"It’s like... you know": The Use of Analogies and Heuristics in Teaching Introductory Statistical Methods

Michael A. Martin
Australian National University

Journal of Statistics Education Volume 11, Number 2 (2003),
ww2.amstat.org/publications/jse/v11n2/martin.html

Copyright © 2003 by Michael A. Martin, all rights reserved. This text may be freely shared among individuals, but it may not be republished in any medium without express written consent from the author and advance notification of the editor.

Key Words: Expectation; Graphical displays; Hypothesis testing; Influence diagnostics; Regression models; Structure mapping; "Teaching-With-Analogies."

Abstract

Students often come to their first statistics class with the preconception that statistics is confusing and dull. This problem is compounded when even introductory techniques are steeped in jargon. One approach that can overcome some of these problems is to align the statistical techniques under study with elements from students’ everyday experiences. The use of simple physical analogies is a powerful way to motivate even complicated statistical ideas. In this article, I describe several analogies, some well known and some new, that I have found useful. The analogies are designed to demystify statistical ideas by placing them in a physical context and by appealing to students’ common experiences. As a result, some frequent misconceptions and mistakes about statistical techniques can be addressed.

1. Introduction

Analogy is a powerful cognitive tool for discovery and learning about new areas based on existing knowledge. In this article, the use of analogy for teaching introductory statistics is proposed and discussed in terms of a framework for analogical thinking. The issue is examined initially in the context of current approaches to teaching statistics: where do the difficulties lie, and how can we address them? The use of analogies is one strategy for bringing statistics into focus for students weighed down by negative expectations of the discipline. The main part of the article is concerned with describing and evaluating many analogies I have found useful in teaching introductory statistics.

Undergraduate students can find statistics impenetrable. Part of this difficulty stems from the fact that, as is the case with most mature disciplines, learning statistics requires assimilating a great deal of jargon.
Statistics’ reliance on jargon for its description belies its real world, practical roots. One of the most difficult aspects of teaching undergraduate statistics is that of motivating students to look beyond the jargon, and to focus instead on understanding statistical thinking. Notions such as null and alternative hypotheses, confidence intervals, test statistics, and sampling distributions are abstractions to students, disconnected from their underlying meanings. Traditional approaches to teaching introductory statistics are often heavily formula-based. Many students at this level find such an approach bewildering as the mathematical expressions add a further layer of abstraction to the already cluttered landscape.

A further barrier to effective statistical learning is that many probability and statistical concepts are counter-intuitive, resulting in misconceptions and errors that lower student confidence that they will ever understand the discipline. Some well known, counter-intuitive results include the Monty Hall problem, the birthday problem, and Simpson’s paradox, but even simpler ideas such as the failure of an average (that is, a sample mean) to truthfully represent a distribution can lead to confusion and despair. Some statistics educators have found that counter-intuitive examples can be very useful tools to motivate students to think beyond superficial paradigms (see, for example Lesser 1994 and Sowey 2001) and explore statistical concepts at a deeper level. Nevertheless, given that statistics is a very practical science, the counter-intuitive nature of some basic statistical concepts can be very discouraging for students. It is therefore important that statistical educators develop new approaches to motivate students to pursue statistics and statistical thinking.

1.1 Some Useful Approaches to Teaching Statistics

A modern approach to statistics teaching focuses on the exploration of real data - the "real world" approach - and it presents statistics as a way of thinking about data. This idea is powerful, and it speaks directly to the necessity for and the importance of statistics as a highly practical discipline. Moore (1997) champions the "data-based" approach to teaching statistics, and his books are delightfully entertaining examples of how statistics can be taught with verve and skill. Nevertheless, for many students, even this motivated practice of statistics is foreign and unfamiliar, a totally new way of thinking that may fail to link with their existing reasoning and decision-making skills.

Another approach to teaching statistics that can be very effective is that of telling evocative stories that intrinsically involve statistical thinking. The best such stories are interesting, topical, and have a moral - for example, failure to use statistics properly usually results in disaster, or, more positively, proper use of statistics often achieves desirable outcomes or correct decisions. Two excellent examples of statistical story-telling are contained in Tufte (1997). First, he told of how good statistical thinking by John Snow in 1854 led to an understanding of how cholera is transmitted. Snow examined maps of a badly affected area of London and isolated a particular water pump (the Broad Street pump) as the center of transmission. Snow’s data were messy and, at times, confusing (for example, no workers from a brewery close to the pump contracted cholera - perhaps because they predominantly drank liquor, not water!). John Snow was a "doctor," not a "statistician," and his thinking was not only good "statistical thinking," it was good visual thinking and good critical thinking. Moreover, although Snow’s explanations on the cause of the cholera transmission transpired to be correct, mass acceptance of them as valid at the time owed as much to the fact that the cholera epidemic was already on the wane when he acted and had the Broad Street pump replaced.

Tufte’s second story related to the 1986 space shuttle Challenger disaster, and how poor - even sloppy - statistical thinking failed to expose reasons that should have stopped the launch going ahead. In that case, the failure was not so much one of poor data but poor exposition. The relationship between O-ring failure and temperature was obfuscated, and hence never adequately explored prior to launch. Even Richard Feynman’s later, memorable demonstration of the link between O-ring resiliency and temperature before a Congressional hearing could be regarded as a poor statistical experiment given the
lack of a statistical control (a fact that Feynman himself recognized - see footnote 48 on page 52 of Tufte 1997).

Story telling as a pedagogical device goes back to cave painting, and its use in statistical teaching goes back at least as far as Huff’s immensely enjoyable book How to Lie with Statistics in 1954 (Huff 1954). More recent, similarly flavored books include those by Campbell (1974), Kimble (1978), and Hollander and Proschan (1984). Each has much to recommend it. Students respond well to interesting and humorous stories, and each of these books is written in a funny, lively, and highly readable style. Articles by Chanter (1983) and Brewer (1989) also describe some successful story telling or anecdotal approaches to motivating undergraduate statistics. Sowey (1995; 2001) describes the mechanism that allows these approaches to be successful as "making statistics memorable," and to a certain extent, the use of analogies is aimed at this same target - "memorability" - as a teaching tool.

1.2 The Case for Analogy

For many of us, learning is most effective when the concepts under consideration can be aligned with our present understanding or knowledge. Analogies allow students to learn intuitively. We can think of using analogies in two ways. First, we can use analogies as a bridge between familiar situations and new situations. The success of this strategy depends on both the students’ understanding of the familiar situations and on the persuasiveness of the argument that draws out similarities between the two situations. The second use of analogies is in representing abstract ideas in terms of concrete or physical structures. The success of the second strategy depends largely on whether the concrete structures employed fall within most students’ life experiences, as well as on the persuasiveness of the argument that maps them to the abstract idea under study.

A simple example in statistics of the first use of analogy is the way a lecturer might introduce the idea of a small-sample $T$ statistic by noting its structural similarity to the familiar large-sample $Z$ statistic. Of course, students who failed to grasp the rationale behind the form of the test statistic in the large-sample case will likely fail to see how the small-sample case works beyond noting the surface similarity between the statistics. In particular, students for whom the large-sample case remains mysterious may find it burdensome to have to remember both the $Z$ and the $T$ statistic.

An example of the second use of analogy lies in how the idea of a standardized statistic might be justified. We might describe a customer going to a store and noting that a price rise seems unusually large in comparison with her usual experience of price fluctuations for that type of item. Then, we can think of how the shopper came to think of the price rise as "large" and realise that the price differential is measured with respect to "standard" behaviour: "large" must be interpreted both in terms of the raw size of "typical" price fluctuations (perhaps as measured by an average) and in terms of how much they usually vary (perhaps as measured by a standard deviation). The notion of a standardized distance then follows as a natural extension of thinking about what people usually mean when they use phrases like "unusually large" to describe behaviour that is not only different in raw size but also falls outside a typical range. The main element in this use of analogy is the focus on a concrete example to which every member of the class can relate. All students have experience in buying objects and basing purchase decisions on price comparisons. By tying their common experience to a statistical notion, the analogy draws relevance to the statistical idea.

The use of analogies in teaching statistical concepts is certainly not new. Perhaps the most commonly known example is the likening of a statistical hypothesis test to the process of a criminal trial in which the "presumption of innocence" plays the role of assuming the truth of the null hypothesis. This analogy seems to have originated with Feinberg (1971), although it was probably used in classrooms before that time. Feinberg’s discussion related in particular to description of Type I and II errors in testing, so it
presupposed some familiarity with the probability-based approach to testing. Bangdiwala (1989) extended the analogy to more completely describe the links between statistical hypothesis testing and the judicial system. The analogy was also described by Chanter (1983) and Brewer (1989), and has found common usage in a large number of introductory statistics texts, including Larsen and Marx (1990), Johnson and Tsui (1998), Kitchens (1998), and Freund (2001), among dozens of others.

