum_i (u(f_i/f) + v)$$
$$= (d+1)^2/(u + v(d+1)).$$

Here we used the fact that the arithmetic mean is greater than or equal to the harmonic mean, and that $\sum_i f_i/f = 1$. One could also add the less interesting generalization of part (i) of Theorem 1, namely

$$\sum_i \frac{uc_i + vq_i}{q_i} = u + v(d+1).$$

The special case $d = 3$, with B_i the foot of the altitude from A_i, and Q the incenter appears in [**20**] and [**15**]. In the former, $u = 3$ and $v = 1$, so that the lower bound is $16/7$; Murray Klamkin was one of the solvers of [**20**]. In [**15**], $v = 0$ and $u = 1$, hence the lower bound is 16.

Acknowledgment. The authors are indebted to the referees for suggestions that greatly improved the presentation of the present paper.

REFERENCES

1. N. Altshiller-Court, *College Geometry*, 2nd ed., Barnes & Noble, New York, 1952.
2. N. Altshiller-Court, *Modern Pure Solid Geometry*, Macmillan, New York, 1935.
3. N. Altshiller-Court, *Modern Pure Solid Geometry*, 2nd ed., Chelsea, New York, 1964.
4. H. S. M. Coxeter, M. S. Longuet-Higgins and J. C. P. Miller, Uniform polyhedra, *Philos. Trans. Roy. Soc. London* (A) **246** (1953/54), 401–450.
5. L. Euler, Geometrica et sphaerica quedam, *Mémoirs de l'Académie des Sciences de St-Pétersburg* **5** (1812), pp. 96–114. Reprinted in Leonhardi Euleri Commentationes Geometricae Vol. 1 (=Leonhardi Euleri Opera Omnia, Ser. Prima, Opera Mathematica, Vol. 26), A. Speiser, ed., Orell Füssli Publ., Lausanne 1953, pp. 344–358; comments (in German) by the editor, *ibid.* pp. XXXIV–XXXV.
6. J. D. Gergonne, *Annales de mathématiques* **9** (1818–1819), 277. (Not available to the authors; quoted according to Zacharias [**21**], pages 988 and 1057.)
7. B. Grünbaum, Cyclic ratio sums and products, *Crux Math.* **24** (1998), 20–25.
8. B. Grünbaum, Are your polyhedra the same as my polyhedra? *Discrete and Computational Geometry: The Goodman-Pollack Festschrift*, B. Aronov, S. Basu, J. Pach, and M. Sharir, eds., Springer, New York, 2003, pp. 461–488.
9. B. Grünbaum, "New" uniform polyhedra, *Discrete Geometry: In Honor of W. Kuperberg's 60th Birthday*, Monographs and Textbooks, Marcel Dekker, New York 2003, pp. 331–350.
10. B. Grünbaum and G. C. Shephard, Ceva, Menelaus, and the Area Principle, this MAGAZINE **68** (1997), 179–192.
11. B. Grünbaum and G. C. Shephard, Ceva, Menelaus, and Selftransversality, *Geometriae Dedicata* **65** (1997), 179–192.
12. H. Gülicher, Problem E3231, *Amer. Math. Monthly* **94** (1987), 876; solution, ibid. **96** (1989), 449–451.
13. M.S. Klamkin, Relations among segments of concurrent cevians of a simplex, *Crux Math.* **13** (1987), 274–275.
14. M. S. Klamkin, A centroidal equality, Solution of Problem 1630, this MAGAZINE **75** (2002), 320–321.
15. F. Leuenberger, Extremaleigenschaften des Summe der wichtigsten Ecktransversalen des n-dimensionalen Simplex, *Elem. Math.* **15** (1960), 81–82.
16. G. C. Shephard, Cyclic sums for polygons, this MAGAZINE **72** (1999), 126–132.
17. G. C. Shephard, Euler's triangle theorem. *Crux Math.* **25** (1999), 148–153.

18. M. Simon, Über die Entwicklung der Elementar-Geometrie im XIX Jahrhundert, *Jahresberichte der Deutschen Mathematiker-Vereinigung*, Ergänzungsband 1, Teubner, Leipzig 1906.
19. M. J. Wenninger, *Polyhedron Models*, Cambridge University Press, 1971.
20. Z. Yun, Problem 1699, A tetrahedron inequality, this MAGAZINE **77** (2004); solution *ibid.* **78** (2005), 242–243.
21. M. Zacharias, Elementargeometrie und elementare nicht-Euclideische Geometrie in synthetischer Behandlung, *Encyklopädie der mathematischen Wissenschaften, Geometrie*, **1** (1914), 859–962.

To appear in *The College Mathematics Journal* May 2006

Articles:

Complementary Coffee Cups *by Thomas Banchoff*

Folding Beauties *by Leah Berman*

Synchronizing Fireflies *by Ying Zhou, Walter Gall, and Karen Nabb*

From Chebyshev to Bernstein: A Tour of Polynomials Large and Small *by Matthew Boelkins, Jennifer Miller, and Benjamin Vugteveen*

The Tippy Trough *by Donald Francis Young*

Classroom Capsules

Revised: $\arctan 1 + \arctan 2 + \arctan 3 = \pi$ *by Michael W. Ecker*

Streaks and Generalized Fibonacci Sequences *by Shahla Ahdout, Sheldon Rothman, and Helen Strassberg*

A Bug Problem *by Aaron Melman*

The Non-Attacking Queens Game *by Hassan Noon and Glen Van Brummelen*

The Existence of Multiplicative Inverses *by Ricardo Alfaro and Steven Althoen*

Cover Image: Pentadecagon Dissection *by Ernest Irving Freese*

Ernest Irving Freese was an architect from Los Angeles who compiled a manuscript of clever geometric dissections such as this one that cuts a pentadecagon into 31 pieces that can be reassembled into two pentadecagons of area equal to half of the original area. This was likely found shortly before his death in 1957. His extensive manuscript was never published.

In the article on page 87 of this issue, Greg Frederickson reports on the charming constructions of Freese and others, and finds a general dissection using fewer pieces than these earlier works.

NOTES

Do You Know Your Relative Driving Speed?

MARK F. SCHILLING
California State University at Northridge
Northridge, CA 91330
mark.schilling@csun.edu

Drivers on a busy highway or freeway typically select a driving speed based not only on the posted speed limit, but also on the velocities of nearby vehicles. Many drivers attempt to stay at or near the "flow of traffic," while others prefer to go a bit faster or slower. Traffic safety engineers have stated that the safest speed to travel on a busy freeway is at the 85th percentile of traffic speeds. A natural question arises: How can individuals gauge their speed percentiles from observing traffic in the vicinity?

In this note, I utilize a simple idealized model for traffic flow: Assume that each vehicle travels at a constant speed and that the locations of vehicles and their speeds are described by what is called a *marked Poisson process*. This means that vehicles are randomly spaced along the highway, and that their speeds are independent of their locations and of all other vehicles' speeds and locations. Assume also that the distribution of traffic speeds (the "marks") has density function $f(x)$, defined for speeds $x > 0$, continuous and positive on its support and not changing over time. That is, the number of vehicles per mile of road traveling between speeds x and $x + \Delta x$ is approximately proportional to $f(x)\,\Delta x$, when Δx is small. For example, if $f(x)$ is the uniform density on some interval $[a, b]$, then we expect an equal number of vehicles traveling at each speed between a and b.

The naïve estimate of one's percentile rank in the distribution of speeds on the highway is simply the observed proportion of vehicles passed out of the total number that one passes or is passed by. In particular, if the number of vehicles passing is equal to the number of vehicles being passed, then a driver might conclude that (s)he is driving at the median speed. Clevenson, Schilling, Watkins, and Watkins **[1]** recently showed that this is not the case. In actuality, when the number of vehicles passing equals the number being passed, the driver is, surprisingly, traveling at the *mean* speed rather than the median. More generally, the driver's speed percentile cannot be obtained merely by counting the vehicles passing and being passed.

This article explores the relationship between the naïve estimate, based on counting vehicles passing or being passed, and the actual speed percentile rank of a driver under the assumed model. We show that a driver traveling at a relatively slow or relatively fast speed will, (perhaps subconsciously) using the naïve estimate, judge his or her speed to be in a more extreme percentile than is actually the case. For instance, a person driving at the 85th percentile may perceive that (s)he is driving in an even higher speed percentile.

To begin, suppose that a particular vehicle V is traveling at speed s. Then the *actual speed percentile* of this vehicle's speed under the assumed model is $p = F(s) = \int_0^s f(x)\,dx$, while the driver's *observed speed percentile* is equal to the proportion of passed vehicles out of all other vehicles seen (either passing V or passed by V). Since vehicles at speed x will be encountered at a rate proportional to both the number of vehicles at speed x and the absolute difference between x and s, this observed percentile will converge over time to

$$p_{\text{obs}} = \frac{\int_0^s (s - x)\, f(x)\, dx}{\int_0^\infty |s - x|\, f(x)\, dx}.$$

Since F is an increasing function of s, it has an inverse, which we use to express p_{obs} as a function of p:

$$p_{\text{obs}}(p) = \frac{\int_0^{F^{-1}(p)} (F^{-1}(p) - x)\, f(x)\, dx}{\int_0^\infty |F^{-1}(p) - x|\, f(x)\, dx}. \tag{1}$$

For the simplest case, when traffic speeds are uniformly distributed, we obtain from (1) that $p_{\text{obs}}(p) = p^2/(2p^2 - 2p + 1)$. Since $p_{\text{obs}}(p)$ is not the identity function, the driver's observed percentile does not generally match his or her actual speed percentile. In fact, we find that $p_{\text{obs}}(p) < p$ for $0 < p < .5$ and $p_{\text{obs}}(p) > p$ for $.5 < p < 1$. For example, if a car is traveling in the 75th percentile of speeds, then $p_{\text{obs}}(.75) = .90$, so the driver will likely perceive that s(he) is in approximately the 90th speed percentile. A car moving in the 75th percentile of speeds will typically pass not *three* times as many vehicles as it is passed by, as one might at first expect, but *nine* times as many!

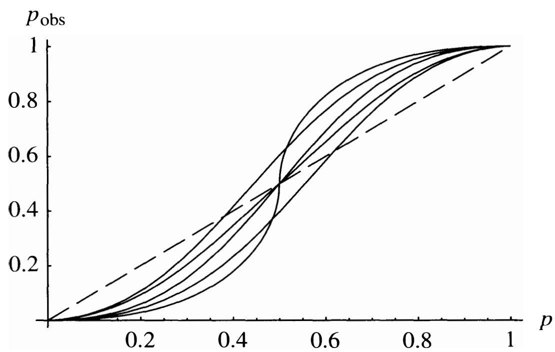


Figure 1 The observed percentile function $p_{\text{obs}}(p)$ for several traffic speed density functions

FIGURE 1 shows p_{obs} as a function of p for five different traffic speed density functions, which are shown in FIGURE 2 below. These simple models cover a wide variety of possible situations, including traffic speed distributions that are uniform, or strongly skewed towards either faster or slower speeds, or in which most drivers travel at medium speeds or at extreme speeds (either very fast or very slow). The diagonal $p_{\text{obs}} = p$ (dashed line) is shown for reference. Looking at the left half of FIGURE 1 ($0 < p < 0.5$), the curves from left to right correspond to the traffic speed density functions $f_3(x) = 2x$, $f_4(x) = 2 - 4|x - 0.5|$, $f_1(x) = 1$ (the uniform distribution), $f_2(x) = 2(1 - x)$, and $f_5(x) = 4|x - 0.5|$. The relationship between p_{obs} and p is invariant with respect to linear transformations of the traffic speed density function, so these results apply to any interval of possible speeds, for example [45 mph, 75 mph], by simply transforming the density functions in FIGURE 2 appropriately. Density functions with infinite support (such as $(0, \infty)$) show similar results, although their relevance as models for traffic speeds is dubious.

