A Statistical Guide for the Ethically Perplexed

Lawrence Hubert and Howard Wainer

Preamble

The meaning of "ethical" adopted here is one of being in accordance with the accepted rules or standards for right conduct that govern the practice of some profession. The professions we have in mind are statistics and the behavioral sciences, and the standards for ethical practice are what we try to instill in our students through the methodology courses we offer, with particular emphasis on the graduate statistics sequence generally required for all the behavioral sciences. Our hope is that the principal general education payoff for competent statistics instruction is an increase in people's ability to be critical and ethical consumers and producers of the statistical reasoning and analyses faced in various applied contexts over the course of their careers.

Introduction

Generations of graduate students in the behavioral and social sciences have completed mandatory year-long course sequences in statistics, sometimes with difficulty and possibly with less than positive regard for the content and how it was taught. Prior to the 1960's, such a sequence usually emphasized a cookbook approach where formulas were applied unthinkingly using mechanically operated calculators. The instructional method could be best characterized as "plug and chug," where there was no need to worry about the meaning of what one was doing, only that the numbers could be put in and an answer generated. This process would hopefully lead to numbers that could then be looked up in tables; in turn, p-values were sought that were less than the magical .05, giving some hope of getting an attendant paper published.

The situation began to change for the behavioral sciences in 1963 with the publication of *Statistics for Psychologists* by William Hays. For the first time, graduate students could be provided both the needed recipes and some deeper understanding of and appreciation for the whole enterprise of inference in the face of uncertainty and fallibility. Currently, the

Hays text is in its fifth edition, with a shortened title of *Statistics* (1994); the name of Hays itself stands as the eponym for what kind of methodology instruction might be required for graduate students – i.e., at the level of Hays, and cover-to-cover. Although now augmented by other sources for related computational work (e.g., by SAS, SPSS, or SYSTAT), the Hays text remains a standard of clarity and completeness. Many methodologists have based their teaching on this resource for more than four decades. Hays typifies books that although containing lucid explanations of statistical procedures, are too often used by students only as a cookbook of statistical recipes. The widespread availability of statistical software has made it clear that we no longer have a need for cookbooks, and instead, require a *Guide to Gastronomy.*

In teaching graduate statistics, there are multiple goals:

1) to be capable of designing and analyzing one's own studies, including doing the computational "heavy lifting" by one's self, and the ability to verify what others attached to a project may be doing;

2) to understand and consume other research intelligently, both in one's own area, but more generally as a statistically and numerically literate citizen;

3) to argue for and justify analyses when questioned by journal and grant reviewers or others, and to understand the basic justification for what was done. For example, an ability to reproduce a formal proof of the Central Limit Theorem is unnecessary, but a general idea of how it is formulated and functions is relevant, and that it might help justify assertions of robustness being made for the methods used. These skills in understanding are not "theoretical" in a pejorative sense, although they do require more thought than just being content to run the SPSS machine blindly. They are absolutely crucial in developing both the type of reflective teaching and research careers we would hope to nurture in graduate students, and more generally for the quantitatively literate citizenry we would wish to make up our society.

Graduate instruction in statistics requires the presentation of general frameworks and how to reason from these. These frameworks can be conceptual: for example, (a) the Fisherian view that provided the evidence of success in the Salk Polio vaccine trials where the physical act of randomization lead to credible causal inferences; or (b) to the unification given by the notion of maximum likelihood estimation and likelihood ratio tests both for our general statistical modeling as well as for more directed formal modeling in a behavioral science subdomain, such as image processing or cognitive neuroscience. These frameworks can also be based on more quantitatively formal structures: for example, (a) the general linear model and its special cases of analysis-of-variance (ANOVA), analysis-of-covariance, and so on, along with model comparisons through full and reduced models; (b) the general principles behind prediction/selection/correlation in simple two-variable systems, with extensions to multiple-variable contexts; and (c) the various dimensionality reduction techniques of principal component/factor analysis, multidimensional scaling, cluster analysis, and discriminant analysis.

The remainder of the sections in this essay will attempt to sketch some basic structures typically introduced in the graduate statistics sequence in the behavioral sciences, along with some necessary cautionary comments on usage and interpretation. The purpose is to provide a small part of the formal scaffolding needed in reasoning ethically about what we see in the course of our careers, both in our own work and that of others, or what might be expected of a statistically literate populace generally. Armed with this deeper understanding, graduates can be expected to deal more effectively with whatever ethically charged situations they might face.

Probability Theory

The formalism of thought offered by probability theory is one of the more useful portions of any beginning course in statistics in helping to promote ethical reasoning. As typically presented, we speak of an event represented by a capital letter, say A, and the probability of the event as some number in the range from 0 to 1, written as $P(A)$. The value of 0 is assigned to the "impossible" event that can never occur; 1 is assigned to the "sure" event that will always occur. The driving condition for the complete edifice of all probability theory is one postulate: for two mutually exclusive events, A and B (where mutually exclusivity implies that both events cannot occur at the same time), $P(A \text{ or } B) = P(A) + P(B)$. As one final beginning definition, we say that two events are independent whenever the probability of the joint event, $P(A \text{ and } B)$, factors as the product of the individual probabilities, $P(A)P(B)$.

The idea of statistical independence and the factoring of the joint event probability, immediately provides a formal tool for understanding several historical miscarriages of justice. In particular, if two events are not independent, then the joint probability cannot be generated by a simple product of the individual probabilities. A recent example is the case of Sally Clark; she was convicted in England of killing her two children, partially on the basis of an inappropriate assumption of statistical independence. The purveyor of statistical misinformation in this case was Sir Roy Meadow, famous for Meadow's Law: "one sudden infant death is a tragedy, two is suspicious, and three is murder." We quote part of a news release from the Royal Statistical Society (Tuesday, October 23, 2001):

> The Royal Statistical Society today issued a statement, prompted by issues raised by the Sally Clark case, expressing its concern at the misuse of statistics in the courts.
>
> In the recent highly-publicised case of R v. Sally Clark, a medical expert witness drew on published studies to obtain a figure for the frequency of sudden infant death syndrome (SIDS, or 'cot death') in families having some of the characteristics of the defendant's family. He went on to square this figure to obtain a value of 1 in 73 million for the frequency of two cases of SIDS in such a family.
>
> This approach is, in general, statistically invalid. It would only be valid if SIDS cases arose independently within families, an assumption that would need to be justified empirically. Not only was no such empirical justification provided in the case, but there are very strong a priori reasons for supposing that the assumption will be false. There may well be unknown genetic or environmental factors that predispose families to SIDS, so that a second case within the family becomes much more likely.
>
> The well-publicised figure of 1 in 73 million thus has no statistical basis. Its use cannot reasonably be justified as a 'ballpark' figure because the error involved is likely to be very large, and in one particular direction. The true frequency of families with two cases of SIDS may be very much less incriminating than the figure presented to the jury at trial.

Numerous other examples for a misuse of the idea of statistical independence exist in the legal literature, such as the notorious 1968 jury trial in California, People v. Collins. Here, the prosecutor suggested that the jury merely multiply several probabilities together, which he conveniently provided, to ascertain the guilt of the defendant. In overturning the conviction, the Supreme Court of California criticized both the statistical reasoning and the framing of the decision for the jury:

> We deal here with the novel question whether evidence of mathematical probability has been properly introduced and used by the prosecution in a criminal case. ... Mathematics, a veritable sorcerer in our computerized society, while assisting the trier of fact in the search of truth, must not cast a spell over him. We conclude that on the record before us, defendant should not have had his guilt determined by the odds and that he is entitled to a new trial. We reverse the judgement.

We will return to both the Clark and Collins cases later when Bayes rule is discussed in the context of conditional probability confusions and what is called the "Prosecutor's Fallacy."

Besides the concept of independence, the definition of conditional probability plays a central role in all our uses of probability theory; in fact, most misapplications of statistical/probabilistic reasoning involve confusions of some sort regarding conditional probabilities. Formally, the conditional probability of some event A given that B has already occurred, denoted $P(A|B)$, is defined generally as $P(A \text{ and } B)/P(B)$; when A and B are independent, $P(A|B) = P(A)P(B)/P(B) = P(A)$; or in words, knowing that B has occurred does not alter the probability of A occurring. If $P(A|B) > P(A)$, we will say that B is "facilitative" of A; when $P(A|B) < P(A)$, B is said to be "inhibitive" of A. As a small example, suppose A is the event of receiving a basketball scholarship; B, the event of being seven feet tall; and C, the event of being five feet tall. One obviously expects B to be facilitative of A (i.e., $P(A|B) > P(A)$); and of C to be inhibitive of A (i.e., $P(A|C) < P(A)$). In any case, the size and sign of the difference between $P(A|B)$ and $P(A)$ is an obvious raw descriptive measure of how much the occurrence of B is associated with an increased or decreased probability of A, with a value of zero corresponding to statistical independence.

One convenient device for interpreting probabilities and understanding how events can be "facilitative" or "inhibitive" is through the use of a simple 2×2 table that cross-classifies a set of objects according to the events A and $\bar{A}$, and B and $\bar{B}$. For example, suppose we have a collection of N balls placed in a container; each ball is labeled with A or $\bar{A}$, and also with B or $\bar{B}$, according to the notationally self-evident table of frequencies below:

	A	$\bar{A}$	
B	N_{AB}	$N_{\bar{A}B}$	N_B
$\bar{B}$	$N_{A\bar{B}}$	$N_{\bar{A}\bar{B}}$	$N_{\bar{B}}$
	N_A	$N_{\bar{A}}$	N

The process we consider is one of picking a ball blindly from the container (where the balls are assumed to be mixed thoroughly), and noting the occurrence of the events A or $\bar{A}$ and B or $\bar{B}$. Based on this physical idealization of such a selection process, it is intuitively reasonable to assign probabilities according to the proportion of balls in the container satisfying the attendant conditions:

$P(A) = N_A/N; P(\bar{A}) = N_{\bar{A}}/N; P(B) = N_B/N; P(\bar{B}) = N_{\bar{B}}/N;$

$P(A|B) = N_{AB}/N_B; P(B|A) = N_{AB}/N_A;$

$P(\bar{A}|B) = N_{\bar{A}B}/N_B; P(B|\bar{A}) = N_{\bar{A}B}/N_{\bar{A}};$

$P(\bar{B}|A) = N_{A\bar{B}}/N_A; P(A|\bar{B}) = N_{A\bar{B}}/N_{\bar{B}};$

$P(\bar{A}|\bar{B}) = N_{\bar{A}\bar{B}}/N_{\bar{B}}; P(\bar{B}|\bar{A}) = N_{\bar{A}\bar{B}}/N_{\bar{A}}$.

By noting the relationships: $N_B = N_{AB} + N_{\bar{A}B}$; $N_{\bar{B}} = N_{A\bar{B}} + N_{\bar{A}\bar{B}}$; $N_A = N_{AB} + N_{A\bar{B}}$; $N_{\bar{A}} = N_{\bar{A}B} + N_{\bar{A}\bar{B}}$; $N_B + N_{\bar{B}} = N_A + N_{\bar{A}} = N$, a variety of interesting connections can be derived and understood that can assist immensely in our probabilistic reasoning. We present a short numerical example below on how these ideas might be used in a realistic context; several such uses are then expanded upon in the subsections to follow.

As a numerical example of using a 2 × 2 contingency table to help explicate probabilistic reasoning, suppose we have an assumed population of 10,000, cross-classified according to the presence or absence of Colorectal Cancer (CC) [A: +CC; $\bar{A}$: −CC], and the status of a Fecal Occult Blood Test (FOBT) [B: +FOBT; $\bar{B}$: −FOBT]. Using the data from Gerd Gigerenzer, *Calculated Risks*, we have the following 2 × 2 table:

	+CC	−CC	
+FOBT	15	299	314
−FOBT	15	9671	9686
	30	9970	10,000

The probability, $P(+CC \mid +FOBT)$, is simply 15/314 = .048, using the frequency value of 15 for the cell (+FOBT, +CC), and the +FOBT row sum of 314. The marginal probability, P(+CC), is 30/10,000 = .003, and thus, a positive FOBT is "facilitative" of a positive CC because .048 is greater than .003. The size of the difference, $P(+CC \mid +FBOT) - P(+CC) =$ +.045, may not be large in any absolute sense, but the change does represent a fifteen-fold increase over the marginal probability of .003. (But note that if you have a positive FOBT, over 95% of the time you don't have cancer – i.e., there are 95% ($\frac{299}{314}$) false positives.)

There are many day-to-day contexts faced where our decisions might best be made from conditional probabilities (if we knew them), instead of from marginal information. When deciding on a particular medical course of action, for example, it is important to condition on our own circumstances of age, risk factors, family medical history, our own psychological needs and makeup, and so on. A recent and controversial instance of this, where the conditioning information is "age," is reported in *The New York Times* article by Gina Kolata, *In Reversal, Panel Urges Mammograms at 50, Not 40* (November 17, 2009).

There are a variety of probability results that prove useful throughout our attempt to reason probabilistically and follow the field of statistical inference. We list some of these below, with uses given throughout this essay.

1) For the complementary event, $\bar{A}$, which occurs when A does not, $P(\bar{A}) = 1 - P(A)$.

2) For events A and B that are not necessarily mutually exclusive,

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \ .$$

3) The *rule of total probability*: given a collection of mutually exclusive and exhaustive events, $B_1, \ldots, B_K$ (i.e., all are pairwise mutually exclusive and their union gives the sure event),

$$P(A) = \sum_{k=1}^{K} P(A|B_k)P(B_k) \ .$$

4) Bayes Theorem (or Rule) for two events, A and B:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\bar{A})P(\bar{A})} \ .$$

5) Bonferroni inequality: for a collection of events, $A_1, \ldots, A_K$,

$$P(A_1 \text{ or } A_2 \text{ or } \cdots \text{ or } A_K) \leq \sum_{k=1}^{K} P(A_k) \ .$$

6) $P(A \text{ and } B) \leq P(A \text{ or } B) \leq P(A) + P(B)$.

In words, the first inequality results from the event "A and B" being wholly contained within the event "A or B"; the second obtains from the Bonferroni inequality restricted to two events.

7) $P(A \text{ and } B) \leq \text{minimum}(P(A), P(B)) \leq P(A) \text{ or } \leq P(B)$.

In words, the first inequality results from the event "A and B" being wholly contained both within A and within B; the second inequalities are more generally appropriate – the minimum of any two numbers is always less than or equal to either of the two numbers.

The (mis)assignment of probabilities

Although the assignment of probabilities to events consistent with the disjoint rule may lead to an internally valid system mathematically, there is still no assurance that this assignment is "meaningful," or bears any empirical validity for observable long-run expected frequencies. There seems to be a never-ending string of misunderstandings in the way probabilities can be generated that are either blatantly wrong, or more subtly incorrect, irrespective of the internally consistent system they might lead to. Some of these problems are briefly sketched below, but we can only hope to be representative of a few possibilities, not exhaustive.

One inappropriate way of generating probabilities is to compute the likelihood of some joint occurrence after some of the outcomes are already known. There is the story about the statistician who takes a bomb aboard a plane, reasoning that if the probability of one bomb on board is small, the probability of two is infinitesimal. Or, during World War I, soldiers were actively encouraged to use fresh shell holes as shelter because it was very unlikely for two shells to hit the same spot during the same day. And the (Minnesota Twins) baseball manager who bats for an individual who earlier in the game hit a home run because it would be very unlikely for him to hit two home runs in the same game. Although these (slightly) amusing stories may provide obvious misassignments of probabilities, other related situations are more subtle. For example, whenever coincidences are culled or "hot spots" identified from some search of available information, the probabilities that are then regenerated for these situations may not be valid. There are several ways of saying this: when some set of observations is the source of an initial suspicion, those same observations should not be used in a calculation that then tests the validity of the suspicion. In Bayesian terms, you don't get the Posterior from the same information that gave you the Prior.

Alternatively said, it makes no sense to do formal hypothesis assessment (by finding estimated probabilities) when the data themselves have suggested the hypothesis in the first place. Some cross-validation strategy is necessary, e.g., collecting independent data. Generally, when some process of search or optimization has been used to identify an unusual situation (e.g., when a "good" regression equation is found through a step-wise procedure [see Freedman, 1983, for a devastating critique]; when data are "mined" and unusual patterns identified; when DNA databases are searched for "cold-hits" against evidence left at a crime scene; when geographic "hot spots" are identified for, say, some particularly unusual cancer, and so on), the same methods for assigning probabilities before the particular situation was identified, are generally no longer appropriate post-hoc.