The judicial analogy for hypothesis testing is a particularly powerful one, as many of the facets of the legal process have a direct counterpart (map) in the formal statistical procedure. The relevant elements might be summarized as follows:

**Criminal Trial**
- Defendant is innocent
- Defendant is guilty
- Gathering of evidence
- Summary of evidence
- Cross-examination
- Jury deliberation and decision
- Verdict
- Verdict is to acquit
- Verdict is to convict
- Presumption of innocence
- Conviction of an innocent person
- Acquittal of a guilty person
- Beyond reasonable doubt
- High probability of convicting a guilty person
- Mistrial

**Hypothesis Test**
- Null hypothesis
- Alternative hypothesis
- Gathering of data
- Calculation of the test statistic
- Application of the decision rule
- Decision
- Failure to reject the null hypothesis
- Rejection of the null hypothesis
- Assumption that the null hypothesis is true
- Type I error
- Type II error
- Fixed (small) probability of Type I error
- High power
- No equivalent - perhaps the role of data snooping?

The analogy works well because mappings exist between so many of the individual elements making up the respective cases, and because the relationships between the mapped elements are also largely preserved. Moreover, the process of a criminal trial mirrors that of a statistical hypothesis test to a large degree. The analogy is also successful because the analogous setting (a criminal trial) is so familiar. Most students already possess a firm understanding of the normal legal process, having doubtless been exposed to its workings numerous times on television and in movies. Nevertheless, this analogy, like all analogies, is not perfect. Some elements in criminal trials with which students are undoubtedly familiar are unmapped. For example, there are no equivalents to ideas like cross-examination, circumstantial evidence, or a mistrial in statistical hypothesis testing. Students seeking to find maps for these elements may think testing an even more complicated notion. Also, even though statistical elements like Type I and Type II errors have quite good maps in their equivalent judicial metaphors, the metaphor does not incorporate the idea of sampling distributions which are present in the formal definitions of these errors. Nevertheless, although students often find the idea of hypothesis testing a rather complicated and formal one, the fact that it is mirrored so closely by a very common process suggests to them that it is neither as arcane nor as arbitrary as it may first appear.

The analogies described in the article have all been used in classes I have taught over the last few years. I have found them to be effective in the face-to-face setting of the lecture theatre, but they could also be used in other settings, such as in a distance education setting. In that context, the analogies themselves could be augmented with pictures or, where possible, multimedia elements. Illustrations have been provided for several of the analogies presented later. The main issue in all formats of presentation is that
the mappings between elements and relationships of the source and target domains be clearly described.

1.3 Structure of the Article

This article is structured as follows. In Section 2, a theoretical framework within which to consider analogy is described, against which individual analogies will be later compared. Sections 3 and 4 describe some experimental work carried out in the cognitive sciences discipline that demonstrates the usefulness of analogies in teaching science, including mathematics and statistics, and which in particular suggests that the use of analogy is most effective in introductory courses. In the remainder of the article, several analogies found useful in teaching introductory statistics will be presented and evaluated in terms of the framework described in Section 2. In the main, the examples discussed here correspond to the second use of analogies described above. Some of the analogies are very strong, mapping both individual elements and preserving relationships between the elements with the original idea. Other analogies are weaker, failing to correspond closely to the original setting if one pursues the analogy too far. Recognising when an analogy fails is also an important part of teaching, since the failure of the analogy must be revealed to the students lest they unwittingly accept the analogy as a complete representation of the new abstract idea. As indicated earlier, some of the analogies here are old and well known, and where possible their origin is cited; other analogies are new. Without exception, however, I have found them all to be useful, and the goal of this article is both to construct a list of useful analogies and to remind teachers of statistics of the power of this device in assisting discovery.

2. Analogical Thinking as a Cognitive Tool - A Framework for Describing Analogy

The use of analogies as a cognitive tool can be found in almost all societies and cultures, and at almost any time in human history. Analogical thinking is a way of understanding new or novel situations by building models based on existing knowledge of familiar scenarios. The new situation may, at least superficially, be quite unlike the familiar domain, but relationships among elements of the new scenario may closely resemble relationships shared by elements of the familiar domain. The familiar domain is referred to as the source analog, the new domain as the target analog. There is an extensive literature, both theoretical and empirical, in the cognitive science discipline on the use and effectiveness of analogies for enhancing cognition, memory, and learning. An excellent summary of the modern view of analogical thinking can be found in Holyoak and Thagard (1995). Another great collection of current ideas on analogy is the series of articles by Gentner and Holyoak (1997), Holyoak and Thagard (1997), Gentner and Markman, and Kolodner (1997) in the 1997 volume of American Psychologist. Here I will summarize the main features of analogical thinking that impact its use in a learning environment, as well as describe some experimental studies demonstrating that the use of analogies is effective in teaching contexts, particularly of science, mathematics, probability, and statistics.

The modern study of analogies began with Hesse (1966), in which she argued that analogies were very useful tools for discovery and conceptualization of scientific principles. Within the cognitive science discipline, the predominant theoretical framework within which analogy has been studied is Gentner’s (1983) structure-mapping theory. Gentner’s theory recognizes objects, attributes, and relationships in each of the source and target domains. Her argument is that similarity only at the level of objects or attributes (superficial similarity) is insufficient for successful analogy, but rather that similarity at the level of relationships (structural similarity) is at the core of analogical thinking. Holyoak and Thagard (1995) discuss three main constraints on analogical thinking: similarity (common properties of objects or attributes); structure (parallels between relationships in source and target domains); and purpose (goals that drive the need for analogical thinking in the first place - often simply a desire to understand an unfamiliar situation). Holyoak and Thagard describe these constraints as flexible rather than rigid,
interacting with and perhaps overriding one another in a potentially complex way to reach a satisfactory
cognitive conclusion. The inclusion of purpose as a constraint on analogical thinking reflected Holyoak
and Thagard’s concern for the use of analogies for problem solving within the target domain rather than
merely for understanding the target. In particular, they highlighted four common purposes for the use of
analogy in scientific domains: discovery, development, evaluation, and exposition. The approach of
Gentner and the later, broader approach of Holyoak and Thagard, form the basis for the modern view of
analogy as a cognitive tool.

Current views of analogical thinking break the process down into four main parts:

- access - one or more relevant source analogs must be accessed from memory;
- mapping - correspondences must be drawn between objects, attributes and, particularly,
  relationships in the source and target domains;
- evaluation - inferences drawn arising from the mappings generated in the second part must be
  evaluated and adapted where necessary to take into account unique features of the target domain;
- learning - new knowledge about the target domain is adopted, new items are added to memory,
  and new understanding resulting from analogical thinking allows easier future access of source
  analogs.

The interplay of similarity, structure, and purpose in analogical thinking is a complex one. For instance,
several researchers (see Gick and Holyoak 1980, 1983; Keane 1987; Gentner, Ratterman and Forbus
1993, and Catrambone 1997) have demonstrated that while superficial similarity is a critical factor in
remembering or retrieving source analogs, structural similarity - similarity between underlying
relationships - is most important in applying analogies for learning about the target domain. Novick
(1988) discovered that subjects tend to choose structural features as the basis for remembering (and
evaluating) if they already have training in an area related to the target domain. Blanchette and Dunbar
(2000) found that the primacy of superficial similarity in driving recall in laboratory studies on analogy
was not necessarily replicated in real world settings, where people tend to generate analogies based on
deep structural features. The results of Blanchette and Dunbar reinforce those of Novick (1988) to the
extent that analogies developed in real world settings are often generated by people with at least some
knowledge of the target domain. Central to the study of analogies in all research has been the importance
of carefully "mapping" elements, both superficial and structural, of the source and target domains, and
carefully evaluating these maps to discover the "depth" of the mapping. Superficial maps tend to lead to
"memorability" - deep, structural maps tend to lead to more "useful" analogies in terms of the goals of
understanding and development of new knowledge.

3. Do Analogies Work? Evidence from Some Empirical Studies

Apart from those described in the preceding section, many other experimental studies have been
reported in the cognitive sciences literature. They investigate the efficacy of analogy both as a cognitive
tool, and more specifically for use in teaching science, mathematics, probability and statistics. In this
section, I describe some of this research. This discussion is by no means exhaustive - the cognitive
science literature contains an enormous amount of research in the broad area of understanding analogical
transfer. Some of it has been directed at the question of the effect of prior knowledge of a target domain
on peoples’ ability to generate and use analogies to enhance or develop knowledge in that domain.
Schustack and Anderson (1979) conducted experimental studies whose outcomes suggested that
providing subjects with a good analogy to prior knowledge both improved their ability to recall that
knowledge but also biased them towards building false models for statements consistent with their prior knowledge. They concluded that the use of analogy based on prior knowledge paid off in terms of improved memory, but argued that learning by analogy was only likely to be effective if there is a close, structural relationship between the prior knowledge and the new material. Work by Wharton, et al. (1994) argued further that although analogs having primarily superficial similarities to the target were easily recalled, analogs whose similarities were structural rather than superficial were also accessible. That is, they found that people are capable of making a "mental leap" to a superficially different source analog in the presence of parallel high-order relationships between the source and target.