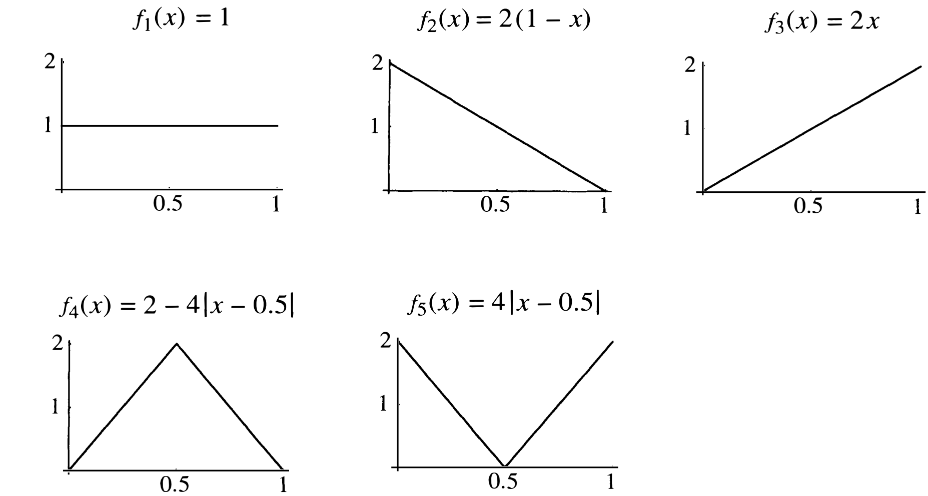


Figure 2 Traffic density functions used in FIGURE 1

We see from FIGURE 1 that in each instance there is only one speed percentile, say $\tilde{p}$, for which $p_{\text{obs}} = p$. This is nearly always the case, although exceptions do exist. The value of $\tilde{p}$ will be close to 0.5 (corresponding to the median speed of traffic) unless the traffic speed density function is highly skewed. In general, FIGURE 1 shows that a driver will tend to overestimate his or her extremity in the speed distribution, sometimes quite substantially. A driver in a high speed percentile ($> \tilde{p}$) will perceive that (s)he is in an even higher one, while a driver in a low speed percentile ($< \tilde{p}$) will think (s)he is in an even lower one.

The driver's observed speed percentile is thus a biased representation of the true speed percentile p of V unless $p = \tilde{p}$. This bias is due to the overweighting of speeds very different from s as compared to those close to s. That is, the number of vehicles a driver will see whose speeds are very different from his or her own speed overrepresents the actual number of vehicles traveling at those speeds when compared to the number of vehicles seen that are traveling at speeds similar to the driver's. For example, the number of much slower vehicles that a relatively fast driver passes is out of proportion to the actual number of such vehicles, making the driver perceive that (s)he is in an even higher speed percentile than (s)he really is.

We now show that for the stated model for traffic speeds, $p_{\text{obs}}(p)$ always takes the form shown in FIGURE 1:

THEOREM. *For any continuous density function $f(x)$ positive on its support, $p_{\text{obs}}(p)$ is a strictly increasing function, and for some p^*, $p^{**} \in (0, 1)$, we have $p_{\text{obs}}(p) < p$ for $0 < p < p^*$ and $p_{\text{obs}}(p) > p$ for $p^{**} < p < 1$.*

Proof. Write the inverse of F as $h(p) = F^{-1}(p)$. Making the substitution $u = F(x)$ in (1) yields

$$p_{\text{obs}}(p) = \frac{\int_0^p (h(p) - h(u))\,du}{\int_0^1 |h(p) - h(u)|\,du}$$

$$= \frac{\int_0^p (h(p) - h(u))\,du}{\int_0^p (h(p) - h(u))\,du + \int_p^1 (h(u) - h(p))\,du}.$$

Since $h(p)$ is an increasing function, it follows at once that $\int_0^p (h(p) - h(u))\,du$ is increasing and $\int_p^1 (h(u) - h(p))\,du$ is decreasing. A little work with inequalities shows that $p_{\text{obs}}(p)$ is increasing as claimed.

That $p_{\text{obs}}(p) < p$ for $0 < p < p^*$ for some $p^* \in (0, 1)$ follows from the fact that

$$\lim_{p \to 0^+} \frac{p_{\text{obs}}(p)}{p} = \lim_{p \to 0^+} \frac{\frac{1}{p}\int_0^p (h(p) - h(u))\,du}{\int_0^1 |h(p) - h(u)|\,du} = \frac{h(0) - h(0)}{\int_0^1 |h(0) - h(u)|\,du} = 0.$$

Similarly, to show that $p_{\text{obs}}(p) > p$ for $p^{**} < p < 1$ for some $p^{**} \in (0, 1)$, we use

$$\lim_{p \to 1^-} \frac{1 - p_{\text{obs}}(p)}{1 - p} = \lim_{p \to 1^-} \frac{\frac{1}{1-p}\int_p^1 (h(p) - h(u))\,du}{\int_0^1 |h(p) - h(u)|\,du} = \frac{h(1) - h(1)}{\int_0^1 |h(1) - h(u)|\,du} = 0. \quad \blacksquare$$

Note that we can take $p^* = p^{**}$ for each density function given in FIGURE 2. Although density functions for which p^* must be less than p^{**} exist, they have unusual forms that are unlikely to represent plausible traffic speed scenarios. FIGURE 3 shows an example:

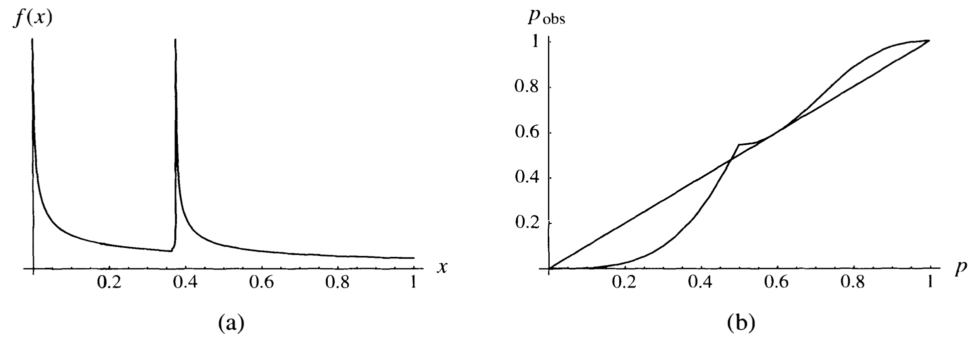


Figure 3 (a) A density function $f(x)$ with $p^* < p^{**}$; (b) $p_{\text{obs}}(p)$ for f shown in Fig. 3a ($p^* = .48$, $p^{**} = .61$)

Conclusion

Our theorem describes in simple terms the perception bias that may occur when a driver estimates the speed percentile p at which (s)he is driving by counting the number of vehicles passing or being passed. At low speeds relative to traffic, one will underestimate p, while at high speeds, one will overestimate p, regardless of the specific distribution of traffic speeds. The perception bias is greatest in situations when there is a large variation in traffic speeds (for instance, $f_5(x)$), and least when there is small variation (for instance, $f_4(x)$).

Could it be that this perception bias encourages faster drivers to slow down, and slower drivers to speed up? If so, this is not the only way in which a driver's misperceptions may affect the way he or she drives. Redelmeier and Tibrishani **[2]** showed that a driver in congested traffic may mistakenly judge that an adjacent lane is faster, perhaps leading to a needless lane change, when in fact the average speed of vehicles in that lane is just the same as in the driver's own. The phenomenon occurs because when the speeds of the two lanes fluctuate, with the same average speed for each, more of the driver's time is spent being passed by vehicles in the next lane than is spent passing such vehicles. Evidently, when it comes to judging one's speed relative to traffic, things are not generally what they seem!

REFERENCES

1. L. Clevenson, M. F. Schilling, A. Watkins, and W. Watkins, The average speed on the highway, *Coll. Math. J.*, **32**:3 (2001), 169–171.
2. D. A. Redelmeier, R. J. Tibrishani, Are those other drivers really going faster?, *Chance*, **13**:3 (2000), 8–14.

Territorial Dynamics: Persistence in Territorial Species

ROLAND H. LAMBERSON
Humboldt State University
Arcata, CA 95521
rhl1@axe.humboldt.edu

The widely studied and very controversial northern spotted owl, along with many other threatened and endangered species, exhibits territorial behavior. That is, adult pairs claim and defend a home range encompassing sufficient resources and of sufficient size to allow the pair to survive and reproduce successfully. Readers may be familiar with population models such as the logistic growth model, the Gompertz model, the Ricker model, and the Beverton-Holt model. These all capture the basic concept of limited growth (carrying capacity); however, they fail to exhibit some fundamental characteristics of the dynamics of territorial species. In particular, they do not exhibit a threshold in the density of suitable habitat below which the species is destined for extinction even if some suitable habitat is still available.

In this paper, we develop a model first proposed by Lamberson and Carroll [**1**] for the dynamics of a territorial animal or bird population. It consists of a continuous model for dispersal which distinguishes between *adults*—individuals who hold territories—and *juveniles*—those (nonterritorial) individuals that have not yet secured a home range. Here we think of birth not as the time of physical birth, but the time at which juveniles leave their natal territory and begin the search for their own home range. The model explicitly considers the cost of dispersal by including an ongoing rate of mortality due to predation and starvation while animals search for a territory. We establish that there is a threshold for density of suitable habitat, below which the population must decrease to extinction and above which the population tends to a stable positive equilibrium size.

Within populations of territorial animals we frequently find individuals that have not had the good fortune to secure a home range. These individuals, sometimes called floaters, usually occupy habitat of marginal quality and not suitable for attracting a mate. Usually, they eke out a secretive existence on the fringes of territories already claimed by other individuals. The dynamics of this floater population is important in understanding the overall behavior of the species, especially if the species is threatened or endangered. In our model, the floaters are considered part of the population of juveniles since they have not yet secured a suitable territory.

In this paper, we use a simple system of differential equations to describe the dynamics of a territorial species. The behavior of the system will be studied under both equilibrium and nonequilibrium conditions. For equilibrium conditions, we will establish: the fixed points, their stability, and the critical threshold in habitat density for persistence of the population.

An adult/juvenile model

We study an all-female model with no age structure except a distinction between territorial (adult) and nonterritorial (juvenile) animals.* It posits different survival rates for the two classes, and only the territorial individuals are considered to reproduce (though we could easily extend the model to include differential reproduction rates). Our model is distinguished from others by the inclusion of a continuous search-success term, which determines the rate at which nonterritorial females secure home ranges and move into the territorial category. We assume that the landscape consists of an array of potential territories, some of which are habitable by our animal and some not. The animal will only settle on suitable site, and only one female may occupy a particular site.

We let $T(t)$ and $J(t)$, respectively, represent the size of territorial and nonterritorial (juvenile) populations at time t. Among our territories (or sites), we let H be the number of suitable sites, those that support permanent habitation and reproduction for one female, and let S be the total number of sites, including both suitable and unsuitable sites. Since a territory is assumed to support one adult female, $T(t)$ is the number of suitable sites that are occupied at time t. The rates of change for the two segments of the population are then given by

$$\frac{dJ}{dt} = bT - qJ - \alpha\frac{(H-T)}{S}J$$

$$\frac{dT}{dt} = \alpha\frac{(H-T)}{S}J - pT.$$

The rate of change of the nonterritorial population has three terms, the birth rate, the death rate for juveniles, and what we call a *search term*—the rate at which nonterritorial animals find territories and move to the territorial class. The growth of the territorial population is given by the difference between the rate at which nonterritorial animals secure sites and the natural mortality in the territorial population. To formulate the search term, we assume that search is a process of sampling potential home ranges (with replacement). Suitable habitat is assumed to be uniformly or randomly distributed throughout the range of the population. Thus, $(H-T)/S$ is the probability of finding a suitable, unoccupied site in one draw; multiplying this probability by J, we get the approximate number of successful draws among the nonterritorial females that search during one time interval. The parameter α provides measure of the rate at which juveniles can search for suitable unoccupied sites. The parameters p and q are mortality rates for territorial and nonterritorial animals, respectively, and the fledge rate, b, comes from the mean number of female young (that survive to disperse from the natal territory) per territorial female.

Analysis of the model—fixed points and their stability

Since we have no means for solving this system of nonlinear differential equations explicitly, we will instead find the fixed points (equilibrium points) and then determine their stability; this means that we will discover whether solutions are attracted to or repelled from each of the fixed points. If we set $dJ/dt = 0$ and $dT/dt = 0$, it is a simple exercise to compute the two fixed points $J^* = T^* = 0$ (extinction), and

$$J^* = \frac{\alpha H(b-p) - Spq}{\alpha q} \quad \text{and} \quad T^* = \frac{\alpha H(b-p) - Spq}{\alpha(b-p)}.$$

*Editor's Note: Although the words *territoried* and *unterritoried* might be more descriptive, biologists use the author's terminology.

The first interesting result comes immediately from the nontrivial fixed point: The ratio of nonterritorial to territorial animals at equilibrium is a function of vital rates (birth and death rates) only and does not include habitat parameters; specifically, $J^*/T^* = (b - p)/q$. This confirms simulation results obtained by Lamberson and Brooks [**2**]. Prior to this work, the conventional wisdom had been that, as habitat diminished, the ratio of nonterritorial to territorial populations should increase.