A second general area of inappropriate probability assessment concerns the model postulated to aggregate probabilities over several events. Campbell (1974) cites an article in the *New York Herald Tribune* (May, 1954) stating that if the probability of knocking down an attacking airplane was .15 at each of five defense positions before reaching the target, then the probability of knocking down the plane before it passed all five barriers would be .75 ($5 \times .15$), this last value being the simple sum of the probabilities, and an inappropriate model. If we could correctly assume independence between the Bernoulli trials at each of the five positions, a more justifiable value would be one minus the probability of passing all barriers successfully: $1.0 - (.85)^5 \approx .56$. The use of similar binomial modeling possibilities, however, may be specious – for example, when dichotomous events occur simultaneously in groups (e.g., the World Trade Center disaster on 9/11/01); when the success proportions are not valid; when the success proportions change in value over the course of the trials;

when time dependencies are present in the trials (e.g., tracking observations above and below a median over time), and so on. In general, when wrong models are used to generate probabilities, the resulting values may have little to do with empirical reality. For instance, in throwing dice and counting the sum of spots that result, it is not true that each of the integers from two through twelve are equally likely. The model of what is equally likely may be reasonable at a different level (e.g., pairs of integers appearing on the two dice), but not at all aggregated levels. There are some stories, probably apocryphal, of methodologists meeting their demises by making these mistakes for their gambling patrons.

Flawed calculations of probability can have dire consequences within our legal systems, as the case of Sally Clark and related others, makes clear. One broad and current area of possible misunderstanding of probabilities is in the context of DNA evidence (which is exacerbated in the older and much more fallible system of identification through fingerprints). In the use of DNA evidence (and with fingerprints), one must be concerned with the Random Match Probability (RMP): the likelihood that a randomly selected unrelated person from the population would match a given DNA profile. Again, the use of independence in RMP estimation is questionable; also, how does the RMP relate to, and is it relevant for, "cold-hit" searches in DNA databases. In a confirmatory identification case, a suspect is first identified by non-DNA evidence; DNA evidence is then used to corroborate traditional police investigation. In a "cold-hit" framework, the suspect is first identified by a search of DNA databases; the DNA evidence is thus used to identify the suspect as perpetrator, to the exclusion of others, directly from the outset (this is somewhat akin to shooting an arrow into a tree and then drawing a target around it). Here, traditional police work is no longer the focus. For a thorough discussion of the probabilistic context surrounding DNA evidence (which extends with even greater force to fingerprints), the article by Jonathan Koehler is recommended (*Error and Exaggeration in the Presentation of DNA Evidence at Trial*, *Jurimetrics Journal*, *34*, 1993–1994, 21–39).

In 1989, and based on urging from the FBI, the National Research Council (NRC) formed the Committee on DNA Technology in Forensic Science, which issued its report in 1992 (*DNA Technology in Forensic Science*; or more briefly, NRC I). The NRC I recommendation about the cold-hit process was as follows:

> The distinction between finding a match between an evidence sample and a suspect sample and finding a match between an evidence sample and one of many entries in a DNA profile databank is important. The chance of finding a match in the second case is considerably higher. ... The initial match should be used as probable cause to obtain a blood sample from the suspect, but only the statistical frequency associated with the additional loci should be presented at trial (to prevent the selection bias that is inherent in searching a databank).

A followup report by a second NRC panel was published in 1996 (*The Evaluation of Forensic DNA Evidence*; or more briefly, NRC II), having the following main recommendation about cold-hit probabilities and using what has been called the "database match probability" or DMP:

> When the suspect is found by a search of DNA databases, the random-match probability should be multiplied by N, the number of persons in the database.

The term "database match probability" (DMP) is somewhat unfortunate; this is not a real probability but more of an expected number of matches given the RMP. A more legitimate value for the probability that another person matches the defendant's DNA profile would be $1 - (1 - \frac{1}{\text{RMP}})^N$, for a database of size N, i.e., one minus the probability of no matches over N trials. For example, for an RMP of 1/1,000,000 and an N of 1,000,000, the above probability of another match is .632; the DMP (not a probability) number is 1.00, being the product of N and RMP. In any case, NRC II made the recommendation of using the DMP to give a measure of the accuracy of a cold-hit match (and did not support the more legitimate "probability of another match" using the formula given above [possibly because it was considered too difficult?]):

> A special circumstance arises when the suspect is identified not by an eyewitness or by circumstantial evidence but rather by a search through a large DNA database. If the only reason that the person becomes a suspect is that his DNA profile turned up in a database, the calculations must be modified. There are several approaches, of which we discuss two. The first, advocated by the 1992 NRC report, is to base probability calculations solely on loci not used in the search. That is a sound procedure, but it wastes information, and if too many loci are used for identification of the suspect, not enough might be left for an adequate subsequent analysis. ... A second procedure is to apply a simple correction: Multiply the match probability by the size of the database searched. This is the procedure we recommend.

The probabilistic generalizations of logical fallacies are no longer fallacies

In our roles as instructors in beginning statistics, we commonly introduce some simple logical considerations early on that revolve around the usual "if p, then q" statements, where p and q are two propositions. As an example, we might let p be "the animal is a Yellow Labrador Retriever," and q, "the animal is in the order *Carnivora*." Continuing, we note that if the statement "if p, then q" is true (which it is), then logically, so must be the contrapositive of "if not q, then not p," i.e., if "the animal is not in the order *Carnivora*," then "the animal is not a Yellow Labrador Retriever." However, there are two fallacies awaiting the unsuspecting:

denying the antecedent: if not p, then not q (if "the animal is not a Yellow Labrador Retriever," then "the animal is not in the order *Carnivora*");

affirming the consequent: if q, then p (if "the animal is in the order *Carnivora*," then "the animal is a Yellow Labrador Retriever").

Also, when we consider definitions given in the form of "p if and only if q," (for example, "the animal is a domesticated dog" if and only if "the animal is a member of the subspecies *Canis lupus familiaris*"), or equivalently, "p is necessary and sufficient for q," these separate into two parts:

"if p, then q" (i.e., p is a sufficient condition for q);

"if q, then p" (i.e., p is a necessary condition for q).

So, for definitions, the two fallacies are not present.

In a probabilistic context, we reinterpret the phrase "if p, then q" as B being facilitative of A, i.e., $P(A|B) > P(A)$, where p is identified with B and q with A. With such a probabilistic reinterpretation, we no longer have the fallacies of denying the antecedent (i.e., $P(\bar{A}|\bar{B}) > P(\bar{A})$), or of affirming the consequent (i.e., $P(B|A) > P(B)$). Both of the latter two probability statements can be algebraically shown true using the simple 2 × 2 cross-classification frequency table and the equivalences among frequency sums given earlier:

(original statement) $P(A|B) > P(A) \Leftrightarrow N_{AB}/N_B > N_A/N \Leftrightarrow$

(denying the antecedent) $P(\bar{A}|\bar{B}) > P(\bar{A}) \Leftrightarrow N_{\bar{A}\bar{B}}/N_{\bar{B}} > N_{\bar{A}}/N \Leftrightarrow$

(affirming the consequent) $P(B|A) > P(B) \Leftrightarrow N_{AB}/N_A > N_B/N \Leftrightarrow$

(contrapositive) $P(\bar{B}|\bar{A}) > P(\bar{B}) \Leftrightarrow N_{\bar{A}\bar{B}}/N_{\bar{A}} > N_{\bar{B}}/N$

Another way of understanding these results is to note that the original statement of $P(A|B) > P(A)$, is equivalent to $N_{AB} > N_A N_B/N$, or in the usual terminology of a 2×2 contingency table, the frequency in the cell labeled (A, B) is greater than the typical expected value constructed under independence of the attributes based on the row total, N_B, times the column total, N_A, divided by the grand total, N. The other probability results follow from the observation that with fixed marginal frequencies, a 2×2 contingency table has only one degree-of-freedom. These results derived from the original of B being facilitative for A, $P(A|B) > P(A)$, could have been restated as $\bar{B}$ being inhibitive of A, or as $\bar{A}$ being inhibitive of B.

In reasoning logically about some situation, it would be rare to have a context that would be so cut-and-dried as to lend itself to the simple logic of "if p, then q," and where we could look for the attendant fallacies to refute some causal claim. More likely, we are given problems characterized by fallible data, and subject to other types of probabilistic processes. For example, even though someone may have some genetic marker that has a greater presence in individuals who have developed some disease (e.g., breast cancer and the BRAC1 gene), it is not typically an unadulterated causal necessity; in other words, it is not true that "if you have the marker, then you must get the disease." In fact, many of these situations might be best reasoned through using our simple 2×2 tables — A and $\bar{A}$ denote the presence/absence of the marker; B and $\bar{B}$ denote the presence/absence of the disease. Assuming A is facilitative of B, we could go on to ask about the strength of the facilitation by looking at, say, the difference, $P(B|A) - P(B)$.

The idea of arguing probabilistic causation is, in effect, our notion of one event being facilitative or inhibitive of another. If we observe a collection of "q" conditions that would be the consequence of a single "p," we may be more prone to conjecture the presence of "p." Although this process may seem like merely affirming the consequent, in a probabilistic context this could be referred to as "inference to the best explanation," or as a variant of the

Charles Pierce notion of abductive reasoning. In any case, with a probabilistic reinterpretation, the assumed fallacies of logic may not be such; moreover, most uses of information in contexts that are legal (forensic) or medical (through screening), or that might, for example, involve academic or workplace selection, need to be assessed probabilistically.

Using Bayes rule to assess the consequences of screening for rare events

Bayes theorem or rule was given in a form appropriate for two events, A and B; it allows the computation of one conditional probability, $P(A|B)$, from two other conditional probabilities, $P(B|A)$ and $P(B|\bar{A})$, and the prior probability for the event A, $P(A)$. A general example might help show the importance of Bayes rule in assessing the value of screening for the occurrence of rare events:

Suppose we have a test that assesses some relatively rare quantity (e.g., disease, ability, talent, terrorism propensity, drug/steriod usage, antibody presence, being a liar [where the test is a polygraph], and so forth). Let B be the event that the test says the person has "it," whatever that may be; A is the event that the person really does have "it." Two "reliabilities" are needed:

a) the probability, $P(B|A)$, that the test is positive if the person has "it"; this is called the *sensitivity* of the test;

b) the probability, $P(\bar{B}|\bar{A})$, that the test is negative if the person doesn't have "it"; this is the *specificity* of the test. The conditional probability used in the denominator of Bayes rule, $P(B|\bar{A})$, is merely $1 - P(\bar{B}|\bar{A})$, and is the probability of a "false positive."

The quantity of prime interest, called the *positive predictive value* (PPV), is the probability that a person has "it" given that the test says so, $P(A|B)$, and is obtainable from Bayes rule using the specificity, sensitivity, and prior probability, $P(A)$:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + (1 - P(\bar{B}|\bar{A}))(1 - P(A))} \ .$$

To understand how well the test does, the facilitative effect of B on A needs interpretation, i.e., a comparison of $P(A|B)$ to $P(A)$, plus an absolute assessment of the size of $P(A|B)$ by itself. Here, the situation is usually dismal whenever $P(A)$ is small (i.e., screening for a relatively rare quantity), and the sensitivity and specificity are not perfect. Although $P(A|B)$ will generally be greater than $P(A)$, and thus, B facilitative of A, the absolute size of $P(A|B)$ is typically so small that the value of the screening may be questionable.

As an example, consider the efficacy of mammograms in detecting breast cancer. In the United States, about 180,000 women are found to have breast cancer each year from among the 33.5 million women who annually have a mammogram. Thus, the probability of a tumor is 180,000/33,500,000 = .0054. Mammograms are no more than 90% accurate, implying that $P(\text{positive mammogram} \mid \text{tumor}) = .90$; $P(\text{negative mammogram} \mid \text{no tumor}) = .90$. Because we do not know whether a tumor is present, all we know is whether the test is positive, Bayes theorem must be used to calculate the probability we really care about, the PPV: $P(\text{tumor} \mid \text{positive mammogram})$. All the pieces are available to use Bayes theorem to calculate this probability and we will do so below. But first, as an exercise for the reader, try to estimate the order of magnitude of that probability, keeping in mind that cancer is rare and the test for it is 90% accurate. Do you guess that if you test positive, you have a 90% chance of cancer? Or perhaps 50%, or 30%? How low must this probability drop before we feel that mammograms may be an unjustifiable drain on resources? Using Bayes rule, the PPV of the test is .047:

$$P(\text{tumor} \mid \text{positive mammogram}) = \frac{.90(.0054)}{.90(.0054) + .10(.9946)} = .047 \ ,$$

which is obviously greater than the prior probability of .0054, but still very small in magnitude, i.e., more than 95% of the positive tests that arise turn out to be incorrect.

Whether using a test that is wrong 95% of the time is worth doing is, at least partially, an ethical question, for if we decide that it isn't worth doing, what is the fate of the 5% or so of women who are correctly diagnosed? We will not attempt a full analysis, but some factors considered might be economic, for 33.5 million mammograms cost about \$3.5 billion, and the 3.5 million women incorrectly diagnosed can be, first, dysfunctionally frightened, and second, they must use up another day for a biopsy, in turn costing at least $1,000 and adding another $3.5 billion to the overall diagnostic bill. Is it worth spending $7 billion to detect 180,000 tumors? That is about $39,000/tumor detected. And, not to put too fine a point on it, biopsies have their own risks: 1% yield Staph infections, and they too have false positives implying that some women end up being treated for non-existent cancers. Also, the majority of the cancers detected in the 5% alluded to above, are generally not life-threatening, and just lead to the ills caused by overdiagnois and invasive overtreatment. The statistics just calculated do not make the decision about whether it is ethical or not to do mammograms, but such a decision to be ethical should be based on accurate information. Two recent articles discuss how the American Cancer Society may itself be shifting its stance on screening; the "page one, above the fold" pieces are by Gina Kolata (*In Shift, Cancer Society Has Concerns on Screenings*, *The New York Times*, October 21, 2009; *In Reversal, Panel Urges Mammograms at 50, Not 40*, *The New York Times*, November 17, 2009). A third recent article discusses the odds and economics of screening (with calculations similar to those given here): *Gauging the Odds (and the Costs) in Health Screening* (Richard H. Thaler, *The New York Times*, December 20, 2009).

As we have seen in subsequent reactions to these "new" recommendations regarding screening for breast cancer, it is doubtful whether individual women will comply, or even that their doctors will advise them to. Health recommendations, such as these, pertain to an aggregate populace, possibly subdivided according to various demographic categories. But an individual who seeks some kind of control over (breast) cancer, is not going to give up the only means they have to do so; all women know (at least indirectly) various individuals for whom breast cancer was detected early (and "cured," even though actually the given cancer may not have been harmful); similarly, all women know about individuals who died after a cancer had metastasized before screening located it. What might be justifiable public health policy in the aggregate, may not be so when applied at the level of individuals; also, the issue that trumps all in the mammogram discussion is what women want (or think they want, which amounts to the same thing). It is doubtful whether a reasoned argument for diminished screening could ever be made politically palatable. To many, a statistical argument for a decrease of screening practice would merely be another mechanism by which insurance companies can deny coverage and make yet more money. To paraphrase a quote about General Motors, it is not true that "what is good for the Insurance Industry is good for the country" (or for that matter, for any single individual living in it). Two very cogent articles on these issues of screening both for individuals and the aggregate, appeared on the same day (November 20, 2009) in *The New York Times*: *A Medical Culture Clash* by Kevin Sack; *Addicted to Mammograms* by Robert Aronowitz.

It might be an obvious statement to make, but in our individual dealings with doctors and the medical establishment generally, it is important for all to understand the PPVs for whatever screening tests we now seem to be constantly subjected to, and thus, the number, (1 − PPV), referring to the false positives, i.e., if a patient tests positive, what is the probability that "it" is not actually present. It is a simple task to plot PPV against $P(A)$ from 0 to 1 for any given pair of sensitivity and specificity values. Such a plot can show dramatically the need for highly reliable tests in the presence of low values of $P(A)$ to attain even mediocre PPV values.

Besides a better understanding of how PPVs are determined, there is a need to recognize that even when a true positive exists, not every disease needs to be treated. In the case of another personal favorite of ours, prostate cancer screening (in that its low accuracy makes mammograms look good), where the worst danger is one of overdiagonosis and overtreatment, leading to more harm than good (see, e.g., Gina Kolata, *Studies Show Prostate Test Save Few Lives*, *The New York Times*, March 19, 2009). Armed with this information, we no longer have to hear the snap of a latex glove behind our backs at our yearly physical, nor do we give blood for a PSA screening test. When we so informed our doctors as to our wishes, they agreed completely; the only reason such tests were done routinely was to practice "defensive medicine" on behalf of their clinics, and to prevent possible lawsuits arising from such screening tests not being administered routinely. In other words, clinics get sued for underdiagnosis but not for overdiagnosis and overtreatment.