Several researchers have considered the use of analogies in teaching science. Holyoak and Thagard (1995, p. 199) note that "teaching science and technology... is a challenging endeavor. One basic reason is that the material being taught will often be very unfamiliar to students, and may even contradict erroneous beliefs they have acquired in everyday life. Analogy provides a potential tool for 'jump starting' students by introducing unfamiliar target domains in terms of more familiar source analogs." They go on to argue that the strengths and weaknesses of particular analogies can be understood in terms of the constraints of similarity, structure, and purpose, something we attempt below in describing our own analogies.

Donnelly and McDaniel (1993) carried out a series of experiments whose outcomes suggested that the use of analogies in teaching science students was most effective when the learners had minimal background knowledge of the scientific area being taught. They concluded that analogies are most effective in introductory courses, and that one benefit of analogies lay "in their ability to foster more accurate inferential thinking about the studied material." While this finding is important because of the relevance placed by teachers on inference-level knowledge, Donnelly and McDaniel noted a concomitant reduction in subjects’ ability to remember specific details about the target domain. In later work, Donnelly and McDaniel (2000) noted that for "knowledgeable learners" (subjects with good background knowledge of the target domain) the use of analogies was effective when new material was presented in a way that made application of previous target-specific knowledge difficult. In this case, the learners behaved effectively like novice learners, as their previous knowledge had been rendered less useful in addressing the new problem. But in cases where the new material was easily related to existing knowledge, they found that analogies even hindered learning. They suggested that this phenomenon might occur because analogies can transfer incorrect information (see, for example, Spiro, Feltovich, Coulson, and Anderson 1989), and in attempting to integrate the analogy with their target-specific understanding, subjects are confronted with inconsistencies that make learning more difficult.

Novick and Holyoak (1991) reported the results of an experimental study conducted to assess how mathematical word problems may be solved by analogy to previous, similar problems. The results of their research suggested that students found little difficulty in mapping between source and target analogs, but that adapting the source solution to meet the requirements of the specific target presented the major difficulty for students. Most interestingly, they discovered evidence that successful use of analogy was accompanied by the creation of a schema for the type of problems represented by the source and target problems. The implication of this finding is that students’ acquired abstract knowledge of how to solve a particular class of problems renders them more likely to be able to solve such problems in the future.

In the specific context of statistics education, Quilici and Mayer (1996) investigated the role of examples in how students learn to categorize statistics word problems. They discovered that students given access to a solved example problem were more likely to solve later word problems by analogy, correctly categorizing the new one on the basis of structural similarities to the source. Students who received no sample problems were more likely to sort problems on the basis of superficial similarities (such as features of the story on which the problem was based) than on structural similarities. Hence, they more
often chose an incorrect procedure for solving the target problem.

4. "The Teaching-With-Analogies" Model

In a series of articles, Glynn and co-authors introduced the "Teaching-With-Analogies" model as a structured approach to building on traditional educational tools. The model was first proposed by Glynn (1991), and discussed and further refined by Glynn and Law (1993), Glynn (1994, 1995, 1996), and Glynn, Duit and Thiele (1995). It consists of six steps:

- Introduce the target concept
- Access the source analog
- Identify relevant features of the source and target
- Map similarities (both superficial and structural) of source and target
- Investigate where the analogy breaks down (for example, examine unmapped differences)
- Draw conclusions.

The six steps broadly address the four principles of access, mapping, evaluation and learning described earlier. The research described in Sections 2 and 3 suggest that care should be taken to focus on deep, structural similarities where possible. The notion of examining unmapped differences is particularly important also, and leads to what Holyoak and Thagard (1995) term *adaption*, a key part of the evaluation phase.

In the following sections, a number of analogies used to introduce statistical techniques to undergraduate students are discussed. For the most part, students undertaking courses at first- or second-year undergraduate level might be regarded as "novices" rather than knowledgeable learners, so the findings of Donnelly and McDaniel (1993) are particularly relevant.

5. Graphical Displays of Data - Painting a Picture

Statistical graphics that summarise one-dimensional data are often taken for granted by instructors as trivial exercises. Yet students often struggle with issues such as how histograms differ from bar charts, and why boxplots look the way they do. To many students, the construction of such graphics is the result of a set of essentially arbitrary rules, and so their understanding of the graphics is hampered by the rules governing their construction.

Most students are already familiar with bar charts and pie charts when they first come to their introductory statistics class. Frequency histograms are seen as simply another example of a bar chart. As a result, students frequently make simple mistakes like separating the bars of the histogram, or giving little or no thought to issues like bin width. The following analogies are designed to help students in developing judgment about "good" choices for bin width and number of bins in drawing histograms. The analogies here are tools to aid in visualization rather than deep analogues to the meaning of histograms as estimators of densities, and we caution readers that the "maps" between source and target analogs in this case are primarily superficial. Nevertheless, I have found these analogies useful as guides to histogram construction.

A physical analogy that works nicely to describe the construction of a frequency histogram is that of an apple-grading machine. A conveyor belt carrying apples runs past a series of slots of increasing size sitting above a collection of bins. When an apple can fit through a slot, it drops into a bin, and the apples are thus sorted by size into ordered categories. The number of apples in each bin is analogous to the height of the corresponding bar of a histogram. The fact that the bins are ordered by the size of apples
they contain suggests a set of bins that must completely span the range of possible sizes. If one changes
the focus from the number of apples in each bin to the proportion of apples in each bin, the analogy also
applies to relative frequency histograms. This analogy works quite well in describing the process of
setting up a histogram given a particular set of bins, but it does not really address the issues of number
of bins or bin width.

These issues can be considered in the context of different apple sorting machines with different numbers
and sizes of slots. If the slots are too numerous and the slot sizes rise in increments too small, the
resultant grading is too fine to be commercially useful - customers will be unable to easily differentiate
between apples in consecutive bins and will be confused by the large number of apparently similar
choices. If the slots are too few and the slot size increments too large, apples of very different sizes will
be graded as equivalent, again leaving customers dissatisfied as they find small apples in their box
labelled "large!"

Another analogy that is useful for describing the effects of bin width and number of bins is that of
focussing a camera. Frequency histogram representations of data usually involve loss of information as
observations that are different can be categorized as belonging to the same class. By way of analogy, the
process of setting the focus on a camera can be likened to the problem of setting bin width in a
histogram. Blurry, or out of focus shots tend to lose too much information, and can be likened to the
case where overly large bin widths are chosen. At the other end of the scale, photographs taken with
excellent focus on high-speed film are likely to reveal too much detail - with every blemish and wrinkle
evident - and can prevent us seeing the "big picture." The common phrase "cannot see the wood for the
trees" is also helpful in describing this phenomenon. For histograms, bins that are too small tend to
reveal spurious modes. A good exercise that brings out this point nicely is to produce photographs at
several focus levels and to ask the class which one best conveys the necessary information. Students are
likely to be very good at deciding which photo looks best, since most people have extensive life
experience of viewing photographs. A similar exercise can then be carried out looking at histograms of a
single set of data but with varying bin widths. By drawing on their past experiences with photographs,
students can then learn that it is a similar process that guides them in constructing histograms with
reasonable bin widths.

Neither the sorting machine nor the camera focussing analogies yield much insight into the idea that
histograms drawn on the density scale can be interpreted as density estimates. As a result, neither
analogy explains the nature of a histogram in describing data. Nevertheless, each focusses on a
particular element of histogram construction (number of bins, bin width), and motivates the need for
wise choices of design elements such as these. A common problem for beginning statistics students is
the perception that "rules" for selecting graphical parameters merely reflect arbitrary design choices.
These analogies suggest that such choices are far from arbitrary; rather they reflect preferences or
judgments similar to those made in the physical examples. Ultimately, the analogies yield the principle
that in choosing the number of bins and the bin width in a histogram, the choices must strike a balance
between having too many small bins (undersmoothing, leading to jagged or spiky graphics) and having
too few, large bins (oversmoothing, where even prominent data features may go unnoticed). By thinking
of the physical analogy, students can make statistical judgments on the basis of how they would react to
the analogous physical situation.