We would next like to determine the stability of the two equilibrium points. To accomplish this we look at a linear approximation of the system centered at the fixed point (J^*, T^*).

$$\begin{bmatrix} \dfrac{dJ}{dt} \\ \dfrac{dT}{dt} \end{bmatrix} = \begin{bmatrix} -q - \alpha\left(\dfrac{H - T^*}{S}\right) & b + \alpha\dfrac{J^*}{S} \\ \alpha\left(\dfrac{H - T^*}{S}\right) & -\alpha\dfrac{J^*}{S} - p \end{bmatrix} \begin{bmatrix} J - J^* \\ T - T^* \end{bmatrix}$$

Since the system is linear, if the two eigenvalues, λ_1 and λ_2, for the matrix (the Jacobian) are distinct, solutions will have form

$$J - J^* = a_1 e^{\lambda_1 t} + a_2 e^{\lambda_2 t}$$
$$T - T^* = c_1 e^{\lambda_1 t} + c_2 e^{\lambda_2 t}$$

If both eigenvalues are negative (or, if they are complex, have negative real part) then, for solutions beginning sufficiently near the fixed point (J^*, T^*), we see that $J - J^*$ and $T - T^*$ both tend to zero as t goes to infinity, thus the fixed point is stable. For a 2×2 matrix to have both eigenvalues negative (or have negative real part) it is sufficient that the trace of the matrix be negative and the determinant be positive. We compute

$$\text{Tr}(J^*, T^*) = -\left[q + \alpha\left(\frac{H - T^*}{S}\right) + \alpha\frac{J^*}{S} + p\right]$$
$$\text{Det}(J^*, T^*) = \frac{\alpha(p - b)(H - T^*)}{S} + \frac{\alpha q J^*}{S} + pq.$$

The trace is always negative, since T cannot exceed H (we cannot have more territorial individuals than territories). We now consider the determinant at this point. Note that we must have $H/S > pq/[\alpha(b - p)]$ for the fixed point to have positive coordinates (and thus be physically meaningful). Evaluating the determinant here, we get:

$$\text{Det}(J^*, T^*) = \frac{\alpha(b - p)H}{S} - pq.$$

We see that the nontrivial fixed point is a stable attractor when

$$\frac{H}{S} > \frac{pq}{\alpha(b - p)}.$$

Thus, the nontrivial fixed point becomes a stable attractor as soon as H/S gets sufficiently large so that J^* and T^* have positive coordinates. This is equivalent to saying that extinction is avoided when the fraction of the landscape that is suitable habitat rises above $pq/[\alpha(b - p)]$.

Now, consider the determinant at the fixed point $(0, 0)$. We find that

$$\text{Det}(0, 0) = \frac{\alpha(p - b)H}{S} + pq.$$

For this to be positive, we must have

$$\frac{H}{S} < \frac{pq}{\alpha(b-p)}.$$

If suitable habitat is not sufficiently dense then extinction is inevitable. It is at this same point, when H/S passes through $pq/[\alpha(b-p])$, that (J^*, T^*) becomes unstable. We have a planar transcritical bifurcation with the two fixed points exchanging stability.

We can go a step further and examine the relationship between occupancy of suitable territories and the fraction of the habitat that is suitable. First, we see that we should not expect 100% occupancy of suitable habitat even when the entire landscape is suitable. And, when the fraction of the landscape that is suitable declines, so does occupancy of the suitable territories. In addition, this decline has a rather sharp shoulder as the density of suitable habitat gets small. All of these characteristics are also seen in the results of two quite different models that have been developed to study particular territorial species [**3, 4**].

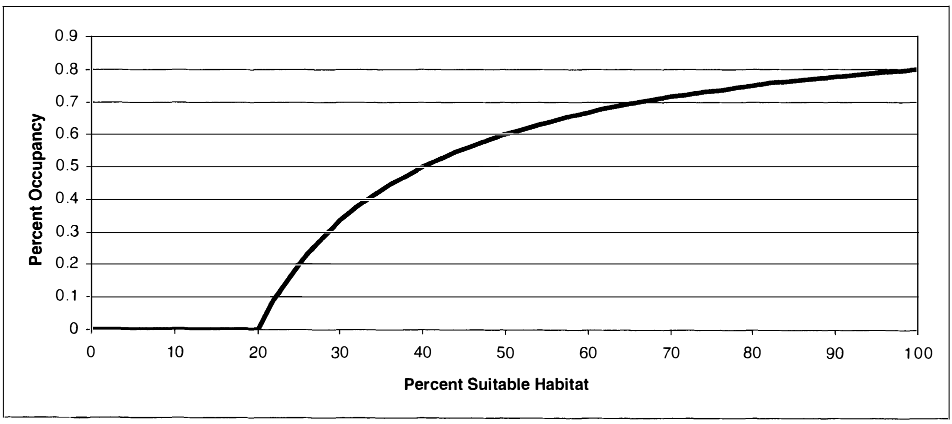


Figure 1 Using parameters from the threatened northern spotted owl we see a persistence threshold at about 20% of the landscape as suitable habitat

Conclusions

We have established the important theoretical result that for a territorial population there will be a threshold in density of suitable habitat below which the population is destined for extinction. However, if the density of suitable habitat remains above that level, then the population size should move to a stable positive equilibrium. Does this hold in reality?

There is a nice example of this in the ovenbird population in the Midwest [**5, 6**]. The ovenbird evolved in forest interiors, so its preferred habitat is large tracts of forested land. If we begin at the Minnesota-Canada border and move south, there is a steady decrease in the size and frequency of forest tracts. As we cross Iowa there are still remnants of forest habitat, but the density of such habitat is low. But, as we move further south across Missouri the density of forest begins to increase until we have nearly continuous forest in parts of the Ozarks of southern Missouri. This is depicted in FIGURE 2.

If we follow the ovenbird populations across these three states, we see high populations of ovenbirds in northern Minnesota, declining to near extinction across most of

Iowa, and then increasing again to high populations in southern Missouri (as in FIGURE 3). Similar phenomena can be seen with the northern spotted owl in the Pacific Northwest.

Figure 2 A transect of estimated forest cover from the Minnesota/Canada border (left) to the Missouri/Arkansas border (right) with the model predicted ovenbird population density (T^*)

Figure 3 A comparison of model predicted population density of ovenbirds (T^*) with estimated ovenbird densities extracted from field studies

There is a biological phenomenon called the Allee effect, which may lead to extinction for a small or low density population. It results from reduced reproductive success frequently due to difficulty finding mates or inbreeding. In our model, the Allee effect would amount to reducing the birth rate, b, in our system of equations when the population was small. However, in our case the birth rate is always fixed, so the threshold we are seeing in this model is not an Allee effect. It is a spatial effect resulting not from changes in biological parameters, but in environmental ones—loss or fragmentation of suitable habitat.

REFERENCES

1. R. H. Lamberson and J. Carroll, Thresholds for persistence in territorial species, in *Topics on Biomathematics*, I. Varbieri, E. Grassi, G. Pallotti, and P. Pettazzoni, editors. World Scientific Publishing Co., Singapore, 1993, pp. 55–62.
2. R. H. Lamberson and S. J. Brooks, The role of nonterritorial birds in a population of northern spotted owls, unpublished technical report, Humboldt State University, Arcata, CA.
3. R. H. Lamberson, R. McKelvey, B. R. Noon, and C. Voss, A dynamic analysis of northern spotted owl viability in a fragmented forest landscape, *Conservation Biology*, **6** (Dec. 1992), 505–512.
4. R. Lande, Demographic models of the northern spotted owl (*Strix occidentalis caurina*), *Oecologia*, **75** (1988), 601–07.
5. T. Donovan, F. Thompson III, John Faaborg, and J.R. Probst, Reproductive success of migratory birds in habitat sources and sinks, *Conservation Biology*, **9** (Dec. 1995), 1380–95.
6. T. Donovan, R. H. Lamberson, A. Kimber, F. Thompson III, and John Faaborg, Modeling the Effects of Habitat Fragmentation on Source and Sink Demography of Neotropical Migrant Birds, *Conservation Biology*, **9** (Dec. 1995), 1396–1407.

Some $a^n \pm b^n$ Problems in Number Theory

MIHAI MANEA
Harvard University
Cambridge, MA 02138
mmanea@fas.harvard.edu

Problems involving numbers of the form $a^n \pm b^n$ appear frequently in mathematical competitions. We offer a unified approach to solving them, which can prove to be quite efficacious and rewarding. Our theory is based on two related results, which are usually hidden in the solutions, without being clearly stated. The camouflage of the general results in particular applications leads to solutions that look highly artificial to someone lacking the appropriate background.

The theory We begin with some notation: For a prime p and a nonzero integer x, we denote by $e_p(x)$, *the exponent of p in x*, the greatest integer k for which $p^k | x$. For two integers a and b, not both zero, we denote by (a, b) their greatest common divisor.

Our first result features elegant and easily applicable formulas for the exponent of those primes that divide $a \pm b$ in numbers of the form $a^n \pm b^n$.

THEOREM 1. *Let a and b be two distinct integers, p be a prime number that does not divide ab, and n be a positive integer. Then*

(A) *if $p \neq 2$ and $p | a - b$, then*

$$e_p(a^n - b^n) = e_p(n) + e_p(a - b);$$

(B) *if n is odd, $a + b \neq 0$ and $p | a + b$, then*

$$e_p(a^n + b^n) = e_p(n) + e_p(a + b).$$

Proof. We prove (A) by induction on n. The base case, $n = 1$, is obviously true. Assume we have proved that the property holds for all the positive integers $k < n$. We shall prove that the property also holds for n.

We distinguish two cases:

Case 1. $p \nmid n$.

Let $d = a - b$. We have

$$a^n - b^n = (b+d)^n - b^n = d\left(nb^{n-1} + d\sum_{i=2}^{n}\binom{n}{i}d^{i-2}b^{n-i}\right).$$

Since p divides d, but not n or b, then

$$p \nmid nb^{n-1} + d\sum_{i=2}^{n}\binom{n}{i}d^{i-2}b^{n-i}.$$

Therefore, the exponent of p in $a^n - b^n$ is the same as the exponent of p in d. Since $e_p(n) = 0$, we have that $e_p(a^n - b^n) = e_p(a-b) + e_p(n)$ and we are done with this case.

Case 2. $p|n$.

Let $m = n/p$, $d = a^m - b^m$ and $c = b^m$. We have:

$$\begin{aligned} a^n - b^n &= (a^m)^p - (b^m)^p = (d+c)^p - c^p \\ &= pdc^{p-1} + d^p + \sum_{i=2}^{p-1}\binom{p}{i}d^i c^{p-i}. \end{aligned}$$

But

$$pd^2 | \sum_{i=2}^{p-1}\binom{p}{i}d^i c^{p-i},$$

since for all i, $2 \le i \le p-1$, $p|\binom{p}{i}$ and $d^2|d^i$.

As $p|a-b$ and $a-b|a^m-b^m$ we obtain $p|d$. Since $p \ge 3$ it follows that $pd^2|d^p$. Therefore, there exists an integer k such that

$$d^p + \sum_{i=2}^{p-1}\binom{p}{i}d^i c^{p-i} = kpd^2.$$

We get $a^n - b^n = pd(c^{p-1} + kd)$. As p does not divide c and p divides d it follows that p does not divide $c^{p-1} + kd$, so

$$e_p(a^n - b^n) = e_p(pd).$$

We recall that $m = n/p < n$ and $d = a^m - b^m$. By the induction hypothesis, $e_p(a^m - b^m) = e_p(a-b) + e_p(m)$. Therefore,

$$\begin{aligned} e_p(a^n - b^n) &= e_p(pd) = 1 + e_p(a^m - b^m) \\ &= 1 + e_p(m) + e_p(a-b) = e_p(n) + e_p(a-b) \end{aligned}$$

and the induction is complete.

To prove part (B), note that if $p \neq 2$, the statement follows from part (A) applied to the numbers a and $-b$.

For $p = 2$, we have to prove that $e_2(a^n + b^n) = e_2(n) + e_2(a+b)$ provided that a, b and n are odd. This holds because

$$a^n + b^n = (a+b)(a^{n-1} - a^{n-2}b + \cdots - ab^{n-2} + b^{n-1}).$$

The second factor of the right side of the equality above is the sum of n odd numbers; therefore, as n is odd, $e_2(a^{n-1} - a^{n-2}b + \cdots - ab^{n-2} + b^{n-1}) = 0$. Thus

$$\begin{aligned} e_2(a^n + b^n) &= e_2(a+b) + e_2(a^{n-1} - a^{n-2}b + \cdots - ab^{n-2} + b^{n-1}) \\ &= e_2(a+b) = e_2(a+b) + e_2(n) \end{aligned}$$

as desired. ■

Our second result displays a robust divisibility property of sequences of the type $\{x_n = a^n - b^n\}_{n\geq 1}$.