Bayes rule and the confusion of conditional probabilities

One way of rewriting Bayes rule is to use a ratio of probabilities, $P(A)/P(B)$, to relate the two conditional probabilities of interest, $P(B|A)$ (test sensitivity) and $P(A|B)$ (positive predictive value):

$$P(A|B) = P(B|A)\ \frac{P(A)}{P(B)}\ .$$

With this rewriting, it is obvious that $P(A|B)$ and $P(B|A)$ will be equal only when the prior probabilities, $P(A)$ and $P(B)$, are the same. Yet, this confusion error is so common in the forensic literature that it is given the special name of the "Prosecutor's Fallacy." In the behavioral sciences research literature, this "Prosector's Fallacy" is sometimes called the "Fallacy of the Transposed Conditional" or the "Inversion Fallacy." In the context of statistical inference, it appears when the probability of seeing a particular data result conditional on the null hypothesis being true, $P(\text{data} \mid H_o)$, is confused with $P(H_o \mid \text{data})$, i.e., the probability that the null hypothesis is true given that a particular data result has occurred.

As a case in point, we return to the Sally Clark conviction where the invalidly constructed probability of 1 in 73 million was used to successfully argue for Sally Clark's guilt. Let A be the event of innocence, and B the event of two "cot deaths" within the same family. The invalid probability of 1 in 73 million was considered to be for $P(B|A)$; a simple equating with $P(A|B)$, the probability of innocence given the two cot deaths, led directly to Sally Clark's conviction. We continue with the Royal Statistical Society Press Release:

> Aside from its invalidity, figures such as the 1 in 73 million are very easily misinterpreted. Some press reports at the time stated that this was the chance that the deaths of Sally Clark's two children were accidental. This (mis-)interpretation is a serious error of logic known as the Prosecutor's Fallacy.
>
> The Court of Appeal has recognised these dangers (R v. Deen 1993, R v. Doheny/Adams 1996) in connection with probabilities used for DNA profile evidence, and has put in place clear guidelines for the presentation of such evidence. The dangers extend more widely, and there is a real possibility that without proper guidance, and well-informed presentation, frequency estimates presented in court could be misinterpreted by the jury in ways that are very prejudicial to defendants.
>
> Society does not tolerate doctors making serious clinical errors because it is widely understood that such errors could mean the difference between life and death. The case of R v. Sally Clark is one example of a medical expert witness making a serious statistical error, one which may have had a profound effect on the

outcome of the case.

> Although many scientists have some familiarity with statistical methods, statistics remains a specialised area. The Society urges the Courts to ensure that statistical evidence is presented only by appropriately qualified statistical experts, as would be the case for any other form of expert evidence.

The situation with Sally Clark and the Collins case in California (where both involved the Prosecutor's Fallacy), is not isolated. There was the recent miscarriage of justice in The Netherlands involving a nurse, Lucia de Berk, accused of multiple deaths at the hospitals where she worked. This case aroused the international community of statisticians to redress the apparent ills visited upon Lucia de Berk. One source for background (although now somewhat dated) is Mark Buchanan at the *The New York Times* Blogs (*The Prosecutor's Fallacy*, May 16, 2007). The Wikipedia article on "Lucia de Berk" provides the details of the case and the attendant probabilistic arguments, up to her complete exoneration in April of 2010.

A much earlier and historically important *fin de sciele* case, is that of Alfred Dreyus, the much maligned French Jew, and Captain in the military, who was falsely imprisoned for espionage. In this instance, the nefarious statistician was the rabid anti-Semite Alphonse Bertillon, who through a convoluted argument, reported a very small probability that Dreyfus was "innocent"; this meretricious probability had no justifiable mathematical basis and was generated from culling coincidences involving a document, the handwritten *bordereau* (without signature) announcing the transmission of French military information. Dreyfus was accused and convicted of penning this document and passing it to the (German) enemy. The "Prosecutor's Fallacy" was more or less invoked to ensure a conviction based on the fallacious small probability given by Bertillon. In addition to Emile Zola's famous article, *J'Accuse*, in the newspaper *L'Aurore* on January 13, 1898, it is interesting to note that well-known turn-of-the-century statisticians and probabilists from the French Academy of Sciences (among them Henri Poincairé) demolished Bertillon's probabilistic arguments, and insisted that any use of such evidence needs to proceed in a fully Bayesian manner, much like our present understanding of evidence in current forensic science and the proper place of probabilistic argumentation. A detailed presentation of all the probabilistic and statistical issues and misuses present in the Dreyfus case is given by Champod, Taroni, and Margot (1999). (Also, see the comprehensive text by Aitken and Taroni [2004], *Statistics and the Evaluation of Evidence for Forensic Scientists.*)

We observe the same general pattern in all of the miscarriages of justice involving the Prosecutor's Fallacy. There is some very small reported probability of "innocence," typically obtained incorrectly either by culling, misapplying the notion of statistical independence, or using an inappropriate statistical model. Such a probability is calculated by a supposed expert with some credibility in court: a community college mathematics instructor for Collins; Roy Meadow for Clark; Henk Elffers for de Berk; Alphonse Bertillon for Dreyfus. The Prosecutor's Fallacy then takes place, leading to a conviction for the crime. Various outrages ensue from the statistically literate community, with the eventual emergence of some "statistical good guys" hoping to redress the wrongs done: an unnamed court-appointed statistician for the California Supreme Court for Collins; Richard Gill for de Berk; Henri Poincairé (among others) for Dreyfus; the Royal Statistical Society for Clark. After long periods of time, convictions are eventually overturned, typically after extensive prison sentences have already been served. We can only hope to avoid similar miscarriages of justice in cases yet to come by recognizing the tell-tale pattern of occurence for the Prosecutor's Fallacy.

There seem to be any number of conditional probability confusions that can arise in important contexts (and possibly when least expected). A famous instance of this is in the O.J. Simpson case, where one conditional probability, say, $P(A|B)$, was confused with another, $P(A|B \text{ and } D)$. We quote the clear explanation of this obfuscation by Krämer and Gigerenzer (2005, p. 228):

> Here is a more recent example from the U.S., where likewise $P(A|B)$ is confused with $P(A|B \text{ and } D)$. This time the confusion is spread by Alan Dershowitz, a renowned Harvard Law professor who advised the O.J. Simpson defense team. The prosecution had argued that Simpson's history of spousal abuse reflected a motive to kill, advancing the premise that "a slap is a prelude to homicide." Dershowitz, however, called this argument "a show of weakness" and said: "We knew that we could prove, if we had to, that an infinitesimal percentage – certainly fewer than 1 of 2,500 – of men who slap or beat their domestic partners go on to murder them." Thus, he argued that the probability of the event K that a husband killed his wife if he battered her was small, $P(K|\text{battered}) = 1/2{,}500$. The relevant probability, however, is not this one, as

Dershowitz would have us believe. Instead, the relevant probability is that of a man murdering his partner given that he battered her and that she was murdered, $P(K|\text{battered and murdered})$. This probability is about 8/9. It must of course not be confused with the probability that O.J. Simpson is guilty; a jury must take into account much more evidence than battering. But it shows that battering is a fairly good predictor of guilt for murder, contrary to Dershowitz's assertions.

The Basic Sampling Model and Related Issues

From *The New York Times* article by David Stout (April 3, 2009), *Obama's Census Choice Unsettles Republicans*:

Robert M. Groves, a former census official and now a sociology professor at the University of Michigan, was nominated Thursday by President Obama to run the Census Bureau, a choice that instantly made Republicans nervous.

Republicans expressed alarm because of one of Mr. Groves's specialties, statistical sampling – roughly speaking, the process of extrapolating from the numbers of people actually counted to arrive at estimates of those uncounted and, presumably, arriving at a realistic total.

If minorities, immigrants, the poor and the homeless are those most likely to be missed in an actual head count, and if political stereotypes hold true, then statistical sampling would presumably benefit the Democrats.

Republicans have generally argued that statistical sampling is not as reliable as its devotees insist. ‘Conducting the census is a vital constitutional obligation,’ Representative John A. Boehner of Ohio, the House minority leader, said Thursday. ‘It should be as solid, reliable and accurate as possible in every respect. That is why I am concerned about the White House decision to select Robert Groves as director of the Census Bureau.’

Mr. Boehner, recalling that controversy (from the early 1990s when Mr. Groves pushed for statistically adjusting the 1990 census to make up for an undercount), said Thursday that ‘we will have to watch closely to ensure the 2010 census is conducted without attempting similar statistical sleight of hand.’

We begin by refreshing our memories about the distinctions between *population* and *sample*; *parameters* and *statistics*; *population distributions* and *sampling distributions.* Someone who has successfully completed a year-long graduate sequence in statistics should know these distinctions very well. Here, only a simple univariate framework is considered explicitly, but obvious and straightforward generalizations exist for the multivariate context.

A *population* of interest is posited, operationalized by some random variable, say X. In this *Theory World* framework, X is characterized by *parameters*, such as the expectation of X, $\mu = \mathrm{E}(X)$, or its variance, $\sigma^2 = \mathrm{V}(X)$. The random variable X has a *(population) distribution*, which is often assumed normal. A *sample* is generated by taking observations on X, say, $X_1, \ldots, X_n$, considered independent and identically distributed as X, i.e., they are exact copies of X. In this *Data World* context, statistics are functions of the sample, and therefore, characterize the sample: the sample mean, $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$; the sample variance, $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2$, with some possible variation in dividing by $n-1$ to generate an unbiased estimator for σ^2. The statistics, $\hat{\mu}$ and $\hat{\sigma}^2$, are *point estimators* of μ and σ^2. They are random variables by themselves, so they have distributions called *sampling distributions.* The general problem of statistical inference is to ask what sample statistics, such as $\hat{\mu}$ and $\hat{\sigma}^2$, tell us about their population counterparts, μ and σ^2. In other words, can we obtain a measure of accuracy for estimation from the sampling distributions through, for example, confidence intervals?

Assuming that the population distribution is normally distributed, the sampling distribution of $\hat{\mu}$ is itself normal with expectation μ and variance σ^2/n. Based on this result, an approximate 95% confidence interval for the unknown parameter μ, can be given by

$$\hat{\mu} \ \pm \ 2.0\frac{\hat{\sigma}}{\sqrt{n}} \ .$$

Note that it is the square root of the sample size that determines the length of the interval (and not the sample size per se). This is both good news and bad. Bad, because if you want to double precision, you need a four-fold increase in sample size; good, because sample size can be cut by four with only a halving of precision.

Even when the population distribution is not originally normally distributed, the Central Limit Theorem (CLT) says that $\hat{\mu}$ is approximately normal in form and becomes exactly so as n goes to infinity. Thus, the approximate confidence interval statement remains valid even when the underlying distribution is not normal; such a result underlies many claims of robustness, i.e., when a procedure remains valid even if the assumption under which it was derived may not be true, as long as some particular condition is satisfied — here, that condition is for the sample size to be reasonably large. Although how large is big enough for a normal approximation to be adequate depends generally on the form of the underlying population distribution, a glance at a "t-table" will show that when the degrees-of-freedom are larger than 30, the values given are indistinguishable from that for the normal. Thus, we surmise that sample sizes above 30 should generally be large enough to invoke the benefits that the CLT provides.

Besides the robustness of the confidence interval calculations for μ, the CLT also encompasses what is called the Law of Large Numbers (LLN). As the sample size increases, the estimator, $\hat{\mu}$, gets closer and closer to μ, and converges to μ at the limit of n going to infinity. This is seen most directly in the sampling variance for $\hat{\mu}$, which gets smaller as the sample size gets larger.

The basic results obtainable from the CLT and LLN that averages are both less variable and more normal in distribution than individual observations, and that averages based on larger sample sizes will show less variability than those based on smaller sample sizes, have far ranging and sometimes very subtle influences on our reasoning skills. For example, suppose we would like to study organizations, such as schools, health care units, or governmental agencies, and have some measure of performance on the individuals in the units, and the average for each unit. To identify those units exhibiting best performance (or, in the current jargon, "best practice"), the top 10%, say, of units in terms of performance are identified; a determination is then made of what common factors might characterize these top-performing units. We are pleased when able to isolate one very salient feature — most units in this top tier are small; we proceed on this observation to advise in the break-up of larger units. Is such a policy really justified based on these data? Probably not, if one also observes that the bottom 10% are also small units. Given that smaller entities just tend to be inherently more variable than the larger, would seem to vitiate a recommendation of breaking-up the larger units for performance improvement. Evidence that the now defunct "small schools movement," funded heavily by the Gates Foundation, was a victim of the "square root of n law," was presented by Wainer (2009, Chapter 1).

Another implication of the basic sampling model is that when the size of the population is effectively infinite, this does not affect the accuracy of our estimate, which is driven by sample size. Thus, if we want a more precise estimate, we need only draw a larger sample. For some reason, this confusion resurfaces and is reiterated every ten years when the U.S. Census is planned, where the issues of complete enumeration, as demanded by the Constitution, and the problems of undercount are revisited. The beginning quotations from John Boehner in relation to the 2010 census is a good case in point. And the ethical implications of his statistical reasoning skills should be fairly clear.

An area of almost mythic proportions in which a misunderstanding, or at least, a misappreciation for randomness exists, is in sports. A reasonable model for sports performance is one of "observed performance" being the sum of "intrinsic ability" (or true performance) and "error," leading to natural variability in outcome either at the individual or the team level. Somehow it appears necessary for sports writers, announcers, and other pundits, to continually give reasons for what is most likely just random variability. We hear of team "chemistry," good or bad, being present or not; individuals having a "hot hand" (or a "cold hand," for that matter); someone needing to "pull out of a slump"; why there might be many more .400 hitters early in the season but not later; a player being "due" for a hit; free-throw failure because of "pressure"; and so on. Making decisions based on natural variation being somehow "predictive" or "descriptive" of the truth, is not very smart, to say the least. But it is done all the time — sports managers are fired and CEOs replaced for what may be just the traces of natural variability.

In asking people to generate random sequences, they tend to underestimate the amount of variation present in such a stochastic process — not enough (longer) runs are present; there is a tendency to produce two many short alternations; and so on. In a similar way, we do not see the naturalness in what will be called in a later section, regression toward the mean – where extremes are followed by less extreme observations just because of fallibility in observed performance. And again, causes are sought. We hear about multi-round golf tournaments where a good performance on the first day is followed by a less adequate score the second (due probably to "pressure"); or a bad performance on the first day followed by an improved performance the next (he/she must have been able to "play loose"). Or in baseball, at the start of a season, an under-performing Derek Jeter might be under "pressure" or too much “media scrutiny”, or the difficulties of performing in a “New York Market.” When an individual starts off well but then appears to fade, it must be people trying to stop him/her (i.e., “gunning” for someone). One should always remember that in estimating intrinsic ability, an individual is unlikely to be as good (or as bad) as the pace they are on. It is always a better bet to vote against someone eventually breaking some record, even when they are “on a pace” to so do early in the season. This may be one origin for the phrase “sucker bet” — a gambling wager where your expected return is significantly lower than your wager.

Another area where one expects to see a lot of anomalous results is when the data set is split into ever finer categorizations that end up having very few observations in them, and thus subject to much greater variability. For example, should we be overly surprised if Albert Pujols doesn’t seem to bat well in domed stadiums at night when batting second against left-handed pitching? The pundits look for “causes” for these kinds of extremes when they should just be marveling at the beauty of natural variation and the effects of sample size. A similar and probably more important misleading effect occurs when our data are on the effectiveness of some medical treatment, and we try to attribute positive or negative results to ever finer-grained classifications of our clinical subjects.

Random processes are a fundamental part of nature and are ubiquitous in our day-to-day lives. Most people do not understand them, or worse, fall under an “illusion of control,” where one believes they have influence over how events progress. Thus, we have almost a mystical belief in the ability of a new coach, CEO, or President to “turn things around.” Part of these strong beliefs may result from the operation of regression toward the mean, or the natural unfolding of any random process. We continue to get our erroneous beliefs reconfirmed when we attribute cause when none may actually be present. As humans we all wish to believe we can affect our future, but when events have dominating stochastic components, we are obviously not in complete control. There appears to be a fundamental clash between our ability to recognize the operation of randomness and the need for control in our lives.