Another graphic with which students often have difficulty is the boxplot. The construction of a boxplot
is very simple in theory, but students tend to have trouble with remembering how far the whiskers
extend from the quartile hinges. Some students draw whiskers that extend all the way to the minimum
and maximum, others always extend them to exactly 1.5 interquartile ranges from the hinges, while
others correctly end the whiskers at the most extreme data point within 1.5 interquartile ranges of the
closest quartile. But getting all the students to draw the same (correct) thing has been a constant
The difficulty students encounter is that they see the rule for how the whiskers are drawn as a somewhat arbitrary design feature of an unfamiliar graphic, and so it is difficult for them to remember the rule precisely. A physical analogy is helpful to reinforce the way the rule works. Imagine a dog tethered by an elastic cord to a post situated at the third quartile, and represent the data points more extreme than the quartiles as a sequence of increasingly large bones as you move away from the closest quartile. The dog is hungry, and so sets off for the most extreme (that is, the largest) bone. Unfortunately, the elastic cord has a maximum extension of 1.5 interquartile ranges, so if the largest bone is outside that range, the dog cannot reach it. So it returns to the largest bone it can reach, which represents the most extreme point within 1.5 interquartile ranges from the third quartile. The elastic cord then represents the way one should draw the whisker. The analogy works because students know that a rational dog will always stop at the largest bone it could reach, and so it is easy to remember how to draw the whisker. Figure 1, below, depicts the situation for the upper end of a boxplot. The use of an illustration is powerful, and would work well in a textbook or distance education setting where visualization can be an important feature of communicating information. The map between the whisker and the cord and that between bones and data points are reasonably simple, a feature which makes it more likely that students will remember the analogy easily.
**Figure 1.** A boxplot’s upper whisker extends to the largest data point within 1.5 interquartile ranges. Similarly, a hungry dog will settle for the largest bone its lead will allow it to reach. Here, data points beyond the third quartile are represented as a series of increasingly large bones, while the whisker is represented as the dog’s elastic lead.
6. Numerical Descriptions of Data

Most students are comfortable with the idea that an average (also called a sample mean) is a reasonable measure of the location of a data set. Yet, when pressed as to why that is, many students are unable to answer clearly. The physical analogue of taking an average is that of bringing the data into balance. This physical idea is depicted very nicely in Moore (1997, cover and p. 262). Imagine each data point as a one-pound weight situated along a ruler representing the number line. Now imagine trying to balance that ruler on your finger. The position of your finger that allows the ruler to come into balance is the average of the data set. This idea becomes useful also in describing expectation, and, even better, the analogy can help us build a bridge linking sample means and population means by arguing that each represents a balance point of an analogous object.

Students typically accept and understand location estimation fairly easily, but variance estimation can be another matter. Certainly, it is straightforward to convince students of the need to understand and estimate spread, but the formulae for sample variance and sample standard deviation are far from obvious to many. In particular, several elements of the formula for sample standard deviation need explanation beyond the description of how one might actually calculate it given a set of data. Four major elements of the formula

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

are:

1. Why use deviations $x_i - \bar{x}$ in the formula?
2. Why use squared distances rather than raw distances?
3. Why use the divisor $n - 1$ in the formula instead of a simple average? What are degrees of freedom?
4. Why take a square root?

The last element is the easiest to address: once we base the formula on squared distances, the units related to the measure are squared units, so the square root allows the spread estimate to be reported in the same units as the original data. Each of the previous three elements must be carefully motivated.

The use of deviations can be justified through the notion that spread must be interpreted with regard to some reference point, usually the "middle of the data." As an example, if we wanted to see how spread out an object is, we might measure the distance from its left hand edge to its right hand edge, or we might consider how far each edge is from the middle. In the context of a particular data set, the deviations - the signed distance of each data point from the average - implicitly give us a reference point (the average) from which the issue of spread about that point can be addressed. All of the most obvious measures of spread (the range and interquartile range, for example) explicitly involve distance from a reference point (for example, the distance from the smallest data point to the largest and distance from the first quartile to the third quartile), so it is not surprising that we should want to base the standard deviation on such relative quantities.

The next question is: how should the deviations be combined to produce an overall measure of spread? The obvious answer would be to consider the average of the deviations. However, the fact that the average, $\bar{x}$, is in the "middle" of the data suggests that the deviations will cancel one another out (positives balancing negatives) and so their average will be zero. The fact that the deviations must take
on values that are both positive and negative while spread is an implicitly positive quantity leads to the conclusion that deviations must be used in a manner whereby each one must contribute a positive amount to the spread. Obvious candidates for converting negative deviations into positive contributions to spread are absolute deviations or squared deviations, the latter preferred for somewhat abstract and technical reasons - for example, the use of squared deviations can lead to an unbiased variance estimator.

The third question: why use \( n - 1 \) in the formula instead of a simple average of squared deviations? Unfortunately, the divisor \( n - 1 \) is introduced with the somewhat cryptic name "degrees of freedom." This concept is one of the hardest concepts to get across to first-year students. Technical arguments about unbiasedness of variance estimates are rarely successful in convincing students that degrees of freedom is an important concept, and students often have difficulty extending their understanding of degrees of freedom beyond the standard one-sample estimation of variance setting. The geometric argument describing degrees of freedom is mathematically appealing but impossibly abstract for typical first-year statistics students. I have tried a variety of ways to explain the concept: mathematically; by demonstration; and via physical analogy. One simple idea is to randomly select, say, ten students and ask each in turn to select a random integer, either positive or negative, with no other restriction. Each student is "free" to choose whichever number they wish. Now, randomly select another ten students, and ask each in turn to select a random integer either positive or negative, this time with the restriction that the total of all the choices must be exactly 0. This time, the first nine students can select freely, but the final student has no choice at all: their choice can only be the number that, added to the previous nine "free" choices, makes the sum zero. There are therefore only nine "degrees of freedom" remaining in this second exercise. It is natural to wonder where the restriction that the numbers in the second exercise add to zero comes from. The answer is that in the first case students were selecting data points, while in the second they were specifying deviations. Nevertheless, the idea of degrees of freedom remains an elusive one for many students.

A simple physical heuristic can be used to think about how degrees of freedom can be calculated. Suppose that the data set under discussion is of size \( n \). Now think of a notepad with exactly \( n \) pages in it. As previously argued, in order to estimate spread, one must first estimate location, so imagine calculating the average of the set of data and writing the result on the first sheet of paper on the notepad. Once the page has been used to write the average, the page is removed from the pad and stored somewhere. Since location has now been estimated, we are able to estimate spread, and there are \( n - 1 \) remaining sheets of paper in the notepad on which the result could be written. The number of sheets of paper left on which to write the variance estimate represents the degrees of freedom.

The heuristic introduced here is evocative and carries with it other important relationships. The representation of a set of data as a notepad with \( n \) pages reinforces the idea that data are a resource that carries only a finite amount of information, much as a notepad carries only finitely many pages. Moreover, a data set with \( n \) data points allows us to estimate at most \( n \) independent quantities, while the notepad can be used to write only \( n \) quantities, one per page. The idea also works in relating the number of degrees of freedom for estimating spread in settings more complex than a one-dimensional set of data. For example, in a multiple regression setting with a single response variable and two covariates, the error variance is estimated based on model residuals that require the prior estimation of three parameters (an intercept and two slope parameters) that describe the "location" of the relationship between the response and the predictors. Hence, our notepad would have the intercept estimate written on the first page, the two slope estimates written on the next two pages, leaving \( n - 3 \) pages (degrees of freedom) for writing down (estimating) the error variance. Similarly, for a two-sample \( t \)-test, locations must be estimated for each of two populations prior to estimating the common variance: beginning with a notepad having \( n + m \) pages, two are consumed by the sample averages, leaving \( n + m - 2 \). The notion that information, gets "used up" when parameters are estimated is an important feature of this concept. It
demonstrates to students that notions like degrees of freedom do more than provide mathematical convenience to statistics - they also carry with them important ideas like the role sample size plays in statistical decision-making. The fact that information, like note paper, runs out can be made clearer by considering a sample of size 1. In that case, there are no sheets of paper left after location is written down, and, analogously, variance cannot be estimated.

While the note pad heuristic is evocative, it is also limited, and it is important to recognise those limitations. First, the idea is specific to the case of measuring spread in the traditional squared error loss sense, and the main elements do not carry over to other measures of spread. Second, in the notepad analogy there is no essential difference between one sheet of paper and another. As a result, in cases where estimating location involves estimating several parameters (such as an intercept and a slope), students might attribute degrees of freedom \( n - 1 \) to the second estimated parameter (the slope) because it is estimated after the intercept, while the statistical theory refers to degrees of freedom only in estimating a variance. In other words, the heuristic development possesses a symmetry (the pages are all the same) that the statistical situation lacks (variance estimation differs from location estimation in several ways). Students may have trouble understanding why all elements of location must be estimated before spread can be addressed.

One can also easily come up with examples where the idea fails completely, such as using the median to estimate location, or estimating location and spread when the data arise from, say, a Normal\((\mu, \sigma)\) distribution. This idea is most useful in the context of giving students a reliable way to work out degrees of freedom for variance estimators in traditional moment-based, location-scale models. The heuristic is not a good way to explain the concept underlying degrees of freedom as it has no way of representing important issues related to the geometry of the estimation procedure. Nevertheless, the heuristic serves as a very useful way in which students can remember how to calculate degrees of freedom in different circumstances. It is best thought of as an aid to remembering how to calculate degrees of freedom rather than as an explanation of the underlying concept. Moreover, I have found it most useful as simply a small part of a multi-step discussion of the usual standard deviation estimator, to try to motivate students to remember what is, in their first exposure to statistical theory, quite a challenging concept.