THEOREM 2. *If a and b are different coprime positive integers, the sequence*

$$\{x_n = a^n - b^n\}_{n\geq 1}$$

is a Mersenne sequence, *that is, for any* m, n,

$$(x_n, x_m) = x_{(n,m)}.$$

Proof. We prove the property by induction on $n+m$. The base case, $n+m = 2$ or $n = m = 1$, is trivial.

Assume that we have proved the property for all pairs of positive integers (n, m) with $n + m < k$ and let us prove it for a pair (n, m) with $n + m = k$. For $n = m$ the proof is trivial. Without loss of generality we may assume that $m < n$.

As $a^n - b^n - a^{n-m}(a^m - b^m) = b^m(a^{n-m} - b^{n-m})$ and b^m is relatively prime to $a^m - b^m$ we get that $(a^n - b^n, a^m - b^m) = (a^{n-m} - b^{n-m}, a^m - b^m)$.

Hence

$$\begin{aligned} (x_n, x_m) &= (a^n - b^n, a^m - b^m) = (a^{n-m} - b^{n-m}, a^m - b^m) \\ &= (x_{n-m}, x_m) = x_{(n-m,m)} = x_{(n,m)}, \end{aligned}$$

where the next-to-last equality follows from the induction hypothesis (because we have $(n-m) + m < n + m$) and the last equality follows from the identity $(n - m, m) = (n, m)$. ■

Applications In this section we illustrate the usefulness of our theory by solving a few moderately difficult problems.

PROBLEM 1. *Given a positive integer k, find the positive integers n for which* $3^k | 2^n - 1$. **[8]**

Solution. Since $k \geq 1$, n has to be even. Let $n = 2m$. Because 3 does not divide 1 and 4, but $3|4-1$ we can apply Theorem 1(A) to obtain:

$$\begin{aligned} e_3(2^n - 1) &= e_3(4^m - 1) \\ &= e_3(4-1) + e_3(m) \\ &= 1 + e_3(m). \end{aligned}$$

Therefore, $3^k | 2^n - 1$ if and only if there exists an integer m such that $n = 2m$ and $1 + e_3(m) \geq k$, or $3^k | 2^n - 1$ if and only if $2 \cdot 3^{k-1} | n$.

PROBLEM 2. *Let* $n > 1, a > 0$ *be integers and p be an odd prime such that* $a^p \equiv 1 (\bmod p^n)$. *Show that* $a \equiv 1 (\bmod p^{n-1})$. *(UNESCO Competition, 1995* **[1]***)*

Solution. By assumption $a^p \equiv 1(\text{mod} p)$ and by Fermat's little theorem, $a^p \equiv a(\text{mod} p)$. Hence $a \equiv 1(\text{mod} p)$. We now apply Theorem 1(A) to get

$$\begin{aligned} e_p(a^p - 1) &= e_p(a - 1) + e_p(p) \\ &= e_p(a - 1) + 1. \end{aligned}$$

As $a^p \equiv 1(\text{mod} p^n)$ it follows that $e_p(a^p - 1) \geq n$ or $e_p(a - 1) + 1 \geq n$, which means that $p^{n-1}|a - 1$ and we are done.

PROBLEM 3. *If, for a positive integer n, $3^n - 2^n$ is a power of a prime number, then n is a prime number itself. (Bulgarian Mathematical Olympiad, 1995* **[1]***)*

Solution. Let $3^n - 2^n = p^e$ where p is an (odd) prime number.

First, we prove that n cannot have two different prime divisors. Assume the contrary and let $q_1 \neq q_2$ be two primes dividing n. As $x_n = 3^n - 2^n$ is a Mersenne sequence and $(q_1, q_2) = 1$ Theorem 2 implies that

$$(3^{q_1} - 2^{q_1}, 3^{q_2} - 2^{q_2}) = 3^{(q_1,q_2)} - 2^{(q_1,q_2)} = 3 - 2 = 1.$$

But $q_i|n$, so $3^{q_i} - 2^{q_i}|3^n - 2^n = p^e$, which implies that $3^{q_i} - 2^{q_i}$ is a power of p for $i \in \{1, 2\}$. Because these two numbers need to be coprime, at least one of them will be equal to 1. This implies that either q_1 or q_2 is 1, a contradiction. Therefore n has at most one prime factor and, since $n = 1$ does not satisfy the hypothesis of the problem, there exist q prime and a positive integer l such that $n = q^l$.

Next, we assume to the contrary that $l \geq 2$. Then $3^q - 2^q|3^{q^2} - 2^{q^2}|3^n - 2^n = p^e$, so $3^q - 2^q$ and $3^{q^2} - 2^{q^2}$ are powers of p. As $q > 1$, $p|3^q - 2^q$.

We also have $p \neq 2$ or 3 and by Theorem 1(A), we obtain

$$e_p(3^{q^2} - 2^{q^2}) = e_p(3^q - 2^q) + e_p(q).$$

But $3^q - 2^q, 3^{q^2} - 2^{q^2}$ are different powers of p, so $e_p(3^{q^2} - 2^{q^2}) \neq e_p(3^q - 2^q)$. Because $e_p(q) \in \{0, 1\}$, this inequality implies $e_p(q) = 1$, i.e., $p = q$. We obtain $3^{q^2} - 2^{q^2} = q(3^q - 2^q)$.

Letting $s = 3^q$ and $t = 2^q$, we find that

$$s^q - t^q = q(s - t)$$

or equivalently,

$$\sum_{i=0}^{q-1} s^i t^{q-1-i} = q.$$

But all the q terms of the sum above are larger than 1, which leads to a contradiction. This proves that our assumption, $l \geq 2$, was false. Thus, $l = 1$, so n is a prime number.

PROBLEM 4. *Let x, y, p, n, k be positive integers such that $x^n + y^n = p^k$. Prove that, if n is odd and p is an odd prime, then n is a power of p. (Russian Mathematical Olympiad, 1996* **[1]***)*

Solution. First, it is easy to see that, under the hypothesis, $e_p(x) = e_p(y)$. Denote this common value by e. Let $a = x/p^e$ and $b = y/p^e$. Therefore, $e_p(a) = e_p(b) = 0$. We have $a^n + b^n = p^{k-en}$ (of course, $k > en$).

Assume that n is not a power of p. Let $q \neq p$ be a prime dividing n. Because n is odd, $a^{n/q} + b^{n/q} | a^n + b^n = p^{k-en}$, so $a^{n/q} + b^{n/q}$ is a power of p. As $a^{n/q} + b^{n/q} > 1$ we can apply Theorem 1(B) to obtain

$$e_p(a^n + b^n) = e_p\left(a^{n/q} + b^{n/q}\right) + e_p(q)$$
$$= e_p\left(a^{n/q} + b^{n/q}\right).$$

Here $e_p(q) = 0$ because $p \neq q$.

As $a^n + b^n$ and $a^{n/q} + b^{n/q}$ are powers of p, it follows that $a^n + b^n = a^{n/q} + b^{n/q}$. Since $a^n \geq a^{n/q}$, $b^n \geq b^{n/q}$ this can happen only if $a^n = a^{n/q}$ and $b^n = b^{n/q}$. Since $q > 1$, we need $a = b = 1$. But in this case we obtain $p = 2$, contradicting the hypotheses of the problem, which proves that the assumption that n had a prime factor different from p was false. Hence n is a power of p.

Other problems that can be solved using the theory The reader is invited to test the effectiveness of our theory as a problem solving tool by applying it to the more challenging Olympiad problems stated in this section.

PROBLEM 5. *Let p be a prime and $m \geq 2$ be an integer. Show that if the equation*

$$\frac{x^p + y^p}{2} = \left(\frac{x+y}{2}\right)^m$$

has a solution $(x, y) \neq (1, 1)$ in positive integers then $m = p$. (Balkan Mathematical Olympiad, 1993 **[3]***)*

PROBLEM 6. *Let p be a prime number and a, n positive integers. Prove that if $2^p + 3^p = a^n$, then $n = 1$. (Ireland Olympiad, 1996* **[1]***)*

PROBLEM 7. *Find the largest k for which* $1991^k | 1990^{1991^{1992}} + 1992^{1991^{1990}}$. *(IMO Shortlist, 1991* **[5]***)*

PROBLEM 8. *Find all pairs of positive integers x and y that solve the equation $p^x - y^p = 1$ where p is an odd prime. (Czech-Slovak Match, 1996* **[1]***)*

PROBLEM 9. *Show that if n is a square-free positive integer then there do not exist relatively prime positive integers x and y such that $(x + y)^3$ divides $x^n + y^n$. (Romanian Selection Test for IMO, 1994* **[2]***)*

PROBLEM 10. *Prove that the equation $x^n + y^n = (x + y)^m$ has a unique solution in integers satisfying $x > y, m > 1, n > 1$. (Romanian Selection Test for IMO, 1993* **[2]***)*

PROBLEM 11. *Let b, m, n be positive integers such that $b > 1$ and $m \neq n$. Prove that if the numbers $b^m - 1$ and $b^n - 1$ have the same prime factors then $b + 1$ is a power of 2. (IMO Shortlist, 1997* **[6]***)*

PROBLEM 12. *Prove that if a positive integer n satisfies $n | 2^n + 1$, then either $n = 3$ or $9 | n$.* **[7]**

PROBLEM 13. *Prove that there exists a positive integer $n = p_1^2 p_2 p_3 \dots p_{2000}$—with $p_1, p_2, p_3, \dots, p_{2000}$ different primes—such that $n | 2^n + 1$. (Improvement of problem 5 IMO, 2000* **[7]***)*

PROBLEM 14. *Let n be an even positive integer. Let S be the set of integers a for which $1 < a < n$ and $n | a^{a-1} - 1$. Prove that if $S = \{n - 1\}$ then $n = 2p$ where p is a prime number. (Romanian Selection Test for IMO, 2002* **[4]***)*

Conclusion In this paper we have proved two theorems that offer extensive characterizations for the number theoretical properties of the numbers of the form $a^n \pm b^n$. The theorems were illustrated to provide very effective tools for solving challenging Olympiad problems involving such numbers.

Acknowledgment. We thank Răzvan Gelca for suggestions on improving the style of the paper.

REFERENCES

1. Titu Andreescu and Kiran Kedlaya, M*athematical Contests 1995-1996: Olympiad Problems and Solutions*, MAA, 1997.
2. Mircea Becheanu, *Probleme din Olimpiade Matematice (Mathematics Olympiad Problems)*, GIL, 1995.
3. Andrei Jorza, Math Archive, Balkan Mathematical Olympiad, 1993, http://ajorza.tripod.com/mathfiles/balkan/balkan10.pdf.
4. Andrei Jorza, Math Archive, IMO Romanian Selection Tests, 2002, http://ajorza.tripod.com/mathfiles/selection2002.pdf.
5. Andrei Jorza, Math Archive, IMO Shortlist, 1991, http://ajorza.tripod.com/mathfiles/imo1991.pdf.
6. Andrei Jorza, Math Archive, IMO Shortlist, 1997, http://ajorza.tripod.com/mathfiles/imo1997.pdf.
7. Mihai Manea, Regandind o Problema de la OIM 2000 (Rethinking a problem proposed for IMO 2000), *Revista Matematica din Timisoara*, No. 2/2001.
8. Laurentiu Panaitopol, IMO Training Problems, *Gazeta Matematica*, No. 1/2000.

Two by Two Matrices with Both Eigenvalues in $\mathbb{Z}/p\mathbb{Z}$

MICHAEL P. KNAPP
Loyola College
Baltimore, MD 21210-2699
mpknapp@loyola.edu

Suppose that p is a prime number. In a recent article in the MAGAZINE [**1**], Gregor Olšavský counted the number of 2×2 matrices with entries in the field $\mathbb{Z}/p\mathbb{Z}$ that have the additional property that both eigenvalues are also in $\mathbb{Z}/p\mathbb{Z}$. In particular, he showed that there are

$$\frac{p^2}{2}(p^2 + 2p - 1)$$

such matrices.

When I began reading Olšavský's article, I thought that this would be an interesting theorem to present to my number theory class. Unfortunately for me, the key ingredient in his proof is a theorem from algebra relating the number of elements in a given conjugacy class of a group to the cardinality of the centralizer of an element in that conjugacy class. Since many of my students had not yet taken algebra and would not know about such concepts, I began to look for a proof that could be taught in an undergraduate number theory class. The purpose of this note is to provide such a proof.