Correlation

The association between two variables measured on the same set of objects is commonly referred to as their correlation and often measured by the Pearson Product Moment Correlation Coefficient. Specifically, suppose $Z_{X_1}, \ldots, Z_{X_N}$ and $Z_{Y_1}, \ldots, Z_{Y_N}$ refer to Z-scores (i.e., having mean zero and variance one) calculated for our original observational pairs, (X_i, Y_i), $i = 1, \ldots, N$; then the correlation between the original variables, r_{XY}, is defined as

$$r_{XY} = (\frac{1}{N}) \sum_{i=1}^{N} Z_{X_i} Z_{Y_i} \ ,$$

or the average product of the Z-scores. As usually pointed out early in any statistics sequence, r_{XY} measures the linearity of any relation that might be present; thus, if some other (nonlinear) form of association exists, different means of assessing it are needed.

In any reasoning based on the presence or absence of a correlation between two variables, it is imperative that graphical mechanisms be used in the form of scatterplots. One might go so far to say that if only the value of r_{XY} is provided and nothing else, we have a *primae facie* case of statistical malpractice. Scatterplots are of major assistance in a number of ways: (1) to ascertain the degree to which linearity might be the type of association present between the variables; this assessment could take the form of directly imposing various scatterplot smoothers and using these to help characterize the association present, if any; (2) to identify outliers or data points that for whatever reason are not reflective of the general pattern exhibited in the scatterplot, and to hopefully figure out why; (3) to provide a graphical context for assessing the influence of a data point on a correlation, possibly by the size and/or color of a plotting symbol, or contour lines indicating the change in value for the correlation that would result if it were to be removed.

One of the most shopworn adages we hear in any methodology course is that "correlation does not imply causation." It is usually noted that other "lurking" or third variables might affect both X and Y, producing a spurious association; also, because r_{XY} is a symmetric measure of association, there is no clue in its value as to the directionality of any causal relationship. For example, we have had some recent revisions in our popular views on the positive effects of moderate drinking; it may be that individuals who otherwise lead healthy lifestyles, also drink moderately. Or in a football sports context, “running the ball” does not cause winning; it is more likely that winning causes “running the ball.” Teams that get an early lead try to run the ball frequently because it keeps the clock running and decreases the time for an opponent to catch up.

In any multiple variable context, it is possible to derive the algebraic restrictions present among some subset of the variables based on the given correlations for another subset. The simplest case involves three variables, say X, Y, and W. From the basic formula for the partial correlation between X and Y “holding” W constant, an *algebraic* restriction is present on r_{XY} given the values of r_{XW} and r_{YW}:

$$r_{XW}r_{YW} - \sqrt{(1-r_{XW}^2)(1-r_{YW}^2)} \le r_{XY} \le r_{XW}r_{YW} + \sqrt{(1-r_{XW}^2)(1-r_{YW}^2)} \ .$$

Note that this is not a probabilistic statement (i.e., it is not a confidence interval); it says that no data set exists where the correlation r_{XY} lies outside of the upper and lower bounds provided by $r_{XW}r_{YW} \pm \sqrt{(1-r_{XW}^2)(1-r_{YW}^2)}$. As a numerical example, suppose X and Y refer to height and weight, respectively, and W is some measure of age. If, say, the correlations, r_{XW} and r_{YW} are both .8, then $.28 \le r_{XY} \le 1.00$. In fact, if a high correlation value of .64 were observed for r_{XY}, should we be impressed about the magnitude of the association between X and Y? Probably not; if the partial correlation between X and Y “holding” W constant were computed with $r_{XY} = .64$, a value of zero would be obtained. All of the observed high association between X and Y can be attributed to their association with the developmentally driven variable. These very general restrictions on correlations have been known for a very long time, and appear, for example, in Yule’s First Edition (1911) of *An Introduction to the Theory of Statistics* under the title: *Conditions of Consistence Among Correlation Coefficients.* Also, in this early volume, see Yule’s chapter on *Fallacies in the Interpretation of Correlation Coefficients.*

A related type of algebraic restriction for a correlation is present when the distribution of the values taken on by the variables, include ties. In the extreme, consider a 2 × 2 contingency table, and the four-fold point correlation; this is constructed by using a 0/1 coding of the category information on the two attributes, and calculating the usual Pearson correlation. Because of the "lumpy" marginal frequencies present in the 2×2 table, the four-fold correlation cannot extend over the complete ±1 range. The achievable bounds possible can be computed (see Carroll, 1961); it may be of some interest descriptively to see how far an observed four-fold correlation is away from its achievable bounds, and possibly, to even normalize the observed value by such a bound.

The bounds of ±1 on a Pearson correlation can be achieved only by data sets demonstrating a perfect linear relationship between the two variables. Another measure that achieves the bounds of ±1 whenever the data sets merely have consistent rank orderings, is Guttman's (weak) monotonicity coefficient, μ_2:

$$\mu_2 = \frac{\sum_{i=1}^{n}\sum_{h=1}^{n}(x_h - x_i)(y_h - y_i)}{\sum_{i=1}^{n}\sum_{h=1}^{n}|x_h - x_i||y_h - y_i|} \ ,$$

where (x_h, y_h) denote the pairs of values being "correlated" by μ_2. The coefficient, μ_2, expresses the extent to which values on one variable increase in a particular direction as the values on another variable increases, without assuming that the increase is exactly according to a straight line. It varies between -1 and $+1$, with $+1$ [-1] reflecting a perfect monotonic trend in a positive [negative] direction. The adjective "weak" refers to the untying of one variable without penalty. In contrast to the Pearson correlation, μ_2 can equal $+1$ or -1, even though the marginal distributions of the two variables differ from one another. When the Pearson correlation is $+1.00$ or -1.00, μ_2 will have the same value; in all other cases, the absolute value of μ_2 will be higher than that of the Pearson correlation including the case of a four-fold point correlation. Here, μ_2 reduces to what is called Yule's Q (which is a special case of the Goodman-Kruskal Gamma statistic for a 2×2 contingency table [a measure of rank-order consistency]).

There are several other correlational pitfalls that seem to occur in various forms whenever we try to reason through data sets involving multiple variables. We briefly mention four of these areas in the sections to follow.

Illusory correlation

An illusory correlation is present whenever a relationship is seen in data where none exists. Common examples would be between membership in some minority group and rare and typically negative behavior, or in the endurance of stereotypes and an overestimation of the link between group membership and certain traits. Illusory correlations seem to depend on the novelty or uniqueness of the variables considered. Some four decades ago, Chapman and Chapman (1967, 1969) studied such false associations in relation to psychodiagnostic signs seen in projective tests. For example, in the "Draw-a-Person" test, a client draws a person on a blank piece of paper. Some psychologists believe that drawing a person with big eyes is a sign of paranoia. Such a correlation is illusionary but very persistent. When data that are deliberately uncorrelated are presented to college students, the same diagnostic signs are found that some psychologists still believe in. It is of some historical interest to know that this very same notion of illusory correlation has been around since the early 1900s – see, for example, Yule's First Edition (1911) of *An Introduction to the Theory of Statistics*, and the chapter entitled: *Illusory Associations.*

There are several faulty reasoning relatives for the notion of an illusory correlation. One is *confirmation bias* where there are tendencies to search for, interpret, and remember information only in a way that confirms one's preconceptions or working hypotheses. No one will soon forget the country's collective confirmation bias in identifying "weapons of mass destruction" in the run-up to the Iraq war; this is related to the "I'm not stupid" fallacy that rests on the belief that if one is mistaken, one must therefore be stupid, and we generally believe that we are not stupid – witness the prosecutor who refuses to drop charges against an obviously innocent suspect because otherwise, he or she would need to admit error and wasted effort. At an extreme, we have (the trap of) *apophenia*, or seeing patterns or connections in random or meaningless data. A subnotion is *pareidola*, where vague and random stimuli (often images or sounds) are perceived as significant, e.g., the Virgin Mary is seen on a grilled cheese sandwich. One particular problematic realization of apophenia is in epidemiology when residential cancer-clusters are identified that rarely if ever result in identifiable causes. What seems to be occurring is sometimes labeled the Texas Sharpshooter Fallacy

– like a Texas sharpshooter who shoots at the side of a barn and then draws a bull's-eye around the largest cluster of bullet holes. In residential cancer-clusters, we tend to notice cases first, e.g., multiple cancer patients on the same street, and then define the population base around them. A particularly well-presented piece on these illusory associations is by Atul Gawande in the February 8th, 1998, *New Yorker*: *The Cancer-Cluster Myth*.

Ecological correlation

An ecological correlation is one calculated between variables that are group means, in contrast to obtaining a correlation between variables measured at an individual level. There are several issues with the use of ecological correlations: they tend to be a lot higher than individual-level correlations, and assuming what is seen at the group level also holds at the level of the individual is so pernicious, it has been labeled the "ecological fallacy" by Selvin (1958). The term "ecological correlation" was popularized from a 1950 paper by William Robinson (Robinson, 1950), but the idea has been around for some time (e.g., see the 1939 paper by E. L. Thorndike, *On the Fallacy of Imputing Correlations Found for Groups to the Individuals or Smaller Groups Composing Them*). Robinson computed a correlation of .53 between literacy rate and the proportion of the population born outside the U.S. for the 48 states of the 1930 census. At the individual level, however, the correlation was $-.11$, so immigrants were on average less literate than their native counterparts. The high ecological correlation of .53 was due to immigrants settling in states with a more literate citizenry. A recent discussion of ecological correlation issues in our present political climate, is the entertaining (at least for statisticians) piece in the *Quarterly Journal of Political Science* by Gelman et al. (2007): *Rich State, Poor State, Red State, Blue State: What's the Matter with Connecticut*. An expansion of this article in book form is Gelman et al. (2010; *Red State, Blue State, Rich State, Poor State: Why Americans Vote the Way They Do (Expanded Edition)*.

A problem related to ecological correlation is the modifiable areal unit problem (MAUP), where differences in spatial units used in the aggregation can cause wide variation in the resulting correlations (e.g., anywhere from minus to plus 1.0). Generally, the manifest association between variables depends on the size of areal units used, with increases as areal unit size gets larger. A related "zone" effect concerns the variation in correlation caused by reaggregating data into different configurations at the same scale. Obviously, the MAUP has serious implications for our abilities to reason with data: when strong relationships exist between variables at an individual level, these can be obscured through aggregation; conversely, aggregation can lead to apparently strong association when none is actually present. A thorough discussion of the modifiable unit problem appears in Yule and Kendall (1968; Fourteenth Edition).

Restriction of range for correlations

The famous psychologist, Clark Hull, noted in 1928 that psychological tests did not predict job performance very well, with correlations rarely above .30. The implication taken was that tests could never be of much use in personnel selection because job performance could not be predicted very well. In one of the most famous papers in all of Industrial and Organizational Psychology, Taylor and Russell (1939) responded to Hull, noting the existence of the restriction of range problem: in a group selected on the basis of some test, the correlation between test and performance must be lower than it would be in an unselected group. Taylor and Russell provided tables and charts for estimating an unselected from the selected correlation based on how the selection was done (the infamous Taylor-Russell Charts).

An issue related to the restriction of range in its effect on correlations, is the need to deal continually with fallible measurement. Generally, the more unreliable our measures, the lower (or more attenuated) the correlations. The field of psychometrics has for some many decades provided a mechanism for assessing the effects of fallible measurement through its "correction for attenuation": the correlation between "true scores" for our measures is the observed correlation divided by the square roots of their reliabilities. Various ways are available for estimating reliability, so implementing attenuation corrections is an eminently feasible enterprise. Another way of stating this correction is to note that any observed correlation must be bounded above by the square root of the product of the reliabilities.

Obviously, if reliabilities are not very good, observed correlations can never be very high.

Another type of range restriction problem (see Figure 1) is observed in the empirical fact of a negative correlation between Law School Admission Test (LSAT) scores and Undergraduate Grade Point Average (UGPA) within almost all law schools. Does this mean that the worse you perform in college courses, the better you will do on the LSAT? Well, no; it is because if you did very well on both, you went to Harvard, and if you did poorly on both, you didn't get into law school. So at all other law schools, there were admittees who did relatively better on one than on the other. A graph of the LSAT scores versus UGPA shows thin bands running from upper left to lower right representing each law school, with the better schools higher up on both; the overall picture, however, is a very positive data swirl with the lower triangle not admitted.

Odd correlations

A recent article (Vul et al. 2009) in a journal from the Association for Psychological Science, *Perspectives on Psychological Science*, has the intriguing title of *Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition* (renamed from the earlier and more controversial, *Voodoo Correlations in Social Neuroscience*). These authors comment on the extremely high (e.g., $> .8$) correlations reported in the literature between brain activation and personality measures, and point out the fallaciousness of how they were obtained. Typically, huge numbers of separate correlations were calculated, and only the mean of those correlations exceeding some threshold (based on a very small significance level) are reported. It is tautological that these correlations selected for size must be large in their average value. With no cross-validation attempted to see the shrinkage expected in these measures on new samples, we have sophistry at best. Any of the usual understanding of yardsticks provided by the correlation or its square, the proportion of shared variance, are inappropriate. In fact, as noted by Vul et al. (2009), these inflated mean correlations typically exceed the upper bounds provided by the correction for attenuation based on what the reliabilities should be for the measures being correlated.

When a correlation reported in the literature seems odd, it is incumbent on a literate consumer of such information to understand why. Sometimes it is as simple as noting the bias created by the selection process as in the fMRI correlations, and that such selection is not being mitigated by any cross-validation. Or, possibly, inflated or deflated association measures may occur because of the use of ecological correlations or modifiable areal units, restriction of range, the fallibility of the behavioral measures, the presence of a nonlinear relationship, and so on. The reason behind apparent correlational artifacts can be subtle and require a careful explication of the processes leading to the measures being correlated and on what objects. For instance, if correlations are being monitored over time, and the group on which the correlations are based changes composition, the effects could be dramatic. Such composition changes might be one of different sex ratios, immigrant influxes, economic effects on the available workforce, age, and so on. One particularly unusual example is discussed by Dawes (1975) on the relation between graduate admission variables and future success. Because admission criteria tend to be compensatory (where good values on certain variables can make up for not so good values on others), the covariance structure among admissions variables in the selected group is unusual in that it involves negative correlations. As argued nicely by Dawes, it must be the case that the variables used to admit graduate students have low correlation with future measures of success.

A related odd correlational effect (see Figure 2) occurs in graduate admissions for departments that specialize in technical subjects – there is a negative correlation of performance in graduate school (as judged by faculty ratings) and Graduate Record Examination – Verbal (GRE-V) scores. Does this imply that faculty judge badly? Or that the poorer your English proficiency, the better you will do in graduate school? The answer is more subtle and is generated by the large number of students with foreign (often Chinese) backgrounds, whose performance on the GRE-V may be relatively poor, but who do well in graduate school. This interpretation is confirmed when we condition on the binary variable 'Native English Speaker' or 'Not' and find that the correlation is strongly positive within either of the two classes. Again, this becomes clear with a graph that shows two tight ovals at different heights corresponding to the two language groups, but the overall regression line runs across the two ovals and in the opposite direction.

Prediction

The attempt to predict the values on some (dependent) variable by a function of (independent) variables is typically approached by simple or multiple regression, for one and more than one predictor, respectively. The most common combination rule is a linear function of the independent variables obtained by least-squares, i.e., the linear combination minimizes the sum of the squared residuals between the actual values on the dependent variable and those predicted from the linear combination. In the case of simple regression, scatterplots again play a major role in assessing linearity of the relationship, the possible effects of outliers on the slope of the least-squares line, and the influence of individual objects in its calculation. The regression slope, in contrast to the correlation, is neither scale invariant nor symmetric in the dependent and independent variables. One usually interprets the least-squares line as one of expecting, for each unit change in the independent variable, a regression slope change in the dependent variable.

There are several topics in prediction that arise continually when we attempt to reason ethically with fallible multivariable data. We discuss briefly four such areas in the subsections to follow: regression toward the mean; the distinction between actuarial (statistical) and clinical prediction; methods involved in using regression for prediction that incorporate corrections for unreliability; and differential prediction effects in selection based on tests.

Regression toward the mean

Regression toward the mean is a phenomenon that will occur whenever dealing with (fallible) measures with a less-than-perfect correlation. The word "regression" was first used by Sir Francis Galton in his 1886 paper, *Regression Toward Mediocrity in Hereditary Stature*, where he showed that heights of children from very tall or short parents would regress toward mediocrity (i.e., toward the mean) — exceptional scores on one variable (parental height) would not be matched with such exceptionality on the second (child height). This observation is purely due to the fallibility for the various measures (i.e., the lack of a perfect correlation between the heights of parents and their children).