7. Standardized Scores - "All Things Being Equal"

One of the most powerful uses of descriptive statistics lies in making comparisons. The concept of standardisation is a useful tool for comparing values arising from different data sets. Strangely, many students are familiar with the term "standardized scores" since they will almost certainly have been presented with standardized versions of exam results during their schooling. However, few actually know what the process of standardisation is supposed to achieve. (A similar phenomenon exists for the idea of "grading on a curve" - students often ask if the course is to be graded on a curve, but cannot tell me what they mean by that phrase.) In fact, some students view standardisation as a sinister plot to depress test scores! A useful heuristic for describing the intention behind standardisation lies in the common phrase "all other things being equal." The idea of this phrase is that comparison between two inherently different objects can be made with reference to a particular characteristic, provided that a majority of other characteristics are the same. For example, this notion is often used by employers in selecting the better of two prospective employees. Of course, in practice it is almost impossible to find two objects or people that are identical in all respects but one. Yet the idea is useful if one is prepared to accept some minor differences in deference to the major point of comparison. In statistics, standardising is a simple way to remove two gross effects from the data set, namely the effects of location and spread. Thereby it renders data sets comparable provided, for example, that the two data distributions are roughly the same shape. Students find that casting standardisation in the context of making a choice between different people applying for a job makes it easier to see the rationale behind the idea.
Moreover, when explained this way, it is easy to see that comparing standardized scores is really just common sense: if two data sets have different locations and spread, then those effects must be accounted for before the relative positions of points can be reliably compared.

8. Probability - A Leap into the Abstract

The progression from empirical descriptions of data to theoretical probability models represents a huge leap for most undergraduates. Probability models might be thought of as existing in a constructed "fantasy" or "model" world, where random behaviour is well understood through theoretical probability distributions. Students often fail to see the relevance of probability models immediately, since they are generally presented as abstractions and most textbooks fail to make clear what connections exist between the real, data-based world and the model world. Indeed, a number of introductory statistics texts now offer the chapter on probability as optional rather than mandatory. Yet statistical inference could not exist without probability models. So it is important for students to appreciate that understanding probability is about understanding randomness, and randomness is at the heart of all applications involving data.

Randomness itself represents a troublesome concept for students. When asked to give examples of "random events" students invariably cite only very unlikely events such as "winning the lottery," or "being involved in a car accident." Although events like "getting out of bed in the morning" or "getting dressed to go to work" are much more common random events, they are usually not regarded by students as "random" simply because they are "ordinary" and "probable." As a result, it is important to stress that probability is about the ordinary as well as about the extraordinary, and that likely events are just as "random" as rare ones. It is also helpful to refer to everyday happenings in terms of random events, as it makes the study of probability a little more concrete. Informally, here we think of a random event as being a member or subset of the sample space.

Some approaches to making probability transparent, such as live experimentation with coins, dice, and cards, can be ineffective when faced with cultural barriers. After what I had thought a particularly evocative and effective lecture involving combinatorial probability with my "props," some multi-sided dice and a pack of cards, a group of students approached and asked, simply, what a pack of cards actually was! Gambling was forbidden in their culture, and so they had no idea, for example, what the significance might be of Kings, Queens, or Jacks.

In describing the idea of expectation both ways of using analogies outlined in the introduction are useful. The idea of expectation as a "balancing point" for the probability distribution is analogous to thinking of the sample mean as a "balancing point" for the data distribution. This analogy is strong, since in a frequentist sense probability distributions can be thought of as limiting cases of data distributions as sample sizes grow to infinity, while the sample mean converges to the expectation as sample size grows. Of course, students who failed to understand the balancing point analogy in the case of a sample mean will likely also fail to understand it in the case of expectation. Hence, a physical analogy is also useful in this case, and is best discussed in the context of a simple example. Suppose we wished to find the expectation of a probability distribution having mass 1/4 at 0 and mass 3/4 at 1. By way of analogy, imagine that as a child you weighed one-third of your brother’s weight, and that you and he sat on a seesaw (See Figure 2). How much closer to the middle of the seesaw must your brother sit in order for you and he to balance? The answer is that he must sit one-third the distance you are sitting from the middle; if he is further than that, the seesaw will tip his way, while if he is closer, the seesaw will tip your way.

Most students can understand this example very easily, having played on seesaws as children. Then they
can see that the probability distribution above must have its expectation at 3/4 so that the heavier mass is one-third of the distance from the balance point as the smaller mass. Also, an important principle has been established: in order to bring the seesaw (distribution) into balance, the product of distances and weights (probabilities) on one side of the balance point must equal the product of distances and weights on the other side. In this way, the formula for expectation can be developed naturally through the physical analogy of the seesaw:

$$E(X) = \sum xP(x) \text{ is like } \sum \text{distance} \times \text{weight}$$

The extension to more complicated distributions, including continuous distributions, is more delicate. Nevertheless, the analogy is a strong one in that it not only replicates the behaviour of the statistical situation, but it also suggests why the formula for expectation looks the way it does.

**Figure 2.** Expectation as child’s play! A distribution sitting atop a seesaw (bottom picture) must obey the same law of physical balance as the children sitting atop the seesaw in the top part of the figure.

Sometimes an imperfect argument is acceptable if it can get an important message across, although care must be taken to stress the limitations of the argument. For example, one of the most common errors undergraduates make is to guess that if \(X\) and \(Y\) are independent random variables then \(\text{Var}(X - Y)\) equals
Var($X$) - Var($Y$) instead of Var($X$) + Var($Y$). A simple way to explain why variances compound rather than subtract in this case is to ask the students to pick a number between 0 and 10 and then, independently, pick a second number between 0 and 10. Now consider the difference between those two numbers. The smallest possible value for the difference is -10, while the largest is 10. So the range of possible values for the difference is 20, equal to that for the first number picked plus that for the second number picked. So, the "variability" (here represented by range) of the difference is clearly larger than either of the original ranges.

The heuristic does not explain why the correct formula actually holds for variances, nor does it suggest the correct formula (except by luck). Nevertheless, it does indicate which of the two formulae above for Var($X - Y$) is the more likely to be correct, and often that is what students really need to know. The important realisation in applying this line of argument is that teaching students the right formula can be achieved on several levels: some members of the class will derive the result mathematically and be perfectly happy; others will need to simply remember the right formula. This heuristic gives students a way to remember the formula correctly, so it assists their learning in helping prevent an obvious blunder. The idea that variation "adds up" is an important one. Of course, spread only "adds up" literally in the case of variance, and not in that of standard deviation, but the fact that the end result is larger than either of the original spreads is an important one.

9. Clothes Maketh the ... Distribution?

An imperfect analogy also works quite well in explaining the process of fitting data to a specific distribution. Sometimes it is difficult for students to make the link between theoretical probability distributions and data descriptions. Therefore, the idea of using data to estimate the parameters of the model distribution is foreign. A simple physical analogy is to think of the family of distributions like a rack of clothes of varying sizes, and to think of a customer trying to find an outfit as representing the data. The goal is to choose the clothes (distribution) with the best fit to the customer (data). If the model distribution has only one unknown parameter, that parameter is like the size of the clothes, while if there are multiple parameters, one can think in terms of separate sizes for the shirt, pants, shoes, and so on. The analogy helps students see the abstract idea of fitting a theoretical distribution to data in concrete, familiar terms. They can be encouraged to think about the consequences of choosing parameters to be too large or too small in terms of how well the distribution fits. In this sense, the analogy is good because it involves a physical situation with which all students will be very familiar.

The weakness in the analogy is that by their nature clothing sizes are discrete, so one has to imagine a continuum of clothing sizes if one wants the analogy to transfer directly to the theoretical situation where parameter values are not restricted to a discrete set. By having to think of a continuum of sizes, we are forced into using an abstraction that can make the example less compelling for some students. I prefer to use the imperfect analogy so that it has maximal impact. Also, care must be taken in saying what is meant by the expression that the clothes "fit." This issue applies in the statistical context as well, since a variety of fitting methods is usually possible.

10. Estimation and Variation

One of the issues that students often find confusing is the dual role the average plays as both a number and a random variable. The problem arises mainly because most of them have never thought of the average as behaving randomly - how can it when it is the result of a formula? The answer, of course, lies in sampling variability, but this concept as well as the idea of repeated sampling is not natural to students. Why? Perhaps the answer to this question lies in the way non-statisticians think about data. Most people accept easily that large samples lead to better inferences, and when pressed to say why they
believe this to be true a common response is "because I have more information in a larger sample." As statisticians, we might rephrase this comment to "because there should be less variability (in the average) from larger samples." Non-statisticians do not necessarily equate "more information" with "less variability." They simply do not think about how the average might behave if the sampling were repeated a large number of times.