Our strategy is to use the quadratic formula to find the roots of the characteristic polynomial of a matrix and then count the number of matrices for which these roots are in $\mathbb{Z}/p\mathbb{Z}$. We will follow Olšavský's notation and abbreviate $\mathbb{Z}/p\mathbb{Z}$ by $\mathcal{F}_p$. Moreover, all of the variables and congruences mentioned in the proof should be interpreted modulo p.

If $p = 2$, then we cannot divide by 2 and so cannot use the quadratic formula to find roots of polynomials. However, we can verify by a direct calculation that of the 16 possible 2×2 matrices with entries in $\mathcal{F}_2$, the only two whose eigenvalues are *not* both in $\mathcal{F}_2$ are

$$\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}.$$

So there are fourteen 2×2 matrices with entries in $\mathcal{F}_2$ and both eigenvalues in $\mathcal{F}_2$, as desired.

If $p > 2$, then suppose that A is the matrix

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

Calculating the characteristic polynomial of A gives us

$$\text{Char}(A) = \lambda^2 - (a + d)\lambda + (ad - bc).$$

We want to count the number of choices of a, b, c, and d such that both roots of this polynomial are in $\mathcal{F}_p$. Since $p \neq 2$, we can use the quadratic formula to find that the roots of Char(A) are

$$\begin{aligned} \lambda &\equiv \frac{a + d \pm \sqrt{(-(a + d))^2 - 4(ad - bc)}}{2} \\ &\equiv \frac{a + d \pm \sqrt{(a - d)^2 + 4bc}}{2}, \end{aligned}$$

where we interpret dividing by 2 to mean multiplying by the inverse and square roots are interpreted modulo p. Clearly, Char(A) has both roots in $\mathcal{F}_p$ if and only if the quantity $(a - d)^2 + 4bc$ is a perfect square in $\mathcal{F}_p$. We will count the number of choices of a, b, c, and d such that this is true.

Let a and d be any fixed elements of $\mathcal{F}_p$. There are p^2 choices for their values. If $b \equiv 0$, then for each of the p possible choices of c, we know that

$$(a - d)^2 + 4bc \equiv (a - d)^2$$

is a perfect square in $\mathcal{F}_p$. So we have $(p^2)(1)(p)$ matrices with entries in $\mathcal{F}_p$, both eigenvalues in $\mathcal{F}_p$ and $b \equiv 0$.

If b is one of the $p - 1$ possible nonzero values, then we use the fact that, including 0, there are precisely $(p + 1)/2$ perfect squares in $\mathcal{F}_p$ (these being

$$0^2, 1^2 \equiv (p - 1)^2, \ldots, \left(\frac{p - 1}{2}\right)^2 \equiv \left(\frac{p + 1}{2}\right)^2).$$

Hence there are $(p + 1)/2$ values that can be added to $(a - d)^2$ to obtain a perfect square modulo p, and each one of these is a unique multiple of $4b$. Thus for each of the $p - 1$ nonzero values of b, we see that there are $(p + 1)/2$ values of c such that $(a - d)^2 + 4bc$ is a perfect square in $\mathcal{F}_p$. So the total number of matrices with entries in $\mathcal{F}_p$, both eigenvalues in $\mathcal{F}_p$, and $b \not\equiv 0$ is $(p^2)(p - 1)(p + 1)/2$. Therefore the total number of matrices (with any value of b) having entries in $\mathcal{F}_p$ and both eigenvalues in

$\mathcal{F}_p$ is

$$(p^2)(1)(p) + (p^2)(p-1)\binom{p+1}{2} = \frac{p^2}{2}\left(p^2+2p-1\right),$$

as desired.

Acknowledgment. This paper was written while the author was supported by NSF grant DMS-0344082.

REFERENCES

1. G. Olšavský, The number of 2 by 2 matrices over $\mathbb{Z}/p\mathbb{Z}$ with eigenvalues in the same field, this MAGAZINE, **76** (2003), 314–317.

Irrationality of Square Roots

PETER UNGAR
71 Standish Dr
Scarsdale, NY 10583-6728
peterungar@yahoo.com

We present a very simple proof of the irrationality of noninteger square roots of integers. The proof generalizes easily to cover solutions of higher degree monic polynomial equations with integer coefficients. It is based on the following criterion.

A real number α is irrational if there are arbitrarily small positive numbers of the form

$$m + n\alpha; \quad \text{where } m \text{ and } n \text{ are integers.} \tag{1}$$

Indeed, if α were a fraction with denominator q, then $m + n\alpha$ would also be fraction with denominator q. Such a fraction is either zero or at least $1/q$ in magnitude.

Let us first note some previous proofs of irrationality based on this criterion. Arbitrarily small numbers $m + n\alpha$ have been constructed using Euclid's Algorithm, starting with α and 1. To prove that one gets arbitrarily small nonzero numbers, one must show that the sequence of numbers produced by Euclid's algorithm does not terminate. For certain numbers α, this can be done by finding a pair of consecutive numbers whose ratio is the same as the ratio of a previous pair of consecutive numbers. The sequence of ratios of consecutive numbers is periodic from then on.

Kalman, Mena, and Shariari [**1**] give a geometric proof that this sequence is periodic for $\alpha = \sqrt{2}$. Geometric proofs must be tailored to each specific number and they are bound to get very complicated. For instance, one can show by computation that for $\alpha = \sqrt{43}$, the ratios repeat only after 10 steps; a geometric proof would therefore have to contain dozens of points and line segments.

Using algebra, Joseph Louis Lagrange proved that Euclid's algorithm is periodic for all square roots. This result can be found in books discussing continued fractions. For other kinds of irrationals such as cube roots Euclid's algorithm is not periodic and the author does not know of a direct way of showing it will not terminate.

Kalman, Mena, and Shariari [**1**] get around these difficulties by ingenious use of matrix algebra instead of Euclid's algorithm. In the present note we obtain the arbitrarily small numbers using only algebra of the simplest kind.

The simple proof Let $\lfloor x \rfloor$ denote the greatest integer $\leq x$. If $\sqrt{d}$ is not an integer then $\sqrt{d} - \lfloor \sqrt{d} \rfloor$ is positive and < 1. Hence the positive expression $(\sqrt{d} - \lfloor \sqrt{d} \rfloor)^k$ can be made smaller than any given positive number by taking a large enough exponent. Expanding the product creates an expression of the form (1). We can conclude that $\sqrt{d}$ is not a fraction.

Next we prove the irrationality of a noninteger real root α of an nth degree equation with integer coefficients and leading coefficient 1,

$$\alpha^n + a_{n-1}\alpha^{n-1} + \cdots + a_0 = 0. \tag{2}$$

(Note that if the coefficient of α^n is a nonzero integer c instead of 1 then $c\alpha$ satisfies an equation of the form (2). Real or complex numbers that satisfy such an equation are called *algebraic integers.* They were originally studied in the effort to prove Fermat's Theorem.)

If α were rational, $\alpha = p/q$, then a positive sum of the form

$$c_0 + c_1\alpha + \cdots + c_{n-1}\alpha^{n-1}, \tag{3}$$

where $c_0, \ldots, c_{n-1}$ are integers, would have to be $\geq 1/q^{n-1}$. We can construct arbitrarily small positive expressions of the form (3) by expanding $(\alpha - \lfloor \alpha \rfloor)^k$ in powers of α and eliminating terms with exponents $\geq n$ by repeated use of

$$\alpha^n = -a_{n-1}\alpha^{n-1} - \cdots - a_0.$$

We conclude that α is irrational.

Acknowledgment. The author wishes to thank a referee of a previous version of this paper for calling his attention to [**1**]. Any reader who has seen or thought of this proof before is urged to inform the author.

REFERENCES

1. Dan Kalman, Robert Mena, and Shariat Shariari, Variations on an irrational theme—geometry, dynamics, algebra, this MAGAZINE, **70**:2 (April, 1997),93–104. (Also available at http://www.american.edu/academic.depts/cas/mathstat/People/kalman/pdffiles/irrat.pdf

Is Teaching Probability Counterproductive?

There could also be a danger to teaching too much math. In many states, the proceeds from the lottery go to education. If you teach people about probability, they'll be much less likely to play the lottery. So paradoxically, the better you teach math, the less funding you may have to teach it.

Alan Chodos, Associate Executive Officer, American Physical Society

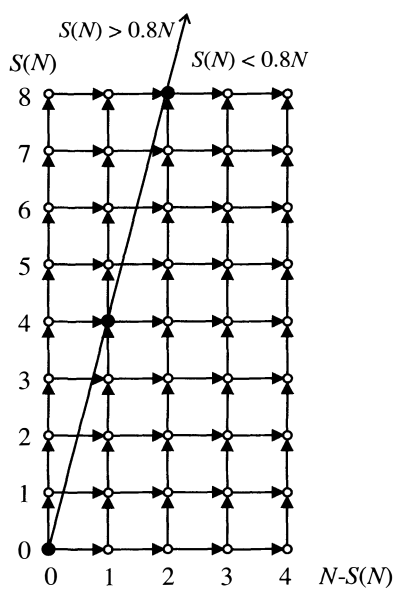
Putnam Proof Without Words

ROBERT J. MACG. DAWSON
Saint Mary's University
Halifax, Nova Scotia, Canada B3H 3C3
rdawson@cs.stmarys.ca

From the 2004 William Lowell Putnam Mathematical Competition:

> Basketball star Shanille O'Keal's team statistician keeps track of the number, $S(N)$, of successful free throws she has made in her first N attempts of the season. Early in the season, $S(N)$ was less than 80% of N, but by the end of the season, $S(N)$ was more than 80%. Was there necessarily a moment in between when $S(N)$ was exactly 80% of N?

Answer: Yes. Proof:



The reader should now also be able to answer the following "riders" that did not form part of the competition question:

1. Answer the same question assuming that Shanille had $S(N) > 0.8N$ early in the season and $S(N) < 0.8N$ at the end.
2. What other values could be substituted for 80% in the original question without changing the answer?

PROBLEMS

ELGIN H. JOHNSTON, *Editor*
Iowa State University

Assistant Editors: RĂZVAN GELCA, Texas Tech University; ROBERT GREGORAC, Iowa State University; GERALD HEUER, Concordia College; VANIA MASCIONI, Ball State University; BYRON WALDEN, Santa Clara University; PAUL ZEITZ, The University of San Francisco

Proposals

To be considered for publication, solutions should be received by September 1, 2006.

1741. *Shahin Amrahov, ARI College, Ankara, Turkey.*

Find all positive integer triples (k, m, n) that solve

a. $2^k + 9^m = 7^n$.

b. $2^k = 9^m + 7^n$.

1742. *Luz M. DeAlba and Jeffrey Langford (student), Drake University, Des Moines, IA.*

Let $\mathcal{C}$ be a circle with center O and diameter $\overline{AC}$ and let B be any point on $\mathcal{C}$ different from A and C. Let D be the point of intersection of the perpendicular bisectors of $\overline{OA}$ and $\overline{OB}$, and let E be the point of intersection of the perpendicular bisectors of $\overline{OC}$ and $\overline{OB}$. Prove that $\triangle DBE$ is similar to $\triangle ABC$.

1743. *David P. Lang, Wentworth Institute of Technology, Boston, MA.*

Let $\{a_n\}_{n=1}^{\infty}$ be a sequence of positive real numbers. For positive integer n, define $s_n = \sum_{j=1}^{n} a_j$, and for $n \geq 2$ define S_n by

$$S_n = \sum_{k=1}^{n} \frac{a_k}{s_n - a_k}.$$

Prove that if $L = \lim_{n\to\infty} S_n$ exists, then $L \geq 1$.

We invite readers to submit problems believed to be new and appealing to students and teachers of advanced undergraduate mathematics. Proposals must, in general, be accompanied by solutions and by any bibliographical information that will assist the editors and referees. A problem submitted as a Quickie should have an unexpected, succinct solution.

Solutions should be written in a style appropriate for this MAGAZINE. Each solution should begin on a separate sheet.

Solutions and new proposals should be mailed to Elgin Johnston, Problems Editor, Department of Mathematics, Iowa State University, Ames IA 50011, or mailed electronically (ideally as a LaTeX file) to ehjohnst@iastate.edu. All communications should include the reader's name, full address, and an e-mail address and/or FAX number.