Regression toward the mean is a ubiquitous phenomenon, and given the name "regressive fallacy" whenever cause is ascribed where none exists. Generally, interventions are undertaken if processes are at an extreme, e.g., a crackdown on speeding or drunk driving as fatalities spike; treatment groups formed from individuals who are seriously depressed; individuals selected because of extreme behaviors, both good or bad; and so on. In all such instances, whatever remediation is carried out will be followed by some more moderate value on a response variable. Whether the remediation was itself causative is problematic to assess given the universality of regression toward the mean.

There are many common instances where regression may lead to invalid reasoning: I went to my doctor and my pain has now lessened; I instituted corporal punishment and behavior has improved; he was jinxed by a *Sports Illustrated* cover because subsequent performance was poorer (i.e., the "sophomore jinx"); although he hadn't had a hit in some time, he was "due", and the coach played him; and on and on. More generally, any time one optimizes with respect to a given sample of data by constructing prediction functions of some kind, there is an implicit use and reliance on data extremities. In other words, the various measures of goodness-of-fit or prediction we might calculate need to be cross-validated either on new data or a clever sample reuse strategy such as the well-known jackknife or bootstrap procedures. The degree of "shrinkage" we see in our measures based on this cross-validation, is an indication of the fallibility of our measures and the adequacy of the given sample sizes.

The misleading interpretive effects engendered by regression toward the mean are legion, particularly when we wish to interpret observational studies for some indication of causality. There is a continual violation of the old adage that "the rich get richer and the poor get poorer," in favor of "when you are at the top, the only way is down." Extreme scores are never quite as extreme as they first appear. Many of these regression artifacts are explicated in the cautionary source, *A Primer on Regression Artifacts* (Campbell and Kenny, 2002), including the various difficulties encountered in trying to equate intact groups by matching or analysis-of-covariance. Statistical equating creates the illusion but not the reality of equivalence. As summarized by Campbell and Kenny, "the failure to understand the likely direction of bias when statistical equating is used, is one of the most serious difficulties in contemporary data analysis."

There are a variety of phrases that seem to get attached whenever regression toward the mean is probably operative. We have the “winner’s curse”, where someone is chosen from a large pool (e.g., of job candidates), who then doesn’t live up to expectation; or when we attribute some observed change to the operation of “spontaneous remission.” As Campbell and Kenny note: “many a quack has made a good living from regression toward the mean.” Or, when a change of diagnostic classification results upon repeat testing for an individual given subsequent one-on-one tutoring (after being placed, for example, in a remedial context); Or, more personally, there is “editorial burn out” when someone is chosen to manage a prestigious journal at the apex of one’s career, and things go quickly downhill from that point forward.

Actuarial versus clinical prediction

Paul Meehl in his classic 1954 monograph, *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*, created quite a stir with his convincing demonstration that mechanical methods of data combination, such as multiple regression, outperform (expert) clinical prediction. The enormous amount of literature produced since the appearance of this seminal contribution, has uniformly supported this general observation; similarly, so have the extensions suggested for combining data in ways other than by multiple regression, e.g., by much simpler unit weighting schemes (Wainer, 1976), or those using other prior weights. It appears that individuals who are conversant in a field are better at selecting and coding information than they are at actually integrating it. Combining such selected information in some more mechanical manner will generally do better than the person choosing such information in the first place. This conclusion can be pushed further: if we formally model the predictions of experts using the same chosen information, we can generally do better than the experts themselves. Such formal representations of what a judge does, are called “paramorphic”.

In an influential review paper, Dawes (1979) discussed what he called proper and improper linear models, and argued for the “robust beauty of improper linear models.” A proper linear model is one obtained by some optimization process, usually least-squares; improper linear models are not "optimal" in this latter sense, and typically have their weighting structures chosen by a simple mechanism, e.g., random or unit weighting. Again, improper linear models generally outperform clinical prediction, but even more surprisingly, improper models typically outperform proper models in cross-validation. What seems to be the reason, is the notorious instability of regression weights with correlated predictor variables, even if sample sizes are very large. Generally, we know that simple averages are more reliable than individual observations, so it may not be so surprising that simple unit weights are likely to do better on cross-validation than those found by squeezing "optimality" out of a sample. Given that the *sine qua non* of any prediction system is its ability to cross-validate, the lesson may be obvious — statistical optimality with respect to a given sample may not be the best answer when we wish to predict well.

The idea that statistical optimality may not lead to the best predictions, seems counterintuitive, but as argued well by Roberts and Pashler (2000), just the achievement of a good fit to observations does not necessarily mean we have found a good model. In fact, because of the overfitting of observations, choosing the model with the absolute best fit is apt to result in poorer predictions. The more flexible the model, the more likely it is to capture not only the underlying pattern but unsystematic patterns such as noise. A single general purpose tool with many adjustable parameters is prone to instability and greater prediction error as a result of high error variance. An observation by John von Neumann is particulary germane: "With four parameters, I can fit an elephant, and with five, I can make him wiggle his trunk." More generally, this notion that "less-is-more" is difficult to get one's head around, but as Gigerenzer and others have argued (e.g., see Gigerenzer and Brighton, 2009), it is clear that simple heuristics, such as "take the best," can at times be more accurate than complex procedures. All of the work emanating from the idea of the "robust beauty of improper linear models" and sequelae may force some reassessment of what the normative ideals of rationality might be; most reduce to simple cautions about overfitting one's observations, and then hoping for better predictions because an emphasis has been placed on immediate optimality instead of the longer-run goal of cross-validation.

Incorporating reliability corrections in prediction

There are two aspects of variable unreliability in the context of prediction that might have consequences for ethical reasoning. One is in estimating a person's true score on a variable; the second is in how regression might be handled when there is measurement error in the independent and/or dependent variables. In both of these instances, there is an implicit underlying model for how any observed score, X, might be constructed additively from a true score, T_X, and an error score, E_X, where E_X is typically assumed uncorrelated with T_X: $X = T_X + E_X$. When we consider the distribution of an observed variable over, say, a population of individuals, there are two sources of variability present in the true and the error scores. If we are interested primarily in structural models among true scores, then some correction must be made because the common regression models implicitly assume that variables are measured without error.

The estimation, $\hat{T}_X$, of a true score from an observed score, X, was derived using the regression model by Kelley in the 1920's (see Kelley, 1947), with a reliance on the observation that the squared correlation between observed and true score is the reliability. If we let $\hat{\rho}$ be estimated reliability, Kelley's equation can be written as $\hat{T}_X = \hat{\rho}X + (1 - \hat{\rho})\bar{X}$, where $\bar{X}$ is the mean of the group to which the individual belongs. In other words, depending on the size of $\hat{\rho}$, a person's estimate is compensated for by where they are in relation to the group — upwards if below the mean; downwards if above. The application of this statistical tautology in the examination of group differences provides such a surprising result to the statistically naive, that this equation has been called "Kelley's Paradox" (Wainer, 2005, Chapter 10). We might note that this notion of being somewhat punitive of performances better than the group to which one supposedly belongs, was not original with Kelley, but was known at least 400 years earlier; in the words of Miguel de Cervantes (1547–1616): "Tell me what company you keep and I'll tell you what you are."

In the topic of errors-in-variables regression, we try to compensate for the tacit assumption in regression that all variables are measured without error. Measurement error in a response variable does not bias the regression coefficients per se, but it does increase standard errors, and thereby reducing power. This is generally a common effect: unreliability attenuates correlations and reduces power even in standard ANOVA paradigms. Measurement error in the predictor variables biases the regression coefficients. For example, for a single predictor, the observed regression coefficient is the "true" value multiplied by the reliability coefficient. Thus, without taking account of measurement error in the predictors, regression coefficients will generally be underestimated, producing a biasing of the structural relationship among the true variables. Such biasing may be particularly troubling when discussing econometric models where unit changes in observed variables are supposedly related to predicted changes in the dependent measure; possibly the unit changes are more desired at the level of the true scores.

Differential prediction effects in selection

One area in which prediction is socially relevant is in selection based on test scores, whether for accreditation, certification, job placement, licensure, educational admission, or other high-stakes endeavors. We note that most of these discussions about fairness of selection need to be phrased in terms of regression models relating a performance measure to a selection test; and whether the regressions are the same over all identified groups of relevance, e.g., ethnic, gender, age, and so on. Specifically, are slopes and intercepts the same; if so or if not, how does this affect the selection mechanism being implemented, and whether it can be considered fair. It is safe to say that depending on the pattern of data within groups, all sorts of things can happen; generally, an understanding of how a regression/selection model works with this kind of variation, is necessary for a literate discussion of its intended or unintended consequences. To obtain a greater sense of the complications that can arise, the reader is referred to Allen and Yen (2001; Chapter 4.4, *Bias in Selection*).

Data Presentation and Interpretation

The goal of statistics is to gain understanding from data; the methods of presentation and analyses used should not only allow us to "tell the story" in the clearest and fairest way possible, but more primarily, to help learn what the story is in the first place. When results are presented, there is a need to be sensitive to the common and maybe not so common missteps that result from a superficial understanding and application of the methods in statistics. It is insufficient to just "copy and paste" without providing context for how good or bad the methods are that are being used, and understanding what is behind the procedures producing the numbers. We will present in this introductory section some of the smaller pitfalls to be avoided; a number of larger areas of concern will be treated in separate subsections:

1) Even very trivial differences will be "significant" when sample sizes are large enough. Also, significance should never be confused with importance; the current emphasis on the use of confidence intervals and the reporting of effect sizes, reflects this point. (For a further discussion of this topic, see Cumming and Fidler [this volume, *Hypothesis testing to parameter estimation: An example of evidence-based practice in statistics*].)

2) As some current textbooks still report inappropriately, a significance test does not evaluate whether a null hypothesis is true. A p-value measures the "surprise value" of a particular observed result conditional on the null hypothesis being true.

3) Degrees-of-freedom do not refer to the number of independent observations within a data set; the term indicates how restricted the quantities are that are being averaged in computing various statistics, e.g., sums of squares between or within groups.

4) Although the Central Limit Theorem comes to the assistance of robustness issues when dealing with means, the same is not true for variances. The common tests on variances are notoriously nonrobust and should never be used; robust alternatives are available in the form of sample-reuse methods such as the jackknife and bootstrap.

5) Do not carry out a test for equality of variances before performing a two-independent samples t-test. A quote, usually attributed to George Box, comments on the good robustness properties of the t-test in relation to the nonrobustness of the usual tests for variances: "to test for equality of variances before carrying out an independent samples t-test, is like putting a row boat out on the ocean to see if it is calm enough for the Queen Mary."

6) Measures of central tendency and dispersion, such as the mean and variance, are not *resistant* in that they are influenced greatly by extreme observations; the median and interquartile range, on the other hand, are resistant, and each observation counts the same in the calculation of the measure.

7) Do not ignore the repeated measures nature of your data and just use methods appropriate for independent samples. For example, don't perform an independent samples t-test on "before" and "after" data in a time-series intervention study. Generally, the standard error of a mean difference must include a correction for correlated observations, as routinely done in a paired (matched samples) t-test. (For more development of these issues, see Goldstein [this volume, *Multilevel modeling*].)

8) The level of the measurement model used for your observations limit the inferences that are meaningful. For example, interpreting the relative sizes of differences makes little sense on data measured with a model yielding only nominal or ordinal level characteristics.

9) Do not issue blanket statements as to the impossibility of carrying out reasonable testing, confidence interval construction, or cross-validation. It is almost always now possible to use resampling methods that do not rely on parametric models or restrictive assumptions, and which are computer implemented for immediate application. The appropriate statement is not that "this can't be done", but rather, "I don't know how to do this as yet."

10) Keep in mind the distinctions between fixed and random effects models and the differing test statistics they may necessitate. The output from some statistical package may use a default understanding of how the factors are to be interpreted. If your context is different, then appropriate calculations must be made, sometimes "by hand." To parody the Capital One Credit Card commercial: "What's in your denominator?"

11) Do not report all of the eight or so decimal places given in typical computer output. Such false precision (or spurious accuracy) is a dead give-away that you really don't know what you are doing. Two decimal places are needed at most, and often, only one is really justified. As an example, consider how large a sample is required to support the reporting of a correlation to more than one decimal place (answer: given the approximate standard error of $\frac{1}{\sqrt{n}}$, a sample size greater than 400 would be needed to give a 95% confidence interval of $\pm$ 0.1).

12) It is wise generally to avoid issuing statements that might appear to be right, but with some deeper understanding, are just misguided:

a) "given the huge size of a population, it is impossible to achieve accuracy with a sample"; this reappears regularly with the discussion of undercount and the census.

b) "it is incumbent on us to always divide by $n - 1$ when calculating a variance to give the 'best' estimator"; well, if you divide by $n + 1$, the estimator has a smaller expected error of estimation, which to many is more important than just being "unbiased." Also, why is it that no one ever really worries that the usual correlation coefficient is a "biased" estimate of its population counterpart?

c) "ANOVA is so robust that all of its assumptions can be violated at will"; although it is true that normality is not that crucial if sample sizes are reasonable in size (and the CLT is of assistance), and homogeneity of variances doesn't really matter as long as cell sizes are close, the independence of errors assumption is critical and one can be lead very far astray when it doesn't hold — for intact groups; spatial contexts; repeated measures. (Again, for further discussion, see Goldstein [this volume, *Multilevel modeling*].)

d) don't lament the dearth of one type of individual from the very upper scores on some test without first noting possible differences in variability. Even though mean scores may be the same for groups, those with even slightly larger variances will tend to have more representatives in both the upper and lower echelons.

13) Avoid using one-tailed tests. Even the carriers of traditional one-tailed hypotheses, the chi-square and F distributions, have two tails, and both ought to be considered. The logic of hypothesis testing is that if an event is sufficiently unlikely, we must reconsider the truth of the null hypothesis. Thus, for example, if an event falls in the lower tail of the chi-square distribution, it implies that the model fits too well. If investigators had used two-tailed tests, the data manipulations of Cyril Burt may have been uncovered much earlier.

In concluding these introductory comments about the smaller missteps to be avoided, we note the observations of Edward Tufte on the ubiquity of PowerPoint (PP) for presenting quantitative data, and the degradation it produces in our ability to communicate (Tufte, 2006, p. 26, his italics):

> *The PP slide format has the worst signal/noise ratio of any known method of communication on paper or computer screen.* Extending PowerPoint to embrace paper and internet screens pollutes those display methods.

Generally PowerPoint is poor at presenting statistical evidence, and is no replacement for more detailed technical reports, data handouts, and the like. It is now part of our "pitch culture," where, for example, we are sold on what drugs to take, but are not provided with the type of detailed numerical evidence we should have for an informed decision about benefits and risks. In commenting on the incredible obscuration of important data that surrounded the use of PowerPoint-type presentations to give the crucial briefings in the first Shuttle accident of Challenger in 1986, Richard Feyman noted (reported in Tufte, 2006, p. 17):

> Then we learned about 'bullets' – little black circles in front of phrases that were supposed to summarize things. There was one after another of these little goddamn bullets in our briefing books and on slides.

Multivariable systems

Whenever results are presented within a multivariate context, it is important to remember there is a system present among the variables, and this has a number of implications for how we proceed: automated systems that cull through collections of independent variables to locate the "best" regression equations (e.g., by forward selection, backward elimination, or the hybrid of stepwise regression), are among the most misused statistical methods available in all the common software packages. They offer a false promise of blind theory-building without user intervention, but the incongruities present in their use are just too great for this to be a reasonable strategy of data analysis: a) one does not necessarily end up with the "best" prediction equations for a given number of variables; b) different implementations of the process don't necessarily end up with the same equations; c) given that a system of interrelated variables is present, the variables not selected cannot be said to be unimportant; d) the order in which variables enter or leave in the process of building the equation does not necessarily reflect their importance; e) all of the attendant significance testing and confidence interval construction methods become completely inappropriate (see Freedman, 1983).

Several methods, such as the use of Mallow's C_p statistic for "all possible subsets (of the independent variables) regression," have some possible mitigating effects on the heuristic nature of the blind methods of stepwise regression. They offer a process of screening all possible equations to find the better ones, with compensation for the differing numbers of parameters that need to be fit. Although these search strategies offer a justifiable mechanism for finding the "best" according to ability to predict a dependent measure, they are somewhat at cross-purposes for how multiple regression is typically used in the behavioral sciences. What is important is in the structure among the variables as reflected by the regression, and not so much in squeezing the very last bit of variance-accounted-for out of our methods. More pointedly, if we find a "best" equation with fewer than the maximum number of available independent variables present, and we cannot say that those not chosen are less important than those that are, then what is the point?