How can we get this idea across? Moore (1997, p. 262) draws an excellent analogy between repeated sampling and target shooting. Think about an archer shooting arrows at a target off into the distance. Each of the archer’s arrows represents a data set; the centre of the target represents the unknown parameter, and the shot from his bow the estimation process. To simulate the usual situation, the archer would have only a single shot at the target using his only arrow (the data). If the archer is somewhat skilled, he should be able to shoot his arrow close to the centre of the target, and the closer he is to the target the more accurate you might expect him to be. His distance from the target represents sample size: the smaller the sample, the further the archer is from the target, while for very large samples he is close to it. By thinking of supplying the archer with a full quiver of arrows, we have an analogy for the idea of repeated sampling. The key features of an estimator might be summarized by thinking about the outcomes of several shots at the target, as in the following picture, which is essentially my own rendering of Moore’s (1997, p. 262) diagram.

![Figure 3. Estimation Analogy - Shooting Arrows at a target.](image)

The fact that in the usual situation the archer has only one arrow is not all that strange an idea for
students. They can see how one could comment on the accuracy (or bias) and precision (or variance) of the archer’s shots by thinking about what he might do with multiple arrows. The notion of repeated sampling, though rarely done with real data, can therefore be considered easily.

Moving from point estimation to interval estimation requires only a slight modification to the above analogy. Now imagine that the archer carries arrows with an exploding tip, so that when they hit the target they remove an area of the target rather than simply a point. The confidence level can be thought of as analogous to the amount of explosive painted on the arrow’s tip, while again distance from the target represents sample size. This analogy is very evocative, but not quite as effective as the preceding one, since the size of the area removed from the target should depend not only on the quantity of explosive (confidence level), but also on the standard deviation of the data set (no analogue). Nevertheless, the analogy forces students to think about different modes and goals of estimation, and so it can still be used reasonably effectively.

11. Testing Times - Hypothesis testing

Hypothesis testing is probably the most daunting topic for undergraduate statistics students to understand, mainly because the related terminology is so rich. As a result, students come to regard the area as unnatural and difficult, while to some extent it mirrors everyday decision-making processes reasonably closely. Of course, the reason statistical hypothesis testing is couched in such a rich a language is that it differs from everyday decision-making in some important respects. Despite this fact, the parallels that exist between students’ everyday experiences of acting on the basis of evidence and formal statistical testing are strong enough to provide excellent motivation to understand the hypothesis testing process. I have already discussed the legal analogy for hypothesis testing. It is an excellent one both because its elements and relationships map to those of the statistical procedure so well, and because students are very familiar with the language and features of the criminal justice process. Nevertheless, some of the more formal elements of hypothesis testing remain difficult for some students, and so another analogy is useful in helping to explain the decision process.

One of the most confusing elements in hypothesis testing is the description of the ultimate decision in terms of either rejecting the null hypothesis or failing to reject the null hypothesis. A common error made by students is to speak rather of accepting or rejecting the null hypothesis, and most texts go to great pains to reflect on the difference between accepting a hypothesis and failing to reject it. The issue of what language should be used in describing the decision can be largely resolved if we take a different view at the outset of the hypothesis test. Instead of thinking in terms of a null and alternative hypothesis, think of two competitors entering a race. Clearly, one of them will emerge victorious over the other, yet a third competitor may beat both. So, we can think of the null and alternative hypotheses as competitors for our decision, keeping in mind that neither may be actually true, but that one will be more supported by the data than the other. Thought of in this way, the whole issue of the difference between "acceptance" and "failure to reject" is avoided. The terms of the contest clearly state that one of the two competitors will be chosen as the "winner" of the contest between them, regardless of how other race participants might perform. Our decision then becomes one of which of the null and alternative hypotheses do we prefer given our data, starting with the premise that the null hypothesis is right (the "favorite"), rather than deciding the absolute truth or otherwise of either hypothesis.

In terms of the legal analogy, this idea simply means that the defendant will be either acquitted or found guilty on the basis of evidence (leaving aside possibilities like hung juries and mistrials), and presuming innocence at the outset. Since only two outcomes are possible in this case the issue of "accepting" innocence versus "failing to reject" innocence doesn’t really arise. Of course, incorrect decisions will be made from time to time, but the statement of the decision rule reflects the lack of certainty through the
Type I error. In the competition analogy this is the chance that the competitor chosen as the winner was not in fact the better of the two.

As pointed out earlier, it is important not to think of any analogy as a "replacement" for the statistical process, as the analogy does not map perfectly to the statistical situation. In the competition example, the analogy fails to grapple with the probabilities of Type I and Type II errors and the relationship between them. Moreover, research by Batanero, Godino, Vallecillos, Green, and Holmes (1994) reports on several fundamental misconceptions commonly held by students about hypothesis testing (see also the article by Shaughnessy (1992)), which this analogy does not specifically address. In particular, according to their research, some students believe that the probabilities of Type I and Type II errors must add to 1. This belief may be reinforced by the structure of the competitor analogy wherein the test is similar to a contest between two exclusive and complementary outcomes. Nevertheless, the analogy is useful in promoting the pragmatic view that hypothesis tests invariably involve a choice between two states, neither of which may be actually correct. It also helps in preventing the misconception that statistical significance is the same as practical significance because most students’ experience of competition will include several examples of cases where the "best" competitor actually lost a contest.

Another instance in which the competitor analogy is useful is in showing students that hypothesis tests should not be interpreted sequentially to make overall inferences. For example, in a sporting context in which three teams are involved in a sequence of games between pairs of teams, the outcome of Team 1 beating Team 2 and that of Team 2 beating Team 3 do not together guarantee that when Team 1 plays Team 3, Team 1 will emerge victorious. Similarly, it is dangerous to combine the results of a sequence of hypothesis tests to make an overall conclusion.

Finally on the matter of hypothesis testing, it is important not to identify the process too closely with everyday decision-making, mainly because of some of the problems human psychology poses for the latter. Efraim Fischbein was among the first researchers to recognize the importance of considering human psychology when looking at approaches to teaching statistics. Fischbein (1975, 1987) are seminal works in the area. Two of his ideas are particularly relevant in the current context. First, each of us in making everyday decisions examines the relevant (and sometimes irrelevant) evidence through a prism of our own experiences and beliefs. Where statistical hypothesis testing is designed to be an objective assessment of data-based evidence, our own decisions are often highly subjective, so any alignment between the two settings is likely to be fraught with risk. Second, many people fail to adequately understand or accept the existence of random events, and this failure leads to such events being interpreted as deterministic rather than random. This interpretation appears, at least in part, to be due to the psychological need people have to find rational reasons for most events in their lives. As a result, students’ intuitions can interfere with their ability to think statistically. In the case of hypothesis testing, these issues can manifest in a number of ways. For example, decisions may be interpreted incorporating students’ own experiences and potential biases, rather than solely on the basis of the data. This outcome is particularly likely when the subject of the test may initiate an emotional or mass media-driven response. Consequently, it is particularly difficult to teach students good statistical thinking by focussing on real world examples where the questions posed are likely to be very topical or controversial. By linking the formal process of hypothesis testing too closely with real world decision-making, we risk opening a Pandora’s Box of potential biases and differences of perception, rendering the decision-making process even more complex. Nevertheless, finding real world analogs for elements of hypothesis testing remains a useful tool for convincing students that the statistical thinking underlying testing is rational and far from arbitrary.

12. Beyond the Introductory Course
Analogies remain a very useful tool in more advanced statistics courses. As statistical concepts to be taught become more advanced, good physical analogies can allow students to relate more easily to the statistical technique. For example, formal notions like the Analysis of Variance are not natural to many students, but the idea of dividing up an object into several parts is, so equivalences drawn between the two processes are evocative and powerful.

12.1 Regression - The Undiscovered Island

In simple one-sample problems it is usually unnecessary to present estimation and inference in terms of formal models. It is much easier and more intuitive to speak of a univariate population as having a mean parameter \( \mu \) without having to formally write the model \( X_i = \mu + \varepsilon_i \) for \( i = 1, \ldots, n \), with \( \varepsilon_i \) distributed as independent and identically distributed Normal(\( \mu, \sigma^2 \)) variates. In regression settings, however, the model is important since it is central to understanding the relationship between the predictor and the response. Of course, it is easy to enhance our explanations of linear regression by drawing pictures. Ideas like least-squares can be motivated by looking at a scatterplot of the data and thinking about how a "best" line might be fit to the data.

Another difference between the case of testing a location parameter in the standard univariate setting and the Analysis of Variance in a regression setting is the fact that in the univariate case, tests for the parameter of interest are usually based on location-scale type statistics. These can be motivated by thinking about how distance between the estimator and the model parameter can be measured. For regression settings, tests are commonly based on the amount of variation explained by the model compared with the total variability inherent in the response. This type of idea leads to a different type of test statistic, an \( F \) ratio. The \( F \) ratio is somewhat harder to motivate than, say, a \( t \) statistic for testing a mean, mainly because the "distance" measure used for the \( F \) statistic is a ratio rather than a more "standardized" measure of distance. As a result, a physical analogy can be very useful in describing it. An excellent analogy is that of a signal-to-noise ratio. Think of a radio station broadcasting a program, and think about listening to that station on your radio some distance away. In between the station and your radio there may be considerable interference, signal degradation, and so on, that means that the signal you hear on your radio is only an approximation to the signal the radio station broadcast. In this particular case, the model fit is very simple: you expect to hear exactly what the station broadcast plus some noise. So a good test of whether this model fit is a reasonable one is the ratio of signal to noise, a common idea from engineering applications. Most people can see quickly that a large signal-to-noise ratio is indicative of you hearing a close approximation to what the station broadcast, while a low one suggests that you will be hearing mainly static. The idea of a ratio estimator to measure the quality of a fit seems quite natural in the context of this analogy, as the comparison makes intuitive sense. Although this analogy is not a particularly strong one in terms of explaining how the \( F \) ratio involves degrees of freedom, it does explain how a ratio statistic can be useful in making comparisons between variabilities.