1744. *Michael Goldenberg and Mark Kaplan, The Ingenuity Project, Baltimore Polytechnic Institute, Baltimore, MD.*

For real number $x \geq 1$, define $a_1 = 2x$ and $a_{n+1} = a_n^2 - 2$, $n = 1, 2, 3, \ldots$. Find a closed form expression for

$$S(x) = \sum_{n=1}^{\infty} \prod_{k=1}^{n} a_k^{-1}.$$

1745. *Gerald A. Edgar, The Ohio State University, Columbus, OH.*

For a positive real number r, define

$$A_0(r) = r, \qquad A_{n+1}(r) = r + \frac{A_n(r)}{A_n(r+1)}, \qquad n = 0, 1, 2, \ldots.$$

Show that $\lim_{n\to\infty} A_n(1)$ exists and find the value of the limit.

Note: Expanding the recursive definition when $r = 1$, one is lead to the elaborate continued fraction-like configuration,

$$1+\cfrac{1+\cfrac{1+\cfrac{1+\cfrac{1+\cdots}{2+\cdots}}{2+\cfrac{2+\cdots}{3+\cdots}}}{2+\cfrac{2+\cfrac{2+\cdots}{3+\cdots}}{3+\cfrac{3+\cdots}{4+\cdots}}}}{2+\cfrac{2+\cfrac{2+\cfrac{2+\cdots}{3+\cdots}}{3+\cfrac{3+\cdots}{4+\cdots}}}{3+\cfrac{3+\cfrac{3+\cdots}{4+\cdots}}{4+\cfrac{4+\cdots}{5+\cdots}}}}$$

Quickies

Answers to the Quickies are on page 155.

Q959. *Clark Kimberling, Evansville, IN., and Peter J.C. Moses, Redditch, England.*

For a random function $f : [n] \to [n]$, let E be the expected number of elements of $[n]$ that are not in the range of f. Prove that

$$\lim_{n\to\infty} \frac{E}{n} = \frac{1}{e}.$$

(Here $[n] = \{1, 2, \ldots, n\}$. Assume that each $f : [n] \to [n]$ occurs with probability $1/n^n$.)

Q960. *Michael W. Botsko, Saint Vincent College, Latrobe, PA.*

Let $\{x_n\}_{n=1}^{\infty}$ be a sequence of real numbers such that

$$|x_n - x_m| < \max\{|x_{n-1} - x_n|, |x_{m-1} - x_m|\},$$

for all $m, n \geq 2$. Prove that the sequence is bounded.

Solutions

Permanent Values **April 2005**

1716. *Café Dalat Problem Solving Group, Washington D.C.*

Let k, n be integers with $n \geq 1$ and $0 \leq k \leq n$. Prove that there is an $n \times n$ matrix A of 0s and 1s with $\text{per}(A) = k$. (Here $\text{per}(A)$ denotes the permanent of A.)

Solution by Eddie Cheng, Oakland University, Rochester, MI.

Let $\text{per}(A)$ denote the permanent of the square matrix A. Because $\text{per}(0_{n\times n}) = 0$ and $\text{per}(I_n) = 1$, the result is true for $k = 0$ and $k = 1$ for any $n \geq k$. For $k \geq 2$, consider the $k \times k$ matrix

$$B_k = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & \cdots & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & \cdots & 0 & 0 & 0 \\ & & & & & \vdots & & & \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 1 & 1 \end{pmatrix},$$

where the first row, main diagonal, and sub diagonal consists of 1s and all other entries are 0s. Then $\text{per}(B_k) = 1 + \text{per}(B_{k-1})$, and it follows by induction that $\text{per}(B_k) = k$. It then follows that the $n \times n$ matrix

$$\begin{pmatrix} B_k & 0_{k\times(n-k)} \\ 0_{(n-k)\times k} & I_{(n-k)} \end{pmatrix}$$

has permanent k.

Also solved by JPV Abad, Tsehaye Andeberhan, Michel Bataille (France), Arthur Benjamin, J. C. Binz (Switzerland), Brian Bradie, Robert Calcaterra, Randall J. Covill, Knut Dale (Norway), Luz M. De Alba, A. K. Desai (India), Michael Goldenberg and Mark Kaplan, G.R.A.20 Math Problems Group (Italy), Jerrold W. Grossman, Amanda Harsy, Russell Jay Hendel, Eugene A. Herman, Houghten College Math Club, Harris Kwong, Peter W. Lindstrom, Thomas C. McMillan and Xiaoshen Wang, Northwestern University Math Problem Solving Group, Kees Onneweer, Frederick G. Schmitt, Byron Siu, W. R. Smythe, Salvatore Tringali (Italy), Li Zhou, and the proposer.

A Triangle Inequality **April 2005**

1717. *Mohammed Aassila, Strasbourg, France*

Let ABC be a triangle, and let A_1, B_1, C_1 be on BC, CA, AB, respectively, with none of A_1, B_1, C_1 coinciding with a vertex of ABC. Show that if

$$AB + BA_1 = AC + CA_1,$$
$$AB + AB_1 = BC + CB_1,$$

and

$$AC + AC_1 = BC + BC_1,$$

then

$$\text{Area}(ABC) \geq 4 \cdot \text{Area}(A_1B_1C_1).$$

Solution by Tom Zerger, Saginaw Valley State University, University Center, MI.

Let $a = BC$, $b = CA$, $c = AB$, and $s = \frac{1}{2}(a + b + c)$. Then the conditions of the problem state that

$$c + BA_1 = s = b + CA_1, \quad c + AB_1 = s = a + CB_1,$$

and

$$b + AC_1 = s = a + BC_1.$$

It follows that

$$\frac{AB_1}{B_1C} \cdot \frac{CA_1}{A_1B} \cdot \frac{BC_1}{C_1A} = 1,$$

and hence, by Ceva's Theorem, $\overline{AA'}$, $\overline{BB'}$ and $\overline{CC'}$ are concurrent at a point P inside of $\triangle ABC$.

We prove the more general result that if $\overline{AA'}$, $\overline{BB'}$ and $\overline{CC'}$ are concurrent at a point P inside of $\triangle ABC$, then the desired inequality holds, with equality if and only if P is the centroid of $\triangle ABC$. (With the additional conditions given in the problem statement, P is the Nagel point of $\triangle ABC$ and A_1, B_1, C_1 are the points of tangency of the excircles to $\triangle ABC$.)

We use barycentric coordinates, with $A = (1, 0, 0)$, $B = (0, 1, 0)$, $C = (0, 0, 1)$. Let $BA_1/A_1C = p/1$, $CB_1/B_1A = q/1$, and $AC_1/C_1B = r/1$. Then, in barycentric coordinates $A_1 = (0, p, 1)$, $B_1 = (q, 0, 1)$, and $C_1 = (1, r, 0)$. Furthermore, the area of $\triangle A_1B_1C_1$ is given by

$$\text{Area}(A_1B_1C_1) = \frac{1}{(p+1)(q+1)(r+1)} \begin{vmatrix} 0 & 1 & p \\ q & 0 & 1 \\ 1 & r & 0 \end{vmatrix} \cdot \text{Area}(ABC). \tag{1}$$

Because the cevians are concurrent at P, we have $pqr = 1$. Evaluating the determinant in (1) we find

$$\text{Area}(A_1B_1C_1) = \frac{2}{(p+1)(q+1)(r+1)}\text{Area}(ABC). \tag{2}$$

If P is inside of $\triangle ABC$ (as is the case when P is the Nagel point), then p, q, r are positive. Then, because $pqr = 1$,

$$\begin{aligned}(p+1)(q+1)(r+1) &= 2 + (p+q+r) + (pq+qr+rp) \\ &= 2 + \left(p + \frac{1}{p}\right) + \left(q + \frac{1}{q}\right) + \left(r + \frac{1}{r}\right) \geq 8,\end{aligned}$$

and equality holds if and only if $p = q = r = 1$. It follows from (2) that

$$\text{Area}(A_1B_1C_1) \leq \frac{1}{4}\text{Area}(ABC),$$

with equality if and only if P is the centroid of $\triangle ABC$.

Also solved by Michel Bataille (France), Herb Bailey, J. C. Binz (Switzerland), Robert Calcaterra, Minh Can, Jesús Jerómino Castro (Mexico), Joseph Cooper, Chip Curtis, Daniele Donini (Italy), Robert L. Doucette, John Ferdinands, Michael Goldenberg and Mark Kaplan, Mowaffaq Hajja (Jordan), John Mangual, Peter E. Nëusch (Switzerland), Northwestern University Math Problem Solving Group, Muneer Ahmad Rashid, Mark de Saint-Rat, Volkhard Schindler (Germany), Raúl Simon (Chile), John W. Spellman and Ricardo M. Torrejón, H. T. Tang, R. S. Tiberio, Michael Vowe (Switzerland), Paul Weisenhorn (Germany), Li Zhou, and the proposer.

Tiling a Frame **April 2005**

1719. *G.R.A.20 Problems Group, Università di Roma, Rome, Italy.*

From an $(n+4) \times (n+4)$ checkerboard of unit squares, the central $n \times n$ square is removed to leave a square frame of width 2. In how many ways can the frame be tiled with 1×2 dominos? (Two different tilings that can be made identical through a rotation of the frame are considered different.)

Solution by Harris Kwong, SUNY Fredonia, Fredonia, NY.

Position the checkerboard in the Cartesian plane so it is inside the region $[0, n+4] \times [0, n+4]$. Each of the four corner squares is covered either by a horizontal 2×1 or a vertical 1×2 domino. Call a tiling "noninterlocking" if the dominos covering the remaining squares can be partitioned into four rectangular checker boards of width 2 or height 2. Otherwise call the tiling "interlocking." It is easy to check that there are only two interlocking tilings. These occur when in the lower left corner we have a domino covering each of $[0,1] \times [0,2]$,$[1,3] \times [0,1]$, $[1,2] \times [1,3]$, or a domino covering each of $[0,2] \times [0,1]$, $[0,1] \times [1,3]$, $[1,3] \times [1,2]$. It is also easy to check that once dominos are laid one of these two patterns, then the rest of the tiling is uniquely determined.

Now let v_k, $0 \le k \le 4$, be the number of noninterlocking tilings in which k of the corner squares are covered by vertical dominos. By symmetry, we have $v_0 = v_4$ and $v_1 = v_3$. We first determine v_0. After covering the four corner squares with horizontal dominos, the remaining squares can be partitioned into four checkerboards, two of size $(n+2) \times 2$ and two of size $2 \times n$. It is well known that a $2 \times r$ checkerboard can be tiled by 1×2 dominos in F_{r+1} ways, where F_m denotes the mth Fibonacci number, defined by $F_0 = 0$, $F_1 = 1$, and $F_m = F_{m-1} + F_{m-2}$ for $m \ge 2$. Thus

$$v_0 = v_4 = F_{n+1}^2 F_{n+3}^2.$$

To determine v_1, note that the vertical domino can be placed in any one of four corners, with horizontal dominos in the other three. The remaining part of the board can then be partitioned into one $2 \times n$, two $2 \times (n+1)$ and one $2 \times (n+2)$ board. Thus

$$v_1 = v_3 = 4F_{n+1}F_{n+2}^2 F_{n+3}.$$

If in the corners are two vertical squares and two horizontal squares, then the vertical squares can be in opposite corners (two ways) or in adjacent corners (four ways.) In the first case the remaining squares can be partitioned into four $2 \times (n+1)$ boards, and in the second case, into one $2 \times n$, two $2 \times (n+1)$ and one $1 \times (n+2)$ boards. Hence

$$v_2 = 2F_{n+2}^4 + 4F_{n+1}F_{n+2}^2 F_{n+3}.$$

Combining all of these counts, we find that the number of ways to tile the frame is

$$T_n = 2F_{n+2}^4 + 12F_{n+1}F_{n+2}^2 F_{n+3} + 2F_{n+1}^2 F_{n+3}^2 + 2.$$

The Cassini-Simson identity states that $F_{r-1}F_{r+1} - F_r^2 = (-1)^r$. This can be applied to simplify the expression for T_n to obtain

$$T_n = 4(2F_{n+2}^2 + (-1)^n)^2.$$

Also solved by Arthur Benjamin, J. C. Binz (Switzerland), Robert Calcaterra, Richard F. McCoart, Kim McInturff, Northwestern University Math Problem Solving Group, The Problem Possee, Li Zhou, and the proposer. There were two incorrect submissions.

Decreasing Expectations **April 2005**

1720. *Stephen J. Herschkorn, Rutgers University, Highland Park, NJ.*

Let X be a standard normal random variable and let a be a positive number. Show that $E[X : |X - a| < t]$ is strictly decreasing in nonnegative t.

Solution by Li Zhou, Polk Community College, Winter Haven, FL.

