The Basic Sampling Model

Psychology (Statistics) 484

Statistics, Ethics, and the Social and Behavioral Sciences

December 11, 2013

Beginning Quotations

As a single atom man is an enigma; as a whole he is a mathematical problem. As an individual he is a free agent; as a species the offspring of necessity.
— Winwood Reade (*The Martyrdom of Man*, 1872)

By a small sample we may judge the whole piece.
— Miguel de Cervantes (1547–1616)

Week 6: The Basic Sampling Model

— the decennial problem posed by the U.S. census; complete enumeration, as required by the Constitution, versus sampling, plus the political issues involved in the problem of “undercount”

— the unfortunate state of the statistical routines present in the widely used Excel program, and the inability of Microsoft to correct errors pointed out by the statistical community

Required Reading:
SGEP (175–225) —
Complete enumeration versus sampling in the Census

Multivariable Systems
Multivariable systems and unidimensional rankings
Graphical Presentation
Problems With Multiple Testing
Issues in Repeated-Measures Analyses
Matching and Blocking
Randomization and Permutation Tests
Pitfalls of Software Implementations
The unfortunate case of Excel
Sample Size Selection
Are large clinical trials in rapidly lethal diseases usually unethical?

Film:

The Plow that Broke the Plains (27 minutes)

The River (31 minutes)

George Stoney Commentary (21 minutes)

Some Statistical Distinctions

We begin by refreshing our memories about the distinctions between *population* and *sample*, *parameters* and *statistics*, and *population distributions* and *sampling distributions*.

Someone who successfully completes a sequence in statistics should know these distinctions very well.

Here, only a simple univariate framework is considered explicitly, but an obvious and straightforward generalization exists for the multivariate context.

Population

A *population* of interest is posited, and 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.

Sample

A *sample* is generated by taking observations on X, say, $X_1, \ldots, X_n$, considered independent and identically distributed as X; that is, 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.

Point Estimators

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 referred to as *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?

Confidence Interval For μ

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

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

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 fourfold increase in sample size; good, because sample size can be cut by four with only a halving of precision.

Central Limit Theorem

Even when the population distribution is not originally normally distributed, the central limit theorem (CLT) (that is, Galton's "Law of Frequency of Error") 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 is the basis for many claims of robustness; that is, when a procedure remains valid even if the assumptions under which it was derived may not be true, as long as some particular condition is satisfied;

here, the condition is that the sample size be reasonably large.

Law of Large Numbers

Besides the robustness of the confidence interval calculations for μ, the CLT also encompasses the law of large numbers (LLN).

As the sample size increases, the estimator, $\hat{\mu}$, gets closer to μ, and converges to μ at the limit as *n* goes to infinity.

This is seen most directly in the variance of the sampling distribution for $\hat{\mu}$, which becomes smaller as the sample size gets larger.

The Importance of Sample Size and Variability

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 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 a measure of performance for 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 we are able to isolate one very salient feature that most units in this top tier are small.

We proceed on this observation and advise the breaking 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. That smaller entities tend to be more variable than the larger entities seems to vitiate a recommendation of breaking up the larger units for performance improvement.

Sports is an area in which there is a great misunderstanding and lack of appreciation for the effects of randomness. A reasonable model for sports performance is one of "observed performance" being the sum of "intrinsic ability" (or true performance) and "error," leading to a natural variability in outcome either at the individual or the team level.

Somehow it appears necessary for sports writers, announcers, and other pundits to 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 .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.

People asked to generate random sequences of numbers tend to underestimate the amount of variation that should be present;

for example, there are not enough longer runs and a tendency to produce too many short alternations.

In a similar way, we do not see the naturalness in regression toward the mean, where extremes are followed by less extreme observations just because of fallibility in observed performance.

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 (the golfer must have been able to "play loose").

Or in baseball, at the start of a season an underperforming Derek Jeter might be under "pressure" or too much "media scrutiny," or subject to the difficulties of performing in a "New York market."

When individuals start off well but then appear to fade, it must be because people are trying to stop them ("gunning" for someone is a common expression).

Subsetting

Another area where one expects to see a lot of anomalous results is when the dataset is split into ever-finer categorizations that end up having 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 the clinical subjects.

The Ubiquity of Random Processes

Random processes are a fundamental part of nature and ubiquitous in our day-to-day lives.

Most people do not understand them, or worse, fall under an "illusion of control" and believe they have influence over how events progress.

Thus, there is an almost 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 cause is attributed 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.

When Randomness is Not Present

An appreciation for how random processes might operate can be helpful in navigating the uncertain world we live in.

When investments with Bernie Madoff give perfect 12% returns, year after year, with no exceptions and no variability, alarms should go off.

If we see a supposed scatterplot of two fallible variables with a least-squares line imposed but where the actual data points have been withdrawn, remember that the relationship is not perfect.

Or when we monitor error in quality assurance and control for various manufacturing or diagnostic processes (for example, application of radiation in medicine), and the tolerances become consistently beyond the region where we should generally expect the process to vary, a need to stop and recalibrate may be necessary.

It is generally important to recognize that data interpretation may be a long-term process, with a need to appreciate variation appearing around a trend line.