A more pertinent analysis was demonstrated by Efron and Gong (1983) in which they bootstrapped the entire model-building process. They showed that by viewing the frequency with which each independent variable finds its way into the model, we can assess the stability of the choice of variables. Examining the structure of the independent variables through, say, a principal component analysis, will alert us to irreducible uncertainty due to high covariance among predictors. This is always a wise step, done in conjunction with bootstrapping, but not instead of it.

The implicit conclusion of the last argument extends more generally to the newer methods of statistical analysis that seem to continually demand our attention, e.g., in hierarchical linear modeling, nonlinear methods of classification, procedures that involve optimal scaling, and so on. When the emphasis is solely on getting better "fit" or increased prediction capability, and thereby, modeling "better," the methods may not be of much use in "telling the story" any more convincingly. And that should be the ultimate purpose of any analysis procedure we choose. Also, as Roberts and Pashler (2000) note rather counterintuitively, "goodness-of-fit" does not necessarily imply "goodness-of-model."

Even without the difficulties presented by a multivariate system when searching through the set of independent variables, there are several admonitions to keep in mind when dealing just with a single equation. The most important may be to remember that regression coefficients cannot be interpreted in isolation for their importance using their sizes, even when based on standardized variables (i.e., those that have been Z-scored). Just because one coefficient is bigger than another, does not imply it is therefore more important. For example, consider the task of comparing the relative usefulness of the Scholastic Aptitude

Test (SAT) scores and High School Grade Point Averages (HSGPA) in predicting freshmen college grades. Both independent variables are highly correlated; so when grades are predicted with SAT scores, a correlation of about 0.7 is found. Correlating the residuals from this prediction with HSGPA, gives a small value. It would be a mistake to conclude from this that SAT is a better predictor of college success than HSGPA. If the order of analysis is reversed, we would find that HSGPA correlates about 0.7 with freshmen grades and the residuals from this analysis have only a small correlation with SAT score.

If we must choose between these two variables, or try to evaluate a claim that one variable is more important than another, it must be from some other basis. For example, SAT scores are like the product of an experiment; they can be manipulated and improved. Flawed test items can be discovered and elided. But HSGPA is like the result of an observational study; they are just found, lying on the ground. We are never sure exactly what they mean. If one teacher harbors a secret bias, and gives students of a particular ilk grades that do not represent their true accomplishments, how are we to know? There are some formal methods that can at times help reduce our ignorance. We will discuss them next, but first remember that no formal procedure guarantees success in the face of an unthinking analysis.

The notion of importance may be explored by comparing models with and without certain variables present, and comparing the changes in variance-accounted-for that ensue. Similarly, the various significance tests for the regression coefficients are not really interpretable independently, e.g., a small number of common factors may underlie all the independent variables, and thus, generate significance for all the regression coefficients. In its starkest form, we have the one, two, and three asterisks scattered around in a correlation matrix, suggesting an ability to evaluate each correlation by itself without consideration of the multivariable system that the correlation matrix reflects in its totality. Finally, for a single equation, the size of the squared multiple correlation (R^2) gets inflated by the process of optimization, and needs to be adjusted, particularly when sample sizes are small. One beginning option is to use the commonly generated Wherry "adjusted R^2," which makes the expected value of R^2 zero when the true squared multiple correlation is itself zero. Note that the name of "Wherry's shrinkage formula" is a misnomer because it is not a measure based on any process of cross-validation. A cross-validation strategy is now routine in software packages, such as SYSTAT, using the “hold out one-at-a-time” mechanism. Given the current ease of implementation, such cross-validation processes should be routinely carried out.

Graphical presentation

The importance of scatterplots in evaluating the association between variables was reiterated several times in our earlier discussions of correlation and prediction. Generally, graphical and other visual methods of data analysis are central to an ability to tell what data may be reflecting and what conclusions are warranted. In a time when graphical presentation may have been more expensive than it is now, it was common to only use summary statistics, even when various reporting rules were followed, e.g., “never present just a measure of central tendency without a corresponding measure of dispersion.” Or, in providing the results of a poll, always give the margin of error (usually, the 95% confidence interval) to reflect the accuracy of the estimate based on the sample size being used. If data are not nicely unimodal, however, more is needed than just means and variances. Both “stem-and-leaf” and “box-and-whisker” plots are helpful in this regard, and should be routinely used for data presentation.

Several egregious uses of graphs for misleading presentations were documented many years ago in the very popular book by Darrell Huff, *How to Lie with Statistics* (1954), and up-dated in Wainer’s oft-cited 1984 classic from *The American Statistician*, *How to Display Data Badly* (also, see Chapter 1 in Wainer, 1997/2000). Both of these deal with visual representation and how graphs can be used to distort, e.g., by truncating bottoms of line or bar charts, so differences are artificially magnified, or using two- and three-dimensional objects to compare values on a unidimensional variable where images do not scale the same way as do univariate quantities. Tufte (e.g., see Tufte, 1983) has lamented on the poor use of graphics that use “chart junk” for questionable visual effect, or gratuitous color or three-dimensions in bar graphs that do not represent anything real at all. In extending some of these methods of misrepresentation to the use of maps, it is particularly easy to deceive given the effects of scale level usage, ecological correlation, and the modifiable areal unit problem. What should be represented generally in our graphs and maps must be as faithful as possible to the data represented, without the distracting application of unnecessary frills that do not communicate any information of value.

There is one particularly insidious use of a graphical format that almost always misleads: the double y-axis plot. In this format there are two vertical axes, one on the left and one on the right depicting two completely different variables — say death rates over time for smokers shown on the left axis (time is on the horizontal axis), and death rates for non-smokers shown on the right axis. Because the scale on the two vertical axes are independent, they can be chosen to show anything the graph maker wants. Compare the first version in Figure 3 (after the Surgeon General's report on the dangers of smoking), to the second in Figure 4, prepared by someone attentive to the needs of big tobacco that uses the double y-axis format. Few other graphic formats lend themselves so easily to the misrepresentation of quantitative phenomena.

In providing data in the form of matrices, such as subject by variable, we should consider the use of "heat maps," where numerical values, assumed commensurable over variables, are mapped into color spectra reflecting magnitude. The further imposing of nondestructively obtained orderings on rows and columns to group similar patches of color together, can lead to useful data displays. A survey of the history of heat maps, particularly as developed in psychology, has been given by Wilkinson and Friendly (2009); this latter article should be mandatory reading in any part of a statistics course concerned with accurate and informative graphical data presentation. Also, see Bertin (1973/1983); Tufte (1983, 1990, 1996); Tukey (1977); Wainer (1997, 2005, 2009).

Problems with multiple testing

A difficulty encountered with the use of automated software analyses is that of multiple testing, where the many significance values provided are all given as if each were obtained individually without regard for how many tests were actually performed. This situation gets exacerbated when the "significant" results are then culled, and only these are used in further analysis. A good case in point was reported earlier in the section on odd correlations where highly inflated correlations get reported in fMRI studies because an average is taken only over those correlations selected to have reached significance according to a stringent threshold. Such a context is a clear violation of a dictum given in any beginning statistics class: you cannot legitimately test a hypothesis on the same data that first suggested it.

Exactly the same issue manifests itself, although in a more subtle, implicit form, in the modern procedure known as data mining. Data mining consists of using powerful graphical methods to view high-dimensional data sets of moderate-to-large size, looking for interesting features. When such a feature is uncovered, it is isolated and saved – a finding! Implicit in the search, however, are many, many comparisons that the viewer makes and decides are not interesting. Because the searching and comparing is done in real-time, it is difficult to keep track of how many insignificant comparisons were discarded before alighting on a significant one. Without knowing how many, we cannot judge the significance of the interesting features found without an independent confirmatory sample. Such independent confirmation is too rarely done.

To be more formal about the problem of multiple testing, suppose there are K hypotheses to test, $H_1, \ldots, H_K$, and, for each, we set the criterion for rejection at the fixed Type I error value of α_k, $k = 1, \ldots, K$. If the events, $A_1, \ldots, A_K$, are defined as: A_k is the incorrect rejection of H_k (i.e., rejection when it is true), the Bonferroni inequality gives:

$$P(A_1 \text{ or } \cdots \text{ or } A_K) \leq \sum_{k=1}^{K} P(A_k) = \sum_{k=1}^{K} \alpha_k \ .$$

Noting that the event (A_1 or $\cdots$ or A_K) can be verbally restated as one of "rejecting incorrectly *one or more* of the hypotheses," the experimentwise (or overall) error-rate is bounded by the sum of the K alpha values set for each hypothesis. Typically, we set $\alpha_1 = \cdots = \alpha_K = \alpha$, and the bound is then $K\alpha$. Thus, the usual rule for controlling the overall error rate through the Bonferroni correction sets the individual alphas at some small value, e.g., $.05/K$; the overall error rate is then guaranteed to be no larger than .05.

The problems of multiple testing and the failure to practice "safe statistics," appears in both blatant and more subtle forms. For example, companies may suppress unfavorable studies until those to their liking occur. There is a possibly apocryphal story that toothpaste companies promoting fluoride in their products in the 1950's, did repeated studies until large effects could be reported for their "look Ma, no cavities" television campaigns. This may be somewhat innocent advertising hype for toothpaste, but when drug or tobacco companies engage in the practice, it is not so innocent and can have a serious impact on our health. It is important to know how many things were tested to assess the importance of those reported. For example, when given only those items from some inventory or survey that produced significant differences between groups, be very wary!

In the framework of multiple testing, there are a number of odd behaviors that people sometimes engage in. We list a few of these below in summary form:

a) it is not legitimate to do a Bonferroni correction post-hoc, i.e., find a set of tests that lead to significance, and then evaluate just this subset with the correction;

b) Scheffe's method (and relatives) are the only true post-hoc procedures to control the overall error rate. An unlimited number of comparisons can be made (no matter whether identified from the given data or not), and the overall error rate remains constant;

c) you cannot look at your data to decide which planned comparisons to do;

d) Tukey's method is not post-hoc because you actually plan to do all possible pairwise comparisons;

e) even though the comparisons you might wish to test are independent (e.g., they are defined by orthogonal comparisons), the problem of inflating the overall error rate remains; similarly, in performing a multifactor ANOVA or testing multiple regression coefficients, all of the tests carried out should have some type of overall error control imposed;

f) it makes no sense to perform a multivariate analysis of variance before you then go on to evaluate each of the component variables one-by-one. Typically, a multivariate-analysis-of-variance (MANOVA) is completely noninformative as to what is really occurring, but people proceed in any case to evaluate the individual univariate ANOVAs irrespective of what occurs at the MANOVA level — we may not reject the null hypothesis at the overall MANOVA level, but then illogically ask where the differences are at the level of the individual variables. Plan to do the individual comparisons beforehand, and avoid the typically noninterpretable overall MANOVA test completely.

We cannot, in good conscience, leave the important topic of multiple comparisons without at least a mention of what is now considered the most useful method — the False Discovery Rate (Benjamini & Hochberg, 1995). But even this strategy is not up to the most vexing of problems of multiplicity. We have already mentioned data mining as one of these; a second arises in the search for genetic markers. A typical paradigm in this crucial area is to isolate a homogeneous group of individuals, some of whom have a genetic disorder and others do not, and then to see if one can determine which genes are likely to be responsible. One such study is currently being carried out with a group of 200 Mennonites in Pennsylvania. Macular degeneration is common among the Mennonites, and this sample was chosen so that 100 of them had macular degeneration and a matched sample of 100 did not. The genetic structure of the two groups was very similar and so the search was on to see which genes were found much more often in the group that had macular degeneration than in the control group. This could be determined with a t-test. Unfortunately, the power of the t-test was diminished considerably when it had to be repeated for more than 100,000 separate genes. The Bonferroni inequality was no help, and the False Discovery Rate, while better, was still not up to the task. The search still goes on to find a better solution to the vexing problem of multiplicity.

(Mis)reporting of data

The Association for Psychological Science publishes a series of timely monographs on *Psychological Science in the Public Interest.* One recent issue was from Gerd Gigerenzer and colleagues, entitled: *Helping Doctors and Patients Make Sense of Health Statistics* (Gigerenzer et al. 2008); it details some issues of statistical literacy as it concerns health, both our own individually as well as societal health policy more generally. Some parts of being statistically literate may be fairly obvious — we know that just making up data, or suppressing information even of supposed outliers without comment, is unethical. The topics touched upon by Gigerenzer et al. (2008), however, are more subtle; if an overall admonition is needed, it is that "context is always important," and the way data and information are presented is absolutely crucial to an ability to reason appropriately and act accordingly. We touch on several of the major issues raised by Gigerenzer et al. in the discussion to follow.

We begin with a quote from Rudy Giuliani from a New Hampshire radio advertisement that aired on October 29, 2007, during his run for the Republican Presidential nomination (this example was also used by Gigerenzer et al.):

> I had prostate cancer, five, six years ago. My chances of surviving prostate cancer and thank God I was cured of it, in the United States, 82 percent. My chances of surviving prostate cancer in England, only 44 percent under socialized medicine.

Not only did Giuliani not receive the Republican Presidential nomination, he was just plain wrong on survival chances for prostate cancer. The problem is a confusion between survival and mortality rates. Basically, higher survival rates with cancer screening do not imply longer life.

To give a more detailed explanation, we define a five-year survival rate and an annual mortality rate:

five-year survival rate = (number of diagnosed patients alive after five years)/(number of diagnosed patients);

annual mortality rate = (number of people who die from a disease over one year)/(number in the group).

The inflation of a five-year survival rate is caused by a *lead-time bias*, where the time of diagnosis is advanced (through screening) even if the time of death is not changed. Moreover, such screening, particularly for cancers such as prostate, leads to an *overdiagnosis bias* — the detection of a pseudodisease that will never progress to cause symptoms in a patient's lifetime. Besides inflating five-year survival statistics over mortality rates, overdiagnosis leads more sinisterly to overtreatment that does more harm than good (e.g., incontinence, impotence, and other health related problems).

It is important to keep in mind that screening does not "prevent cancer," and early detection does not diminish the risk of getting cancer. One can only hope that cancer is caught, either by screening or other symptoms, at an early enough stage to help. It is also relevant to remember that more invasive treatments are not automatically more effective. A recent and informative summary of the dismal state and circumstances surrounding cancer screening generally, appeared in *The New York Times*, page one and "above the fold," article by Natasha Singer (Friday, July 17, 2009), *In Push for Cancer Screening, Limited Benefits.*

A major area of concern in the clarity of reporting health statistics, is in how the data are framed as relative risk reduction or as absolute risk reduction, with the former usually seeming much more important than the latter. As examples that present the same information:

relative risk reduction — if you have this test every two years, it will reduce your chance of dying from the disease by about one third over the next ten years.

absolute risk reduction — if you have this test every two years, it will reduce your chance of dying from the disease from 3 in 1000 to 2 in 1000, over the next ten years.

We also have a useful variant on absolute risk reduction given by its reciprocal, the *number needed to treat* — if 1000 people have this test every two years, one person will be saved from dying from the disease every ten years.

Because bigger numbers garner better headlines and more media attention, it is expected that relative rather than absolute risks are the norm. It is especially disconcerting, however, to have potential benefits (of drugs, screening, treatments, and the like) given in relative terms, but harm in absolute terms that is typically much smaller numerically. The latter has been called "mismatched framing" by Gigerenzer and colleagues (2008).

An ethical presentation of information avoids nontransparent framing of information, whether unintentional or intentional. Intentional efforts to manipulate or persuade people are particularly destructive, and unethical, by definition. As Tversky and Kahneman (e.g., 1981) have noted many times in their published contributions, framing effects and context have major influences on a person's decision processes. Whenever possible, give measures that have operational meanings with respect to the sample at hand (e.g., the Goodman-Kruskal γ), and avoid measures that do not, such as odds-ratios. This advice is not always followed (see, for example, the Institute of Medicine's, *2008 National Healthcare Disparities Report*, in which the efficacy of medical care is compared across various groups in plots with the odds-ratio as the dependent variable. As might be expected, this section's impact on the public consciousness was severely limited).

In a framework of misreporting data, we have the all-to-common occurrence of inflated (and sensational) statistics intended to have some type of dramatic effect. As noted succinctly by Joel Best in his article, *Lies, Calculations and Constructions* (*Statistical Science*, *20*, 2005, 210–214): "Ridiculous statistics live on, long after they've been thoroughly debunked; they are harder to kill than vampires." We typically see a three-stage process in the use of inflated statistics: first, there is some tale of atrocity (think Roman Polanski's *Rosemary's Baby*); the problem is then given a name (e.g., the presence of satanic cults in our midst); and finally, some inflated and, most likely, incorrect statistic is given that is intended to alarm (e.g., there are well over 150,000 active satanic cults throughout the United States and Canada).