By definition

$$E[X : |X - a| < t] = \frac{\int_{a-t}^{a+t} xe^{-x^2/2}\,dx}{\int_{a-t}^{a+t} e^{-x^2/2}\,dx}.$$

This is also the abscissa of the centroid of A_t, where A_t is the region bounded by the graphs of $y = e^{-x^2/2}$, $y = 0$, $x = a - t$, and $x = a + t$. For any $t > 0$ and $\Delta t > 0$, it is clear from the shape of the graph of $y = e^{-x^2/2}$ that as t increases from t to $t + \Delta t$, more area is gained on the left of the region than on the right. Hence the centroid of $A_{t+\Delta t}$ will be to the left of the centroid of A_t. This completes the proof.

Also solved by Robert Calcaterra, Daniele Donini (Italy), James C. Hickman, Peter W. Lindstrom, John Mangual, Paul Weisenhorn (Germany), and the proposer.

Answers

Solutions to the Quickies from page 151.

A959. *Solution by Assistant Editor Byron Walden.* Each value $1, 2, \ldots, n$ is omitted by $(n-1)^n$ of the functions. Thus in considering all functions, there are $n(n-1)^n$ omitted values. It follows that

$$E = n\left(1 - \frac{1}{n}\right)^n, \quad \text{so} \quad \lim_{n\to\infty} \frac{E}{n} = \frac{1}{e}.$$

A960. Note that

$$|x_n - x_{n+1}| < \max\{|x_{n-1} - x_n|, |x_n - x_{n+1}|\},$$

so $|x_n - x_{n+1}| < |x_{n-1} - x_n|$ for $n \geq 2$. Therefore,

$$|x_n - x_{n+1}| < |x_{n-1} - x_n| < |x_{n-2} - x_{n-1}| < \cdots < |x_1 - x_2|.$$

Thus $|x_n - x_{n+1}| < |x_1 - x_2|$ for all $n \geq 2$. Now consider

$$|x_2 - x_n| < \max\{|x_1 - x_2|, |x_{n-1} - x_n|\} = |x_1 - x_2|,$$

so $|x_2 - x_n| < |x_1 - x_2|$ for $n \geq 2$. Therefore $\{x_n\}$ is bounded.

Note. There are sequences that satisfy the conditions in the problem statement. One such is the sequence $\{x_n\}$ defined by $x_n = 1/2^n$.

REVIEWS

PAUL J. CAMPBELL, *Editor*
Beloit College

Assistant Editor: Eric S. Rosenthal, West Orange, NJ. Articles and books are selected for this section to call attention to interesting mathematical exposition that occurs outside the mainstream of mathematics literature. Readers are invited to suggest items for review to the editors.

Hayes, Brian, Unwed numbers, *American Scientist* (January-February 2006). `http://www.americanscientist.org/template/AssetDetail/assetid/48550?` . Knuth, Donald E., Dancing links, P159 at `http://www-cs-faculty.stanford.edu/~knuth/preprints.html` ; DANCE [code for exact cover problem], `http://www-cs-faculty.stanford.edu/~knuth/programs/dance.w` . Bailey, R.A., Letters: Latin squares and Su doku, *Significance* (December 2005) 187.

Solving sudoku—no math required? Well, no arithmetic, algebra, geometry, or trigonometry. But logic is a necessity, particularly as encapsulated in various local and nonlocal "rules" that apply to patterns of increasing sophistication. Such rules try to capture the conclusions of backtracking (systematic trial and error) while still working in a "forward-looking, nonspeculative mode"—that is, without guessing and perhaps later erasing. From a mathematical point of view, a sudoku puzzle is an *exact cover problem*: Given a 0–1 matrix, find a set of rows that contains a single 1 in each column. A sudoku problem translates into a matrix with 729 rows (9 digits times 81 squares) and 243 columns (constraints for rows, columns, and blocks). Knuth applies his dancing links algorithm for an exact cover problem to solve pentomino-packing, queen-placing, and other problems; he wrote in 2000, before the sudoku phenomenon, but others since have realized its relevance and implemented it for sudoku. And just in case you think sudoku is new and/or a useless frenzy: R.A. Bailey cites a 1956 paper on a sudoku combinatorial design for use in agricultural field trials, with further papers in 1990 and 1991.

Adams, Colin, The theorem blaster, *Mathematical Intelligencer* 28 (1) (2006) 17–18.

"Is your theorem overweight? ... we are offering to the public, for the first time ever, the amazing Theorem Blaster®. ... You've seen it on Math TV.... Overweight theorems can be dangerous in the long term. Sluggish and impenetrable... But now you can have the theorem you always dreamed of.... [A]s a special deal for those ordering today, in addition to the Theorem Blaster... you will also receive instructions for the world-famous Analysis Diet®. That's right. The diet where you just drop analysis from your mathematical diet and watch the bloat come off." Better order fast, before eager grad students buy out the limited supply!

Strogatz, Steven H., et al., Crowd synchrony on the Millennium Bridge, *Nature* 438 (3 November 2005) 43–44.

Army officers have known for centuries to order their soldiers to break step in crossing a bridge, but engineers have been slower to learn the corresponding lesson for their side. When London's Millennium Bridge was opened in 2000, a crowd streamed on—and the bridge began to sway. The individuals "spontaneously fell into step with the bridge's vibrations, inadvertently amplifying them." The authors model the bridge as a simple damped harmonic oscillator and also model the effect of its movement on the pedestrians' gait. They explain the dynamics of the resulting simple two-equation system. The "synchronization timescale," the amplitude of the sway, and the number of pedestrians to induce the effect all agree with actual experiments. The bridge recently reopened, after $9 million in repairs to increasing its damping.

Woolsey, Robert E.D., *Real World Operations Research: The Woolsey Papers*, Lionheart Publishing, 2003; vi + 156 pp, $19.95 (P). ISBN 1–931634–25–4.

Gene Woolsey, Prince of the Pragmatic and the antidote to Dilbertism, has been past president of the Institute of Management Sciences (now INFORMS), editor of five journals, professor at eight colleges (in four countries), and author of eight books and over 100 papers. He has graduated more than 200 master's students and 47 Ph.D.'s. No operations research student of his graduates without a thesis on "a real-world problem done for a company/agency that *uses* the results" with "a provable reduction in costs or increase in profits"; if the amount exceeds $1M in the first year, Woolsey awards a diamond stickpin. Saved so far by his graduates, *by using mathematics, people skills, and common sense*: $820M. Most of his provocative and highly witty essays collected here appeared in *Interfaces* over the past 30 years; they are wonderfully entertaining wisdom of the first order. For years, I have given his "On becoming an elephant" to all pre-engineers, to help persuade them of the merits of a liberal arts background. That essay concludes, "Training is the absorption of skills. Education is the acquisition of new hungers."

Schattschneider, Doris, *M.C. Escher: Visions of Symmetry*, 2nd ed., Harry N. Abrams, 2004; xiii + 370 pp, $29.95. ISBN 0–8109–4308–5.

This second edition of the beautiful 1990 book about Escher's work and the mathematics behind it contains an additional 20-page afterword and an expanded bibliography. From the afterword, we learn more about the fruitful early interaction of Escher with George Pólya and the latter's unfulfilled intentions to write a book for lay people on "the symmetry of ornament." There is more known now, from examination of Escher's sketchbooks, about which patterns he discovered and which he did not but might have. Mathematical questions remain, such as categorizing all hypersymmetric tiles (ones with mirror symmetry but in no tiling with the tile is its mirror symmetry a symmetry of the tiling). (That a book with so many color plates can be sold at so low a price is a wonder; then again, it was printed in China.)

Rabinowitz, Stanley, and Mark Bowron (eds.), *Index to Mathematical Problems 1975–1979*, MathPro Press, 1999; x + 518 pp, $69.95. ISBN 0–9626401–2–3.

This book is the second volume of an ambitious but valuable series endeavoring to complete the monumental task of indexing all mathematical problems appearing in journals and contests (but not in problem books, and not including problems in Martin Gardner's columns in *Scientific American*). The problem statements are all given in full. The problems are listed sorted by topic, with indexes by author (including those who provided solutions or comments), problem title, keyword, and more (including a key to pseudonyms). Vol. 1 covered 1980–1984; the publishers also have volumes on the *Leningrad Mathematical Olympiads 1987–1991*, *ARML–NYSML Contests 1989-1994*, and *Problems and Solutions from The Mathematical Visitor 1877–1894*. Moreover, they provide for online search at `http://www.problemcorner.org/` . I was pleased to find my name in this volume as a solver of two problems (that I don't remember at all!), but I could not locate those problems or my name at the online search (which also confusingly appears to interpret the search for a name as "first name" OR "last name"). The journals included are all in English, except for *Nieuw Archief voor Wiskunde*, whose problems are translated. (Despite the copyright date of 1999, this book seems to have appeared in 2005.)

Ellenberg, Jordan, Is math a sport? And what about target shooting, Skee-Ball, and standing on one foot?, *Slate* (15 July 2004) `http://slate.msn.com/id/2103903/` .

The Winter Olympics are beginning as I write this, with assorted snowboarding and other events—some with dubious popularity and a handful of participants worldwide—that did not exist when I was young. But there weren't "mathletes" then, either, in interscholastic competitions where they and their teachers strive for recognition on a par with athletes. "Could mathletes some day compete alongside track stars and basketball players" in the Olympics, where chess and bridge have been "exhibition sports"? What is a sport? or a game? Mathematician Ellenberg suggests that those are not well-defined terms but mathematics is not a sport. It is "something better: a game you can't ever really win." Does that answer satisfy?

NEWS AND LETTERS

34th United States of America Mathematical Olympiad April 19 and 20, 2005

Edited by Zuming Feng, Cecil Rousseau, and Melaine Wood

PROBLEMS

1. Determine all composite positive integers n for which it is possible to arrange all divisors of n that are greater than 1 in a circle so that no two adjacent divisors are relatively prime.
2. Prove that the system

$$x^6 + x^3 + x^3y + y = 147^{157}$$
$$x^3 + x^3y + y^2 + y + z^9 = 157^{147}$$

has no solutions in integers x, y, and z.

3. Let ABC be an acute-angled triangle, and let P and Q be two points on side BC. Construct point C_1 in such a way that convex quadrilateral $APBC_1$ is cyclic, $QC_1 \parallel CA$, and C_1 and Q lie on opposite sides of line AB. Construct point B_1 in such a way that convex quadrilateral $APCB_1$ is cyclic, $QB_1 \parallel BA$, and B_1 and Q lie on opposite sides of line AC. Prove that points B_1, C_1, P, and Q lie on a circle.
4. Legs L_1, L_2, L_3, L_4 of a square table each have length n, where n is a positive integer. For how many ordered 4-tuples (k_1, k_2, k_3, k_4) of nonnegative integers can we cut a piece of length k_i from the end of leg L_i $(i = 1, 2, 3, 4)$ and still have a stable table? (The table is *stable* if it can be placed so that all four of the leg ends touch the floor. Note that a cut leg of length 0 is permitted.)
5. Let n be an integer greater than 1. Suppose $2n$ points are given in the plane, no three of which are collinear. Suppose n of the given $2n$ points are colored blue and the other n colored red. A line in the plane is called a *balancing line* if it passes through one blue and one red point and, for each side of the line, the number of blue points on that side is equal to the number of red points on the same side. Prove that there exist at least two balancing lines.
6. For m a positive integer, let $s(m)$ be the sum of the digits of m. For $n \geq 2$, let $f(n)$ be the minimal k for which there exists a set S of n positive integers such that $s\left(\sum_{x\in X} x\right) = k$ for any nonempty subset $X \subset S$. Prove that there are constants $0 < C_1 < C_2$ with

$$C_1 \log_{10} n \leq f(n) \leq C_2 \log_{10} n.$$

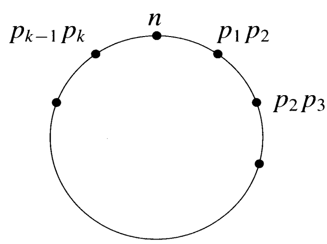
Note: For interested readers, the editors recommend *Mathematical Olympiads 2005*. There the problems are presented together with a collection of selected solutions developed by the examination committees, contestants, and experts, during and after the contests.

SOLUTIONS

1. No such circular arrangement exists for $n = pq$, where p and q are distinct primes. In that case, the numbers to be arranged are p, q and pq, and in any circular arrangement, p and q will be adjacent. We claim that the desired circular arrangement exists in all other cases. If $n = p^e$ where $e \geq 2$, an arbitrary circular arrangement works. Henceforth we assume that n has prime factorization $p_1^{e_1} p_2^{e_2} \cdots p_k^{e_k}$, where $p_1 < p_2 < \cdots < p_k$ and either $k > 2$ or else $\max(e_1, e_2) > 1$. Let D_n denote the set of divisors of n that are greater than 1; that is,

$$D_n := \{d : d \mid n \text{ and } d > 1\}.$$

To construct the desired circular arrangement of the elements of D_n, start with the circular arrangement of $n,\ p_1p_2,\ p_2p_3,\ \ldots,\ p_{k-1}p_k$ as shown below.