Prediction is a very important use of statistical modelling, so much so that I usually preface class discussions on the topic with a comment that the desire to predict is one of the basic human urges! The dangers of extrapolation are particularly important for students to understand; yet although they are mathematically obvious, students often have trouble translating the mathematical perils of extrapolation to a real world setting. An excellent analogy, communicated to me anonymously, is that of thinking of the observed data as being seen through the viewfinder of a camera. The viewer cannot see what is happening outside the viewfinder, but students can very easily imagine that the surroundings could be very different to what the viewer can see. Extrapolation is like predicting what is happening outside the photograph that was taken - a moment’s thought reveals how difficult this might be! Again, the alignment of a statistical process (extrapolating a model fit) with a physical idea (imagining what lies outside a viewfinder) serves to motivate students to think more carefully about what the inherent limitations of models are. This allows them to think less of models as "truth" but more as simply
"guides" for how the data may behave. Further, once students accept the idea that models have validity limited by the extents of data, they may be encouraged to think about model adequacy beyond the simple rubric of good initial fit to the data.

The fact that in multiple regression individual sums of squares relating to specific sets of explanatory variables behave sequentially is a cause of confusion for students. Of particular concern is the fact that regardless of the order in which a model is fit the parameter estimates are the same. Yet the sums of squares for individual explanatory variables vary according to the order in which the explanatory variables entered the model in the fitting algorithm. As a result, the order that the model is fit is critical for testing certain combinations of predictors. One of the most commonly made errors in testing multiple parameters is the failure of students to fit the parameters in question in the order appropriate for testing specific hypotheses.

An analogy I have found really useful in explaining why sums of squares depend on fit order is that of the unexplored island. Imagine a far-off, uninhabited, virgin island, and imagine that different explorers set off from various countries in search of this island. In this case, the explorers are analogous to the explanatory variables in a multiple regression. One of them will discover the island first, and will set about exploring it until his supplies or capabilities are exhausted. The amount of the area explored by the first arriving adventurer is analogous to the sum of squares for the first fit variable. Eventually, the second explorer will arrive. When he sets out to explore the island, he will find that some of it is already mapped, and so he will venture into that part of the island that remains unexplored. The amount of the island that he surveys represents the sequential sum of squares for the second fit variable given the presence in the model of the first fit variable, since the second explorer cannot lay claim to having explored something that has already been mapped by the first. The other explorers arrive in turn, each mapping as much of what has not been previously explored as their supplies will allow, until they have all arrived leaving the most unreachable part of the island uncharted.

The discovery of the island is therefore attributed to the explorers in turn, each one only able to claim what has been left to discover by their predecessors. The analogical maps in this case include the explorers for the explanatory variables, order of arrival for order of fit, territory explored for variation explained, and impenetrable or unreachable terrain for residual or unexplained variability. The analogy is a strong one provided one understands that "exploration" of a part of the island is a one-time activity (so the notion of exploration is exclusive - once explored, the territory cannot be "discovered" again as it is already known). In this case, an illustration is particularly useful in drawing parallels between areas explored and variation explained.
Figure 4. An analogy for sequential sums of squares. Who is credited with exploring what parts of the island depends critically on who arrived first. In the top left of the figure, DaGama claims the lion’s share, having arrived in 1502, while di Varthema must settle for what’s left on his arrival in 1505. Had di Varthema arrived first in 1505 (bottom left), he may have fared significantly better, while DaGama would have to settle for a much smaller portion were he to arrive later, in 1507. Analogously, variables fit first in a model can claim a larger sum of squares than if they were fit after other variables, when they would be competing for a reduced available sum of squares. In either case, regardless of the order of arrival (fit), the amount remaining unexplored (unexplained) is unchanged.

The analogy also captures other important ideas in regression such as multicollinearity. Explorers who set off from the same direction are likely to be competing to explore the same part of the island. Whoever gets there first will explore the bulk of what their competitor would have been able to explore had they arrived first. So the one who arrives second will have only a very small amount of the island that they can claim as their own. By analogy, collinear variables often have either very small or very large extra sums of squares, depending on which variable was fit first in terms of partitioning the explained variability. Also, explorers who approach the island from very different directions are analogous to predictors that behave independently. Each is likely to explore their own separate part of the island regardless of whether their counterpart arrived before them. Again, the analogy is not intended as an explanation of the overall analysis of variance, as it does not capture important parts of that construction, such as the degrees of freedom associated with each segment of the partitioned variability. The idea does, however, motivate students to understand that the order in which variables are specified in a model-fitting algorithm changes the way we think about them contributing to partitioning the explained variation in the response.
12.2 Of Symptoms and Diseases

The problem of locating influential points in multiple regression settings is a difficult one even for experienced statisticians. Correspondingly, there is a plethora of statistics designed to reveal influential points, including leverage, DFFITS, Cook’s distance, DFBETAS, COVRATIO, and so on. The first of these typically encountered by undergraduates is the leverage, and there is a strong tendency for students to confuse the concepts of high leverage and influence. Of course, we can explain that, leverages depend only on the predictor variables and not on the response. As a result, high leverage on its own cannot be equivalent to influence, which must depend on both the predictors and the response. Nevertheless, in my experience the confusion persists and high leverage is often mistaken for influence. A potent analogy that has been successful in resolving this issue in my classes has been the very simple statement that "leverage is a symptom, while influence is a disease." This analogy brings into clear focus that leverage is merely a diagnostic measure, and that it may or may not reveal the presence of a genuine problem with the data. In the same way, a patient may present to a doctor with a cough but it is not necessarily the case that the patient has the ‘flu. Indeed, in order to diagnose the ‘flu, the doctor would assess a number of other possible symptoms. In the same way, a statistician might use high leverage as an indicator that a certain point needs further investigation. For most students, both the words "leverage" and "influence" are new, and this analogy assists them in learning the relationship between the concepts by aligning them with a word-pair whose relationship is analogous. In fact, a deeper map exists between the source and target analogs because each is embedded in a "diagnostic" context within its domain.

Some of students’ confusions can be addressed just by posing a simple question. As an example, after a lecture that covered different types of residuals, a student asked why studentized deleted residuals were computed using a standard error estimate not involving the corresponding data point. My answer was to immediately ask, "Would you allow a person accused of a crime to be allowed to sit on their own jury?" I could almost see the light bulb appear above the student’s head! The idea expressed here is that a data point should not be allowed to exert undue influence on a procedure designed to determine that very point’s effect on the model. The analogy is evocative, but is not a particularly strong one for some of the reasons discussed in the section on hypothesis testing concerning the role of prior beliefs and intuition. For example, in a criminal trial, the accused is certainly permitted to address the jury, and this may be interpreted by some as exerting undue influence on their decision. The benefit of the analogy is that the motive behind using a delete-one estimate of spread in constructing the residual mirrors the motive for not permitting the accused onto the jury - that is, pursuing an objective decision process. The analogy described here is a simple way of communicating the idea in a way that students can easily grasp.

13. Conclusion

Benjamin Disraeli’s oft-quoted maxim "lies, damned lies and statistics" is usually well known to students before they even attend a class in statistics. As a result, statistics is often seen by students as something of a mystic art that can be employed to bend the truth, to make the data say whatever one wants. Of course, nothing could be further from the truth, as suggested by Mosteller’s comment that while it is easy to lie with statistics it is far easier to lie without them. Yet, students often approach undergraduate study in the area with mistrust and fear, and it is this hurdle that we as teachers of statistics must overcome. One approach to tackling the problem is to motivate the study of statistics by relating statistical notions and statistical thinking to realms more familiar to students, and to focus on simple explanations as far as possible. The use of physical analogies is enormously helpful in achieving this goal by appealing to students’ common experiences to expose logic similar to that used in developing good statistical intuition. The importance of analogical thinking for cognition and learning is well established in the psychology literature. As Holyoak and Thagard (1995, p. 137) remark,
The message, then, is simple. Analogy is not a sure-fire shortcut to expertise, far less a substitute for careful thinking and detailed study of a new domain. But when it is used carefully - when a plausible analog is selected and mapped, when inferences are critically evaluated and adapted as needed, and when the deep structure of the analogy is extracted to form a schema - analogy can be a powerful mental tool. At its best, analogy can jump-start creative thinking in a new domain and through successive refinements move the novice along the path to expertise.