Another issue in the reporting of data is when the context for some statement is important but is just not given (or is suppressed), resulting in a misinterpretation (or at least, an over-interpretation). These examples are legion and follow the types illustrated below:

a) The chances of a married man becoming an alcoholic are double those of a bachelor because 66% of souses are married men. (This may not be so dramatic when we also note that 75% of all men over 20 are married.)

b) Among 95% of couples seeking divorce, either one or both do not attend church regularly. (This example needs some base-rate information to effect a comparison, e.g., what is the proportion of couples generally, where one or both do not attend church regularly.)

c) Over 65% of all accidents occur with 25 miles of home, and at a speed of 40 miles per hour or less. (An obvious question to ask is where most of one's driving is done.)

d) Hector Luna, who went 2-for-5 and raised his average to .432, had his fourth straight multi-hit game for the Cardinals, who have won six of seven overall (Associated Press; St. Louis Cardinals vs. Pittsburgh Pirates, April 26, 2004). (Reporting of data should provide a context that is internally consistent; here, the word "raised" is odd.)

Pitfalls of software implementations

Most of our statistical analyses are now done through the use of packages such as SYSTAT, SPSS, or SAS. Because these systems are basically blind to what data you may be analyzing and what questions you may want to ask, it is up to the user to know some of the pitfalls to avoid. For example, just because an ANCOVA is extremely easy to do, doesn't mean it should be done or that it is possible to legitimately equate intact groups statistically. Also, just because output may be provided, doesn't automatically mean it should be used. Cases in point are the inappropriate reporting of indeterminate factor scores, the gratuitous number of decimal places typically given, Durbin-Watson tests when the data are not over time, uninformative overall MANOVAs, nonrobust tests for variances, and so on. We mention two more general traps we've seen repeatedly, and which need to be recognized to avoid embarrassment:

a) In the construction of items or variables, the numbers assigned may at times be open to arbitrary keying. For instance, instead of using a 1 to 10 scale, where '1' means 'best' and '10' 'worst', the keying could be reversed so '1' means 'worst' and '10' best. When an intercorrelation matrix is obtained among a collection of variables subject to this kind of scoring arbitrariness, it is possible to obtain some pretty impressive (two-group) structures in methods of multidimensional scaling and cluster analysis that are merely artifacts of the keying and not of any inherent meaning in the items themselves. In these situations, it is common to "reverse score" a subset of the items, so hopefully an approximate "positive manifold" is obtained for the correlation matrix, i.e., there are few if any negative correlations that can't be attributed to just sampling error. (The topic of reverse scoring for the ubiquitous Likert scales is noted, at least in passing, in a variety of measurement sources; one recent and readable account is given by Dunn-Rankin et al. [2004].)

b) There are certain methods of analysis (e.g., most forms of multidimensional scaling, K-means and mixture model cluster analyses, and some strategies involving optimal scaling) that are prone to local optima, i.e., a result is presented but one that is not the best possible according to the goodness-of-fit measure being optimized. The strategies used in the optimization are not able to guarantee global optimality because of the structure of the functions being optimized (e.g., those that are highly nonconvex). One standard method of local optimality exploration is to repeatedly start (randomly) some specific analysis method, and observe how bad the local optima problem is for a given data set, and to choose the best analysis found for reporting a final result. Unfortunately, none of the current packages (SPSS, SAS, SYSTAT) offer these random start options for all the methods that may be prone to local optima (for a good case in point involving K-means clustering, see Steinley, 2003). These local optimality difficulties are one of the reasons for allowing more than the closed analysis systems in graduate statistics instruction, and the general move (or maybe, we should say rush) toward using environments such as MATLAB and R (or at least to choose packages that allow an exploration of local optima, e.g., MPlus includes a facility for supplying sets of random starting values for model-based mixture analyses).

The ease to which analyses can be done with closed statistical systems requiring little-or-no understanding of what the "point-and-clicks" are really giving, may at times be more of an impediment to clear reasoning than it is of assistance. The user does not need to know much before being swamped with copious amounts of output, and with little or no help on how to wade through the results, or when necessary, to engage in further exploration (e.g., in investigating local minima or alternative analyses). One of the main reasons for now employing some of the newer statistical environments (such a R and MATLAB) is that they do not rely on pull-down menus to do one's thinking; instead, they are built up from functions that take various inputs and provide outputs — but you need to know what to ask for, and the syntax of the function being used. Also, the source code for the routines is available and can be modified if some variant of an analysis is desired — again, this assumes more than a superficial understanding of how the methods work; these are valuable skills to have when attempting to reason from data. The R environment has become the *lingua franca* for framing cutting-edge statistical development and analysis, and is becoming *the* major computational tool we need to develop in the graduate-level statistics sequence. It is also open-source and free, so there are no additional instructional costs incurred with the adoption of R.

Simpson's Paradox

In the presentation of multiway contingency tables, an unusual phenomenon occurs so frequently it has been given the label of "Simpson's Paradox" (Simpson, 1951; Yule, 1903).

Basically, various relationships that appear to be present when data are conditioned on the levels of one variable, either disappear or change "direction" when aggregation occurs over the levels of the conditioning variable. A well-known real-life example is the Berkeley sex bias case involving women applying to graduate school (see Bickel, Hammel, & O'Connell, 1975). The table below shows the aggregate admission figures for the Fall of 1973:

	Number of Applicants	Percent Admitted
Men	8442	44%
Women	4321	35%

There appears to be a *primae facie* case for bias given the lower rate of admission for women as compared to men.

Although there appears to be bias at the aggregate level, the situation becomes less clear once the data are broken down by major (these data are for only the top six majors in number of applicants; the numbers therefore do not add to those in the previous table):

Major	Men		Women	
	Applicants	%Admitted	Applicants	%Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	272	6%	341	7%

Here, no department is significantly biased against women, and in fact, most have a small bias against men; Simpson's paradox has occurred! Apparently, based on the table presented above, women tend to apply to competitive departments with lower rates of admission among qualified applicants (e.g., English); men tend to apply to departments with generally higher rates of admission (e.g., Engineering).

A different example showing a similar point can be given using data on the differential imposition of a death sentence depending on the race of the defendant and the victim. These data are from twenty Florida counties during 1976-7; our source is Radelet (1981), but they are repeated in many categorical data analysis texts (e.g., see Agresti, 2007):

Defendant	Death:Yes	Death:No
White	19 (12%)	141
Black	17 (10%)	149

Because 12% of White defendants receive the Death penalty and only 10% of Blacks, at this aggregate level there appears to be no bias against Blacks. But when the data are disaggregated, the situation appears to change:

Victim	Defendant	Death:Yes	Death:No
White	White	19 (13%)	132
White	Black	11 (17%)	52
Black	White	0 (0%)	9
Black	Black	6 (6%)	97

Because 12% of which defendants receive the death penalty and only 10% of blacks, at this aggregate level, there appears to be no bias against blacks. But when the data are disaggregated, the situation appears to change, for when we condition on the race of the victim, in both cases the Black defendant has the higher probability of receiving the death sentence compared to the White defendant (17% to 13% for White victims; 6% to 0% for Black victims). The conclusion one can reach is disconcerting: the value of a victim is worth more if White than if Black, and because more Whites kill Whites, at the aggregate level, there appears to be a slight bias against Whites. But for both types of victims, Blacks are more likely to receive the death penalty.

Although not explicitly a Simpson's Paradox context, there are similar situations that appear in various forms of multifactor analysis-of-variance that raise cautions about aggregation phenomena. The simplest dictum is that "you cannot interpret main effects in the presence of interaction." Some softening of this admonition is usually given when the interaction is not disordinal, and where the graphs of means don't actually cross. In these instances it may be possible to eliminate the interaction by some relatively simple transformation of the data, and produce an "additive" model. Because of this, non-crossing interactions might be considered "unimportant." Similarly, the absence of parallel profiles (i.e., when interaction is present) may hinder the other tests for the main effects of coincident and horizontal profiles. Possibly, if the profiles again show only an "unimportant" interaction, such evaluations could proceed.

Although Simpson's Paradox has been known by this name only rather recently (as coined by Colin Blyth in 1972), the phenomenon has been recognized and discussed for well over a hundred years; in fact, it has a complete textbook development in Yule's, *An Introduction to the Theory of Statistics*, first published in 1911. In honor of Yule's early contribution (Yule, 1903), we sometimes see the title of the Yule-Simpson effect.

Some Concluding Remarks

A graduate course in statistics hopefully prepares students in a number of areas that have immediate implications for the practice of ethical reasoning. We review six broad areas in this concluding section that should be part of any competently taught sequence in the behavioral sciences: 1) formal tools to help think through ethical situations; 2) a basic understanding of the psychology of reasoning and how it may differ from that based on a normative theory of probability; 3) how to be (dis)honest in the presentation of information, and to avoid obfuscation; 4) some ability to ferret out specious argumentation when it has a supposed statistical basis; (5) the deleterious effects of culling in all its various forms (e.g., the identification of "false positives"), and the subsequent failures to either replicate or cross-validate; (6) identifying plausible but misguided reasoning from data, or from other information presented graphically.

One of the trite quantitative sayings that may at times drive individuals "up a wall" is when someone says condescendingly, "just do the math." Possibly, this saying can become a little less obnoxious when reinterpreted to mean working through a situation formally rather than just giving a quick answer based on first impressions that may be wrong. An example of this may help: In 1990, Craig Whitaker wrote a letter to Marilyn vos Savant and her column in *Parade* magazine stating what has been called the Monte Hall problem:

> Suppose you're on a game show, and you're given the choice of three doors. Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, 'Do you want to pick door No. 2?' Is it to your advantage to switch your choice?

The answer almost universally given to this problem is that switching does not matter, presumably with the reasoning that there is no way for the player to know which of the two unopened doors is the winner, and each of these must then have an equal probability of being the winner. By "doing the math," however, possibly writing down three doors hiding one car and two goats, and working through the options in a short simulation, it becomes clear quickly that the opening of a goat door changes the information one has about the original situation, and that always changing doors doubles the probability of winning from 1/3 to 2/3. (As an interesting historical note, the "Monte Hall" problem has been a fixture of probability theory from at least the 1890's; it is called the problem of the "three caskets" by Henri Poincairé, and is more generally known as (Joseph) Bertrand's Box Paradox.)

Any beginning statistics class should always include a number of formal tools to help "do the math." Several of these have been mentioned in early sections: Bayes theorem and implications for screening using sensitivities, specificities, and prior probabilities; conditional probabilities more generally and how probabilistic reasoning might work for facilitative and inhibitive events; sample sizes and variability in, say, a sample mean, and how a confidence interval might be constructed that could be made as accurate as necessary by just increasing the sample size, and without any need to consider the exact size (assumed to be large) of the original population of interest; how statistical independence operates or doesn't; the pervasiveness of natural variability and the use of simple probability models (as the binomial) to generate stochastic processes; the computations involved in corrections for attenuation; usage of Taylor-Russell charts.

A second area of interest in developing statistical literacy and learning to reason ethically, is the large body of work produced by psychologists regarding the normative theory of choice and decisions derivable from probability theory, and how it may not be the best guide to the actual reasoning processes that individuals engage in. The Nobel Prize level contributions of Tversky and Kahneman (e.g., 1971, 1974, 1981) are particularly germane, and the view that people rely on various simplifying heuristic principles to assess probabilities and engage in judgements under uncertainty; also, that the psychology of choice is dictated to a great extent by the framing of a decision problem. We give two classic Tversky and Kahneman examples to illustrate how reasoning heuristics and framing might operate:

Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

> Which is more probable? 1. Linda is a bank teller. 2. Linda is a bank teller and is active in the feminist movement.

For one group of subjects, 85% choose option 2, even though the conjunction of two events must be less likely than either of the constituent events. Tversky and Kahneman argue that this "conjunction fallacy" occurs because the "representativeness heuristic" is being used to make the judgement — the second option seems more representative of Linda based on the description given for her.

The representativeness heuristic operates where probabilities are evaluated by the degree to which A is representative of B; if highly representative, the probability that A originates from B is assessed to be higher. When representativeness heuristics are in operation, a number of related characteristics of the attendant reasoning processes become apparent: prior probabilities (base-rates) are ignored; insensitivity develops to the operation of sample size on variability; an expectation that a sequence of events generated by some random process, even when the sequence is short, will still possess all the essential characteristics of the process itself. This leads to the "gambler's fallacy" (or, "the doctrine of the maturity of chances"), where certain events must be "due" to bring the string more in line with representativeness — as one should know, corrections are not made in a chance process but only diluted as the process unfolds. When a belief is present in the "law of small numbers," even small samples must be highly representative of the parent population — thus, researchers put too much faith in what is seen in small samples and overestimate replicability; and fail to recognize regression toward the mean because predicted outcomes should be maximally representative of the input, and therefore, be exactly as extreme.

A second powerful reasoning heuristic is *availability.* We quote from Tversky and Kahneman (1974, p. 1128):

> Lifelong experience has taught us that, in general, instances of large classes are recalled better and faster than instances of less frequent classes; that likely occurrences are easier to imagine than unlikely ones; and that the associative connections between events are strengthened when the events frequently co-occur. As a result, man has at his disposal a procedure (the availability heuristic) for estimating the numerosity of a class, the likelihood of an event, or the frequency of co-occurrences, by the ease with which the relevant mental operations of retrieval, construction, or association can be performed.

Because retrievability can be influenced by differential familiarity and saliences, the probability of an event may not be best estimated by the ease to which occurrences come to mind. A third reasoning heuristic is one of *adjustment and anchoring*, which may also be prone to various biasing effects. Here, estimates are made based on some initial value that is then adjusted.

The power of framing in how decision situations are assessed, can be illustrated well though an example and the associated discussion provided by Tversky and Kahneman (1981, p. 453):

> Problem 1 [$N = 152$]: Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimate of the consequences of the programs are as follows:
> If Program A is adopted, 200 people will be saved. [72 percent]
> If Program B is adopted, there is 1/3 probability that 600 people will be saved, and 2/3 probability that no people will be saved. [28 percent]
> Which of the two programs would you favor?
> The majority choice in this problem is risk averse: the prospect of certainly saving 200 lives is more attractive than a risky prospect of equal expected value, that is, a one-in-three chance of saving 600 lives.
>
> A second group of respondents was given the cover story of problem 1 with a different formulation of the alternative programs, as follows:
>
> Problem 2 [$N = 155$]:
> If Program C is adopted, 400 people will die. [22 percent]
> If Program D is adopted, there is 1/3 probability that nobody will die, and 2/3 probability that 600 people will die. [78 percent]
>
> Which of the two programs would you favor?
>
> The majority choice in problem 2 is risk taking: the certain death of 400 people is less acceptable than the two-in-three chance that 600 will die. The preferences in problems 1 and 2 illustrate a common pattern: choices involving gains are often risk averse and choices involving losses are often risk taking. However, it is easy to see that the two problems are effectively identical. The only difference between them is that the outcomes are described in problem 1 by the number of lives saved and in problem 2 by the number of lives lost. The change is accompanied by a pronounced shift from risk aversion to risk taking.

The effects of framing can be very subtle when certain (coded) words are used to provide salient contexts that influence decision processes, either consciously or unconsciously. A recent demonstration of this in the framework of our ongoing climate-change debate is given by Hardisty, Johnson, and Weber (2010) in the journal *Psychological Science*. The article has an interesting title: *A Dirty Word or a Dirty World? Attribute Framing, Political Affiliation, and Query Theory*; an abstract follows that was first posted online:

> Paying more for carbon-producing activities is one way to compensate for carbon dioxide emissions, but new research suggests that policymakers should be mindful of how they describe such initiatives. Volunteers were asked to choose between two identical products, one option including a surcharge for emitted carbon dioxide. When the surcharge was labeled as an "offset," the majority of volunteers chose the more expensive, environmentally friendly product. However, when the surcharge was labeled as a "tax," Republican and Independent volunteers were more likely to choose the less expensive option; Democratic volunteers' preferences did not change.

When required to reason about an individual's motives in some ethical context, it may be best to remember the operation of the *fundamental attribution error*, where people presume that actions of others are indicative of the true ilk of a person, and not just that the situation compels the behavior.