Then between n and p_1p_2, place (in arbitrary order) all other members of D_n that have p_1 as their smallest prime factor. Between p_1p_2 and p_2p_3, place all members of D_n other than p_2p_3 that have p_2 as their smallest prime factor. Continue in this way, ending by placing $p_k, p_k^2, \ldots, p_k^{e_k}$ between $p_{k-1}p_k$ and n. It is easy to see that each element of D_n is placed exactly one time, and any two adjacent elements have a common prime factor. Hence this arrangement has the desired property.

2. We will show there is no solution to the system modulo 13. Add the two equations and add 1 to obtain

$$(x^3 + y + 1)^2 + z^9 = 147^{157} + 157^{147} + 1.$$

By Fermat's Theorem, $a^{12} \equiv 1 \pmod{13}$ when a is not a multiple of 13. Hence we compute $147^{157} \equiv 4^1 \equiv 4 \pmod{13}$ and $157^{147} \equiv 1^3 \equiv 1 \pmod{13}$. Thus

$$(x^3 + y + 1)^2 + z^9 \equiv 6 \pmod{13}.$$

The cubes mod 13 are 0, ± 1, and ± 5. The first of the two given equations yields the congruence

$$(x^3 + 1)(x^3 + y) \equiv 4 \pmod{13}.$$

Thus there is no solution in case $x^3 \equiv -1 \pmod{13}$, and for x^3 congruent to $0, 1, 5, -5$, correspondingly $x^3 + y$ must be congruent to $4, 2, 5, -1$. Hence

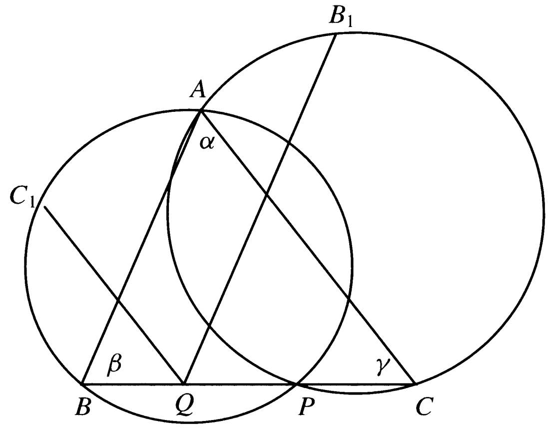
$$(x^3 + y + 1)^2 \equiv 12, 9, 10, \text{ or } 0 \pmod{13}.$$

Also z^9 is a cube, hence z^9 must be 0, 1, 5, 8, or 12 (mod 13). The following table shows that 6 modulo 13 is not obtained by adding one of 0, 9, 10, 12 to one of 0, 1, 5, 8, 12.

+	0	1	5	8	12
0	0	1	5	8	12
9	9	10	1	4	8
10	10	11	2	5	9
12	12	0	4	7	11

Hence the system has no solutions in integers. It is worth noting that this argument shows there is no solution even if z^9 is replaced by z^3.

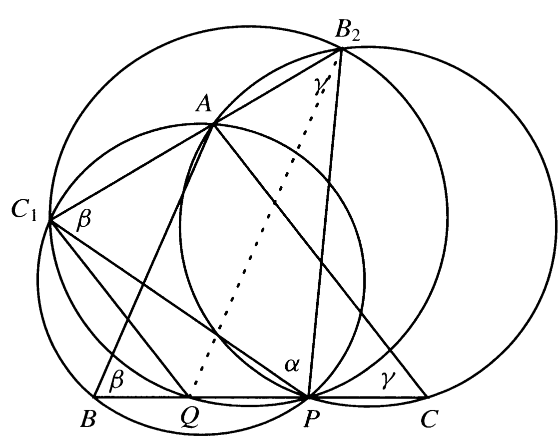
3. Let α, β, γ denote the angles of triangle ABC. Without loss of generality, we assume that Q is on the segment BP.



We guess that B_1 is on the line through C_1 and A. To confirm that our guess is correct and prove that B_1, C_1, P, and Q lie on a circle, we start by letting B_2 be the point other than A that is on the line through C_1 and A, and on the circle through C, P, and A. Because AB_2CP is cyclic, $\angle AB_2P = \angle ACP = \gamma$. From $QC_1 \parallel CA$ we have $\angle PQC_1 = 180° - \gamma$. Hence $\angle PQC_1 + \angle C_1B_2P = \angle PQC_1 + \angle AB_2P = 180°$, and so quadrilateral PQC_1B_2 is cyclic. Because PQC_1B_2 and AC_1BP are cyclic, we conclude that

$$\angle PQB_2 = \angle PC_1B_2 = \angle PC_1A = \angle PBA = \beta,$$

from which it follows that $QB_2 \parallel AB$. Therefore, $B_1 = B_2$ and thus points P, Q, C_1, and B_1 lie on a circle.



4. Turn the table upside down so its surface lies in the xy-plane. We may assume that the corner with leg L_1 is at $(1, 0)$, and the corners with legs L_2, L_3, L_4 are at $(0, 1)$, $(-1, 0)$, and $(0, -1)$, respectively. (We may do this because rescaling the x and y coordinates does not affect the stability of the cut table.) For $i = 1, 2, 3, 4$, let ℓ_i be the length of leg L_i after it is cut. Thus $0 \le \ell_i \le n$ for each i. The table will be stable if and only if the four points $F_1(1, 0, \ell_1)$, $F_2(0, 1, \ell_2)$, $F_3(-1, 0, \ell_3)$, and $F_4(0, -1, \ell_4)$ are coplanar. This will be the case if and only if F_1F_3 intersects F_2F_4, and this will happen if and only if the midpoints of the two segments coincide, that is,

$$\left(0, 0, \frac{\ell_1 + \ell_3}{2}\right) = \left(0, 0, \frac{\ell_2 + \ell_4}{2}\right). \qquad (*)$$

Because each ℓ_i is an integer satisfying $0 \le \ell_i \le n$, the third coordinate for each of these midpoints can be any of the numbers $0, \frac{1}{2}, 1, \frac{3}{2}, \ldots, n$.

For each nonnegative integer $k \le n$, let S_k be the number of solutions of $x + y = k$ where x, y are integers satisfying $0 \le x, y \le n$. The number of stable tables (in other words, the number of solutions of $(*)$) is $N = \sum_{k=0}^{n} S_k^2$.

Next we determine S_k. For $0 \le k \le n$, the solutions to $x + y = k$ are described by the ordered pairs $(j, k - j)$, $0 \le j \le k$. Thus $S_k = k + 1$ in this case. For each $n + 1 \le k \le 2n$, the solutions to $x + y = k$ are given by $(x, y) = (j, k - j)$, $k - n \le j \le n$. Thus $S_k = 2n - k + 1$ in this case. The number of stable tables is therefore

$$\begin{aligned} N &= 1^2 + 2^2 + \cdots n^2 + (n+1)^2 + n^2 + \cdots + 1^2 \\ &= 2\,\frac{n(n+1)(2n+1)}{6} + (n+1)^2 \\ &= \frac{1}{3}(n+1)(2n^2 + 4n + 3). \end{aligned}$$

5. We will show that every vertex of the convex hull of the set of given $2n$ points lies on a balancing line.

Let R be a vertex of the convex hull of the given $2n$ points and assume, without loss of generality, that R is red. Since R is a vertex of the convex hull, there exists a line ℓ through R such that all of the given points (except R) lie on the same side of ℓ. If we rotate ℓ about R in the clockwise direction, we will encounter all of the blue points in some order. Denote the blue points by $B_1, B_2, \ldots, B_n$ in the order in which they are encountered as ℓ is rotated clockwise about R. For $i = 1, \ldots, n$, let b_i and r_i be the numbers of blue points and red points, respectively, that are encountered before the point B_i as ℓ is rotated (in particular, B_i is not counted in b_i and R is never counted). Then

$$b_i = i - 1,$$

for $i = 1, \ldots, n$, and

$$0 \le r_1 \le r_2 \le \cdots \le r_n \le n - 1.$$

We show now that $b_i = r_i$, for some $i = 1, \ldots, n$. Define $d_i = r_i - b_i$, $i = 1, \ldots, n$. Then $d_1 = r_1 \ge 0$ and $d_n = r_n - b_n = r_n - (n - 1) \le 0$. Thus the sequence $d_1, \ldots, d_n$ starts nonnegative and ends nonpositive. As i grows, r_i does not decrease, while b_i always increases by exactly 1. This means that the sequence $d_1, \ldots, d_n$ can never decrease by more than 1 between consecutive terms. Indeed,

$$d_i - d_{i+1} = (r_i - r_{i+1}) + (b_{i+1} - b_i) \leq 0 + 1 = 1,$$

for $i = 1, \ldots, n-1$. Since the integer-valued sequence $d_1, d_2, \ldots, d_n$ starts nonnegative, ends nonpositive, and never decreases by more than 1 (so it never jumps over any integer value on the way down), it must attain the value 0 at some point, i.e., there exists some $i = 1, \ldots, n$ for which $d_i = 0$. For such an i, we have $r_i = b_i$ and RB_i is a balancing line.

Since $n \geq 2$, the convex hull of the $2n$ points has at least 3 vertices, and since each of the vertices of the convex hull lies on a balancing line, there must be at least two distinct balancing lines.

6. For the upper bound, let p be the smallest integer such that $10^p \geq n(n+1)/2$ and let

$$S = \{10^p - 1, 2(10^p - 1), \ldots, n(10^p - 1)\}.$$

The sum of any nonempty set of elements of S will have the form $k(10^p - 1)$ for some $1 \leq k \leq n(n+1)/2$. Write $k(10^p - 1) = [(k-1)10^p] + [(10^p - 1) - (k-1)]$. The second term gives the bottom p digits of the sum and the first term gives at most p top digits. Since the sum of a digit of the second term and the corresponding digit of $k-1$ is always 9, the sum of the digits will be $9p$. Since $10^{p-1} < n(n+1)/2$, this example shows that

$$f(n) \leq 9p < 9\log_{10}(5n(n+1)).$$

Since $n \geq 2, 5(n+1) < n^4$, and hence

$$f(n) < 9\log_{10} n^5 = 45\log_{10} n.$$

For the lower bound, let S be a set of $n \geq 2$ positive integers such that any nonempty $X \subset S$ has $s\left(\sum_{x \in X} x\right) = f(n)$. Since $s(m)$ is always congruent to m modulo 9, $\sum_{x \in X} x \equiv f(n) \pmod 9$ for all nonempty $X \subset S$. Hence every element of S must be a multiple of 9 and $f(n) \geq 9$. Let q be the largest positive integer such that $10^q - 1 \leq n$. Lemma 1 below shows that there is a nonempty subset X of S with $\sum_{x \in X} x$ a multiple of $10^q - 1$, and hence Lemma 2 shows that $f(n) \geq 9q$.

LEMMA 1. *Any set of m positive integers contains a nonempty subset whose sum is a multiple of m.*

Proof. Suppose a set T has no nonempty subset with sum divisible by m. Look at the possible sums mod m of nonempty subsets of T. Adding a new element a to T will give at least one new sum mod m, namely the least multiple of a which does not already occur. Therefore the set T has at least $|T|$ distinct sums mod m of nonempty subsets and $|T| < m$. ∎

LEMMA 2. *Any positive multiple M of $10^q - 1$ has $s(M) \geq 9q$.*

Proof. Suppose on the contrary that M is the smallest positive multiple of $10^q - 1$ with $s(M) < 9q$. Then $M \neq 10^q - 1$, hence $M > 10^q$. Suppose the most significant digit of M is the 10^m digit, $m \geq q$. Then $N = M - 10^{m-q}(10^q - 1)$ is a smaller positive multiple of $10^q - 1$ and has $s(N) \leq s(M) < 9q$, a contradiction. ∎

Finally, since $10^{q+1} > n$, we have $q + 1 > \log_{10} n$. Since $f(n) \geq 9q$ and $f(n) \geq 9$, we have

$$f(n) \geq \frac{9q+9}{2} > \frac{9}{2}\log_{10} n.$$

Weaker versions of Lemmas 1 and 2 are still sufficient to prove the desired type of lower bound.

CONTENTS

ARTICLES

NOTES

PROBLEMS

REVIEWS

NEWS AND LETTERS

THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, NW
Washington, DC 20036