Provided we keep in mind that no analogy is perfect and that they augment rather than replace a good conceptual understanding of statistical thinking, analogies can be extraordinarily powerful tools for enhancing learning outcomes.

Acknowledgements

The author is grateful to the associate editor and the referees for several suggestions that improved the article markedly. Thanks also to Karolina Duklan for invaluable comments and suggestions.

References

Bangdiwala, S. I. (1989), "The teaching of the concepts of statistical tests of hypotheses to non-statisticians," Journal of Applied Statistics, 16, pp. 355-361.

Batanero, C., Godino, J. D., Vallecillos, A., Green, D. R., and Holmes, P. (1994), "Errors and Difficulties in Understanding Elementary Statistical Concepts," International Journal of Mathematical Education in Science and Technology, 25(4), pp. 527-547.

Blanchette, I., and Dunbar, K. (2000), "How analogies are generated: The roles of structural and superficial similarity," Memory and Cognition, 29, pp. 730-735.

Brewer, J. K. (1989), "Analogies and parables in the teaching of statistics," Teaching Statistics, 11, pp. 21-23.

Campbell, S. K. (1974), Flaws and Fallacies in Statistical Thinking, Upper Saddle River, NJ: Prentice Hall.

Catrambone, R. (1997), "Reinvestigating the effects of surface and structural features on analogical access," in Proceedings of the 19th Annual Conference of the Cognitive Science Society, eds. M. G. Shafto and P. Langley, Stanford, CA: Lawrence Erlbaum Associates, Inc. pp. 90-95.

Chanter, D. O. (1983), "Some anecdotes and analogies for illustrating statistical ideas," Teaching Statistics, 5, pp. 14-16.

Donnelly, C. M., and McDaniel, M. A. (1993), "Use of analogy in learning scientific concepts," Journal of Experimental Psychology: Learning, Memory and Cognition, 19, pp. 975-987.
Dunbar, K., and Blanchette, I. (2001), "The invivo/invitro approach to cognition: the case of analogy," *Trends in Cognitive Sciences*, 5, pp. 334-339.

Feinberg, W. E. (1971), "Teaching the type I and II errors: the judicial process," *The American Statistician*, 25, pp. 30-32.

Fischbein, E. (1975), *The Intuitive Sources of Probabilistic Thinking in Children*, Dordrecht, The Netherlands: Reidel.

----- (1987), *Intuition in Science and Mathematics*, Dordrecht, The Netherlands: Reidel.

Freund, J. E. (2001), *Modern Elementary Statistics* (10th ed.), Upper Saddle River, NJ: Prentice Hall.

Gentner, D. (1983), "Structure-mapping: A theoretical framework for analogy," *Cognitive Science*, 7, pp. 155-170.

Gentner, D., Ratterman, M. J., and Forbus, K. (1993), "The roles of similarity in transfer: Separating retrievability from inferential soundness," *Cognitive Psychology*, 25, pp. 524-575.

Gentner, D., and Holyoak, K. J. (1997), "Reasoning and learning by analogy," *American Psychologist*, 52(1), pp. 32-34.

Gentner, D., and Markman, A. B. (1997), "Structure mapping in analogy and similarity," *American Psychologist*, 52(1), pp. 45-56.

Gick, S., and Holyoak, K. J. (1980), "Analogical problem solving," *Cognitive Psychology*, 12, pp. 306-355.

----- (1983), "Schema induction and analogical transfer," *Cognitive Psychology*, 15, pp. 1-38.

Glynn, S. M. (1991), "Explaining science concepts: A teaching-with-analogies model," in *The Psychology of Learning Science*, eds. S. M. Glynn, R. H. Yeany, and B. K. Britton, Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 219-240.

----- (1994), "Teaching science with analogies: A strategy for teachers and textbook authors," Research Report No. 15, Athens, GA: University of Georgia and College Park, MD: University of Maryland, National Reading Research Center.

----- (1995), "Conceptual bridges: Using analogies to explain scientific concepts," *The Science Teacher*, December 1995, pp. 24-27.

----- (1996), "Teaching with analogies: Building on the science textbook," *The Reading Teacher*, 49(6), pp. 490-492.

Glynn, S.M., Duit, R., and Thiele, R. (1995), "Teaching science with analogies: a strategy for transferring knowledge," in *Learning Science in the Schools: Research Reforming Practice*, eds. S. M. Glynn and R. Duit, Mahwah, NJ: Lawrence Erlbaum Associates, pp. 247-273.
Glynn, S. M., and Law, M. (1993), "Teaching science with analogies: Building on the book" [Video], Athens, GA: University of Georgia and College Park, MD: University of Maryland, National Reading Research Center.

Hesse, M. (1966), *Models and Analogies in Science*, South Bend, IN: Notre Dame University Press.

Hollander, M., and Proschan, F. (1984), *The Statistical Exorcist: Dispelling Statistics Anxiety*, New York: Marcel Dekker, Inc.

Holyoak, K. J., and Thagard, P. (1995), *Mental Leaps: Analogy in Creative Thought*, Cambridge, Massachusetts: MIT Press.

----- (1997), "The Analogical Mind," *American Psychologist*, 52(1), pp. 35-44.

Huff, D. (1954), *How to Lie with Statistics*, New York: W. W. Norton and Company.

Johnson, R. A., and Tsui, K.-W. (1998), *Statistical Reasoning and Methods*, New York: John Wiley and Sons, Inc.

Keane, M. T. (1987), "On retrieving analogues when solving problems," *Quarterly Journal of Experimental Psychology*, 39A, pp. 29-41.

Kimble, G. A. (1978), *How to Use (and Misuse) Statistics*, Upper Saddle River, NJ: Prentice Hall.

Kitchens, L. J. (1998), *Exploring Statistics: A Modern Introduction to Data Analysis and Inference*, Pacific Grove, CA: Duxbury Press.

Kolodner, J. L. (1997), "Educational implications of analogy: A view from case-based reasoning," *American Psychologist*, 52, pp. 57-66.

Larsen, R. J., and Marx, M. L. (1990), *Statistics*, Upper Saddle River, NJ: Prentice Hall.

Lesser, L. M. (1994), "The Role of Counterintuitive Examples in Statistics Education," Doctoral Dissertation, University of Texas at Austin.

Moore, D. S. (1997), *Statistics: Concepts and Controversies* (4th ed.), New York: W. H. Freeman and Co., Inc.

Novick, L. R. (1988), "Analogical transfer, problem similarity, and expertise," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14, pp. 510-520.

Novick, L. R., and Holyoak, K. J. (1991), "Mathematical problem solving by analogy," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17, pp. 398-415.

Quilici, J. L., and Mayer, R. E. (1996), "Role of examples in how students learn to categorize statistics word problems," *Journal of Educational Psychology*, 88, pp. 144-161.

Reeves, C. A., and Brewer, J. K. (1980), "Hypothesis testing and proof by contradiction: an analogy," *Teaching Statistics*, 2, pp. 57-59.
Schustack, M. W., and Anderson, J. R. (1979), "Effects of analogy to prior knowledge on memory for new information," *Journal of Verbal Behavior and Verbal Learning*, 18, pp. 565-583.

Shaughnessy, J. M. (1992), "Research in Probability and Statistics: Reflections and Directions," in *Handbook of Research on Mathematics Teaching and Learning*, ed. D. Grouws, New York: Macmillan, pp. 465-494.

Sowey, E. R. (1995), "Teaching Statistics: Making it Memorable," *Journal of Statistics Education* [Online], 3(2). (ww2.amstat.org/publications/jse/v3n2/sowey.html)

----- (2001), "Striking Demonstrations in Teaching Statistics," *Journal of Statistics Education* [Online], 9(1). (ww2.amstat.org/publications/jse/v9n1/sowey.html)

Spiro, R. J., Feltovich, P. J., Coulson, R. L., and Anderson, D. K. (1989), "Multiple analogies for complex concepts: Antidotes for analogy-induced misconception in advanced knowledge acquisition," in *Similarity and analogical reasoning*, eds. S. Vosniadou and A. Ortony, Cambridge: Cambridge University Press, pp. 498-531.

Tufte, E. R. (1997), *Visual Explanations*, Cheshire, CT: Graphics Press.

Wharton, C. M., Holyoak, K. J., Downing, P. E., Lange, T. E., Wickens, T. D., and Melz, E. R. (1994), "Below the surface: Analogical similarity and retrieval competition in reminding," *Cognitive Psychology*, 26, pp. 64-101.

---

**Addendum**

Volume 11, Number 3, of the *Journal of Statistics Education* contains a [Letter to the Editor](mailto:Michael.Martin@anu.edu.au) concerning this article.

---

Michael Martin  
School of Finance and Applied Statistics  
Australian National University  
Canberra, Australian Capital Territory 0200  
AUSTRALIA  
*Michael.Martin@anu.edu.au*