The presentation of data is an obvious area of concern when developing the basics of statistical literacy. Some aspects may be obvious, such as not making up data or suppressing analyses or information that don't conform to prior expectations. At times, however, it is possible to contextualize (or to "frame") the same information in different ways that might lead to differing interpretations. As noted in Gigerenzer et al. (2008), distinctions should be made between survival and mortality rates, absolute versus relative risks, natural frequencies versus probabilities. Generally, the presentation of information should be as honest and clear as possible. An example given by Gigerenzer et al. (2008) suggests the use of frequency statements instead of single-event probabilities, which removes the ambiguity of the reference class being referred to: instead of saying "there is a 30 to 50% probability of developing sexual problems with Prozac," use "out of every ten patients who take Prozac, 3 to 5 experience a sexual problem."

In presenting data to persuade, and because of the so-called "lead-time bias" that medical screening produces, it is unethical to promote any kind of screening based on improved five-year survival rates, or to compare such survival rates across countries where screening practices vary. As a somewhat jaded view of our current health situation, we have physicians practicing defensive medicine because there are no legal consequences for overdiagnosis and overtreatment, but only for underdiagnosis. Or, as the editor of *Lancet* commented (quoted in Gigerenzer et al. 2008): "journals have devolved into information laundering operations for the pharmaceutical industry." The ethical issues involved in medical screening and its associated consequences are socially important; for example, months after false positives for HIV, mammograms, prostate cancer, and the like, considerable and possibly dysfunctional anxiety may still exist.

A fourth statistical literacy concern is to have enough of the formal skills and context to separate legitimate claims from those that might represent more specious arguments. As examples, one should recognize when a case for cause is made in a situation where regression toward the mean is as likely an explanation, or when test unfairness is argued for based on differential performance (i.e., impact) and not on actual test bias (i.e., same ability levels performing differently). A more recent example of the questionable promotion of a methodological approach, called Optimal Data Analysis (ODA), is given in Yarnold and Soltysik (2004). We quote from the preface:

> ... to determine whether ODA is the appropriate method of analysis for any particular data set, it is sufficient to consider the following question: When you make a prediction, would you rather be correct or incorrect? If your answer is "correct," then ODA is the appropriate analytic methodology — by definition. That is because, for any given data set, ODA explicitly obtains a statistical model that yields the theoretical maximum possible level of predictive accuracy (e.g., number of correct predictions) when it is applied to those data. That is the motivation for ODA; that is its purpose. Of course, it is a matter of personal preference whether one desires to make accurate predictions. In contrast, alternative non-ODA statistical models do not explicitly yield theoretical maximum predictive accuracy. Although they sometimes may, it is not guaranteed as it is for ODA models. It is for this reason that we refer to non-ODA models as being *suboptimal*.

Sophistic arguments such as these, have no place in the legitimate methodological literature. It is not ethical to call one's method "optimal" and refer pejoratively to others as therefore "suboptimal". The simplistic approach to classification underlying "optimal data analysis" is known to not cross-validate well (see, for example, Stam, 1997); it is a huge area of operations research where the engineering effort is always to squeeze a little more out of an observed sample. What is most relevant in the behavioral sciences is stability and cross-validation (of the type reviewed in Dawes [1979] on proper and improper linear models); and to know what variables discriminate and how, and to thereby "tell the story" more convincingly and honestly.

The penultimate area of review in this concluding section is a reminder of the ubiquitous effects of searching/selecting/optimization, and the identification of "false-positives." We have mentioned some blatant examples in earlier sections — the weird neuroscience correlations; the small probabilities (mis)reported in various legal cases (such as the Dreyfus small probability for the forgery coincidences; or that for the de Berk hospital fatalities pattern); repeated clinical experimentation until positive results are reached in a drug trial — but there are many more situations that would fail to replicate; we need to be ever-vigilant of results obtained by "culling" and then presented to us as evidence.

A general version of the difficulties encountered when results are culled, is labeled the *file drawer problem*. This refers to the practice of researchers putting away studies with negative outcomes, i.e., those not reaching reasonable statistical significance or when something is found contrary to what the researchers want or expect; or those rejected by journals who will only consider publishing articles demonstrating positive and significant effects. The file drawer problem can seriously bias the results of a meta-analysis (i.e., methods for synthesizing collections of studies in a particular domain), particularly if only published sources are used (and not, for example, unpublished dissertations or all the rejected manuscripts lying on a pile in someone's office). We quote from the abstract of a fairly recent review: *The Scientific Status of Projective Techniques* (Lilienfeld, Wood, and Garb, 2000):

> Although some projective instruments were better than chance at detecting child sexual abuse, there were virtually no replicated findings across independent investigative teams. This meta-analysis also provides the first clear evidence of substantial file drawer effects in the projectives literature, as the effect sizes from published studies markedly exceeded those from unpublished studies.

The subtle effects of culling with subsequent failures to replicate can have serious consequences for the advancement of our understanding of human behavior. A recent important case in point, involves a gene-environment interaction studied by a team lead by Avshalom Caspi. A polymorphism related to the neurotransmitter serotonin was identified that apparently could be triggered to confer susceptibility to life stresses and resulting depression. Needless to say, this behavioral genetic link caused quite a stir in the community devoted to mental health research. Unfortunately, the result could not be replicated in a subsequent meta-analysis (could this possibly be due to the implicit culling over the numerous genes affecting the amount of serotonin in the brain?). Because of the importance of this cautionary tale for all behavioral genetics research, we refer the reader to a *News of the Week* item from *Science*, written by Constance Holden (June 26, 2009): *Back to the Drawing Board for Psychiatric Genetics.*

Our final concluding statistical literacy issue is the importance of developing abilities to spot and avoid falling prey to the trap of specious reasoning known as an “argument from ignorance,” or *argumentum ad ignorantiam*, where a premise is claimed to be true only because it has not been proven false, or that it is false because it has not been proven true. Sometimes this is also referred to as “arguing from a vacuum” (paraphrasing from Dawes, 1994) — what is purported to be true is supported not by direct evidence but by attacking an alternative possibility. Thus, a clinician might say: “because the research results indicate a great deal of uncertainty about what to do, my expert judgement can do better in prescribing treatment than these results.” Or to argue that people “need” drugs just because they haven’t solved their problems before taking them.

A related fallacy is “argument from personal incredulity,” where because one personally finds a premise unlikely or unbelievable, the premise can be assumed false, or that another preferred but unproven premise is true instead. In both of these instances, a person regards the lack of evidence for one view as constituting proof that another is true. Related fallacies are: (a) the *false dilemma* where only two alternatives are considered when there are, in fact, other options. The famous Eldridge Cleaver quote from his 1968 presidential campaign is a case in point: “You’re either part of the solution or part of the problem.” Or, (b) the Latin phrase *falsum in uno, falsum in omnibus* (false in one thing, false in everything) implying that someone found to be wrong on one issue, must be wrong on all others as well. In a more homey form, "when a clock strikes thirteen, it raises doubt not only to that chime, but to the twelve that came before." Unfortunately, we may have a current example of this in the ongoing climate-change debate; the one false statistic proffered by a report from the Intergovernmental Panel on Climate Change (IPCC) on Himalayan-glacier melt may serve to derail the whole science-based argument that climate change is real.

Fallacies with a strong statistical tinge related to *argumentum ad ignorantiam* would be the "margin of error folly," usually attributed to David Rogosa (the name, not the folly itself): if it could be, it is. Or, in a hypothesis testing context, if a difference isn't significant, it is zero. We now can refer to all of these reasoning anomalies under the umbrella term "truthiness," coined by Stephen Cobert from Comedy Central's, *The Cobert Report*. Here, truth comes from the gut, not books, and refers to the preferring of concepts or facts one wishes to be true, rather than concepts or facts known to be true. Thus, in 2009 we have the "birthers," who claim that Obama was not born in the United States, so constitutionally he cannot be President; or that the Health Care Bill includes "death squads" ready to "pull the plug on granny," or that there were weapons of mass destruction that justified the Iraq war; and on and on.

References

Agresti, A. (2007). *An introduction to categorial data analysis* (second edition). New York: Wiley-Interscience.

Aitken, C. G. G., & Taroni, F. (2004). *Statistics and the evaluation of evidence for forensic scientists.* Chichester, England: Wiley.

Allen, M. J., & Yen, W. M. (2001). *Introduction to measurement theory.* Prospect Heights, IL: Waveland Press.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, *57*, 289–300.

Bertin, J. (1973). *Semiologie graphique.* The Hague: Mouton-Gautier. 2nd Ed. (English translation by William Berg and Howard Wainer; published as *Semiology of graphics*, Madison, Wisconsin: University of Wisconsin Press, 1983.)

Best, J. (2005). Lies, calculations and constructions: Beyond "How to Lie with Statistics". *Statistical Science*, *20*, 210–214.

Bickel, P. J., Hammel, E. A., & O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science*, *187*, 398–404.

Blyth, C. R. (1972). On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, *67*, 364–366.

Campbell, D. T., & Kenny, D. A. (2002). *A primer on regression artifacts.* New York: Guilford Press.

Campbell, S. K. (1974). *Flaws and fallacies in statistical thinking.* Englewood Cliffs, NJ: Prentice-Hall.

Carroll, J. B. (1961). The nature of the data, or how to choose a correlation coefficient. *Psychometrika*, *26*, 347–372.

Champod, C., Taroni, F., & Margot, P.-A. (1999). The Dreyfus case — an early debate on expert's conclusions. *International Journal of Forensic Document Examiners*, *5*, 446–459.

Chapman, L. J., & Chapman, J. P. (1967). Genesis of popular but erroneous psychodiagnostic observations. *Journal of Abnormal Psychology*, *72*, 193–204.

Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology*, *74*, 271–280.

Dawes, R. M. (1975). Graduate admissions criteria and future success. *Science*, *187*, 721–723.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*, 571–582.

Dawes, R. M. (1994). *House of cards: Psychology and psychotherapy built on myth.* New York: The Free Press.

Dunn-Rankin, P., Knezek, G. A., Wallace, S. R., & Zhang, S. (2004). *Scaling methods* (Second edition). Mahwah, NJ: Lawrence Erlbaum.

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, *37*, 36–48.

Freedman, D. A. (1983). A note on screening regression coefficients. *The American Statistician*, *37*, 152–155.

Gelman, A., Shor, B., Bafumi, J., & Park, D. (2007). Rich state, poor state, red state, blue state: What's the matter with Connecticut? *Quarterly Journal of Political Science*, *2*, 345–367.

Gelman, A., et al. (2010). *Rich state, blue state, rich state, poor state: Why Americans vote the way they do (Expanded edition).* Princeton, NJ: Princeton University Press.

Gigerenzer, G. (2002). *Calculated risks: How to know when numbers deceive you.* New York: Simon & Schuster.

Gigerenzer, G., & Brighton, H. (2009). *Homo heuristics*: Why biased minds make better inferences. *Topics in Cognitive Science*, *1*, 107–143.

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2008). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, *8*, 53–96.

Hardisty, D. J., Johnson, E. J., & Weber, E. U. (2010). A dirty word or a dirty world? Attribute framing, political affliation, and query-theory. *Psychological Science*, *21*, 86–92.

Hays, W. L. (1994). *Statistics* (fifth edition). Belmont, CA: Wadsworth.

Huff, D. (1954). *How to lie with statistics.* New York: Norton.

Krämer, W., & Gigerenzer, G. (2005). How to confuse with statistics or: The use and misuse of conditional probabilities. *Statistical Science*, *20*, 223–230.

Kelley, T. L. (1947). *Fundamentals of statistics.* Cambridge: Harvard University Press.

Koehler, J. J. (1993). Error and exaggeration in the presentation of DNA evidence at trial. *Jurimetrics Journal*, *34*, 21–39.

Lilienfeld, S. A., Wood, J. M., & Garb, H. N. (2000). The scientific status of projective techniques. *Pschological Science in the Public Interest*, *1*, 27–66.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* Minneapolis: University of Minnesota Press.

Radelet, M. L. (1981). Racial characteristics and the impositon of the death penalty. *American Sociological Review*, *46*, 918–927.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*, 358–367.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, *15*, 351–357.

Selvin, H. C. (1958). Durkheim's *Suicide* and problems of empirical research. *American Journal of Sociology*, *63*, 607–619.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B*, *13*, 238–241.

Stam, A. (1997). MP approaches to classification: Issues and trends. *Annals of Operations Research*, *74*, 1–36.

Steinley, D. (2003). Local optima in K-means clustering: What you don't know may hurt you. *Psycholgical Methods*, *8*, 294–304.

Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology*, *33*, 565–578.

Thorndike, E. L. (1939). On the fallacy of imputing correlations found for groups to the individuals or smaller groups composing them. *The American Journal of Psychology*, *52*, 122–124.

Tufte, E. R. (1983). *The visual display of quantitative information.* Cheshire, CT: Graphics Press.

Tufte, E. R. (1990). *Envisioning information.* Cheshire, CT: Graphics Press.

Tufte, E. R. (1996). *Visual explanations.* Cheshire, CT: Graphics Press.

Tufte, E. (2006). *The cognitive style of PowerPoint: Pitching out corrupts within* (second edition). Cheshire, CT: Graphics Press.

Tukey, J. W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*, 105–110.

Tversky, A., & Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131.

Tversky, A., & Kahnman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*, 453–458.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, *4*, 274–290.

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin*, *83*, 213–217.

Wainer, H. (1984). How to display data badly. *The American Statistician*, *38*, 137–147.

Wainer, H. (1997). *Visual revelations: Graphical tales of fate and deception from Napoleon Bonaparte to Ross Perot.* New York: Copernicus Books (reprinted in 2000, Hillsdale, NJ: Lawrence Erlbaum Associates).

Wainer, H. (2005). *Graphic discovery: A trout in the milk and other visual adventures.* Princeton, N.J.: Princeton University Press.

Wainer, H. (2009). *Picturing the uncertain world: How to understand, communicate and control uncertainty through graphical display.* Princeton, NJ: Princeton University Press.

Wilkinson, L., & Friendly, M. (2009). The history of the cluster heat map. *The American Statistician*, *63*, 179–184.

Yarnold, P. R., & Soltysik, R. C. (2004). *Optimal data analysis.* Washington, DC: American Psychological Association.

Yule, G. U. (1903). Notes on the theory of association of attributes of statistics. *Biometrika*, *2*, 121–134.

Yule, G. U., & Kendall, M. G. (1968). *An introduction to the theory of statistics* (Fourteenth Edition, Fifth Impression). New York: Hafner Publishing Company.

Figure 1: A Restriction of Range Issue Between UGPA and LSAT

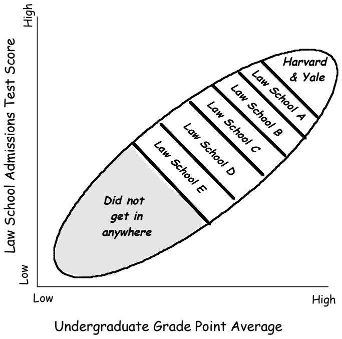


Figure 2: The Relation Between Rating by Graduate Faculty and the GRE Verbal Exam

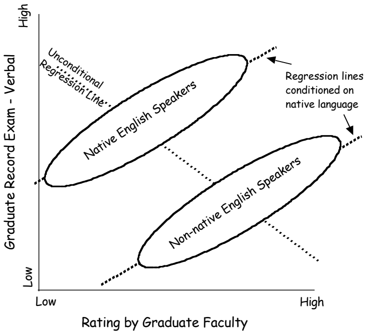

Graduate Record Exam - Verbal
High
Low
Unconditional Regression Line
Native English Speakers
Regression lines conditioned on native language
Non-native English Speakers
Low
High
Rating by Graduate Faculty

Figure 3: A Graph Showing That Smokers Die Sooner Than Nonsmokers

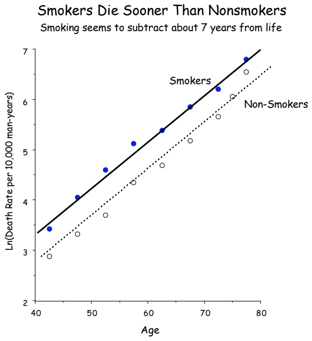
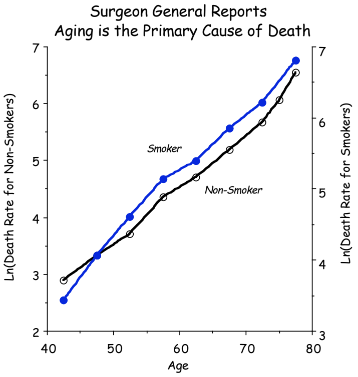


Figure 4: A Graph Showing That the Surgeon General Reports Aging is the Primary Cause of Death


Surgeon General Reports
Aging is the Primary Cause of Death
Smoker
Non-Smoker
Ln(Death Rate for Non-Smokers)
Ln(Death Rate for Smokers)
Age