Thus, the immediacy of some major storms does not vitiate a longer-term perspective on global climate change.

Remember the old meteorological adage: climate is what you expect; weather is what you get.

Relatedly, it is important to monitor processes we have some personal responsibility for (such as our own lipid panels when we go for physicals), and to assess when unacceptable variation appears outside of our normative values.

The Stapel Affair

Besides having an appreciation for randomness in our day-to-day lives, there is also a flip side: if you don't see randomness when you probably should, something is amiss.

There are many such deterministic traps awaiting the gullible. When something seems just too good to be true, most likely it isn't true.

A recent ongoing case in point involves the Dutch social psychologist, Diederik Stapel, and the massive fraud he committed in the very best psychology journals in the field. A news item by G. Vogel in *Science* has the title, "Psychologist Accused of Fraud on 'Astonishing Scale'."

Basically, in dozens of published articles and doctoral dissertations he supervised, Stapel never failed to obtain data showing the clean results he expected to see at the outset.

As any practicing researcher in the behavioral sciences knows, this is just too good to be true.

We give a short quotation from the *Science* news item (October 31, 2011) commenting on the Tilberg University report on the Stapel affair (authored by a committee headed by the well-known Dutch psycholinguist, Willem Levelt):

Stapel was "absolute lord of the data" in his collaborations . . . many of Stapel's datasets have improbable effect sizes and other statistical irregularities, the report says. Among Stapel's colleagues, the description of data as too good to be true "was a heartfelt compliment to his skill and creativity."

Complete Enumeration versus Sampling in the Census

The basic sampling model implies that when the size of the population is effectively infinite, this does not affect the accuracy of our estimate, which is driven solely 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 United States Census is planned, where the issue of complete enumeration, as demanded by the Constitution, and the problems of undercount are revisited.

We begin with a short excerpt from a *New York Times* article by David Stout (April 2, 2009), "Obama's Census Choice Unsettles Republicans."

The quotation it contains from John Boehner in relation to the 2010 census is a good instance of the "resurfacing confusion"; also, the ethical implications of Boehner's statistical reasoning skills should be fairly clear.

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."

The Supreme Court ruling in *Department of Commerce v. United States House of Representatives* (1999) seems to have resolved the issue of sampling versus complete enumeration in a Solomon-like manner.

For purposes of House of Representatives apportionment, complete enumeration is required with all its problems of "undercount."

For other uses of the Census, however, "undercount" corrections that make the demographic information more accurate are permissable And these corrected estimates could be used in differential resource allocation to the states.

Multivariable Systems

In multivariate analysis, it is important to remember that there is systematic covariation possible among the variables, and this has a number of implications for how we proceed.

Automated analysis methods that search through collections of independent variables to locate the "best" regression equations (for example, by forward selection, backward elimination, or the hybrid of stepwise regression) are among the most misused statistical methods available in 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.

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 the structure among the variables as reflected by the regression, and not so much 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?

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 with a single equation.

The most important may be to remember that regression coefficients cannot be interpreted in isolation for their importance using their size, even when based on standardized variables (such as those calculated from z-scores).

That one coefficient is larger than another does not imply it is therefore more important.

The notion of importance can 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;

for example, 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" type of mechanism.

Multivariable Systems and Unidimensional Rankings

A common demand in academics is an adherence to a "truth in teaching" perspective that requires grading practices for a course to be spelled out in advance and in a syllabus.

The following wording is typical: "your grade for this course will depend on two midterm exams and a final; the final will account for 50% of your grade, with 25% for each of the two midterms."

The algorithmic process for assigning final grades might be to first standardize the three exam scores, weight the two midterms by .25 and the final by .5, and then sum. Chosen cut-scores would dictate the final letter grades assigned.

The percentage statements just can't be done so simply whenever there is a system of correlated (exam) scores.

If a weighting strategy for the scores were developed so that the variance in a set of midterm scores reflected 25 percent of the total score, it would most likely have the effect of overweighting pure random variability (the specific error) in the midterm scores.

US News and World Report Rankings

The manner in which the *US News* college rankings are constructed is akin to using a weighting scheme to assign grades.

Seven variables are now used for the college rankings with the *US News* percentage weights given in parentheses:

1. Undergraduate academic reputation (22.5%);
2. Graduation and freshman retention rates (20%);
3. Faculty resources (20%);
4. Student selectivity (15%);
5. Financial resources (10%);
6. Graduation rate performance (7.5%);
7. Alumni giving (5%).

In effect, the *US News* rankings seem driven by the one underlying factor of wealth (much as our exam scores may have been driven by an underlying factor of "subject matter knowledge").

The (former) President of Penn State, Graham Spanier, is quoted in the Gladwell article (mentioned in the readings) to this effect:

"What I find more than anything else is a measure of wealth: institutional wealth, how big is your endowment, what percentage of alumni are donating each year, what are your faculty salaries, how much are you spending per student."

Generally, it would be better to have an assessment of what a college adds to those individuals who attend, rather than just what the individuals themselves bring with them.

In economic terms, colleges should be evaluated by their production functions—what is the output for all combinations of input.

Commensurability

Several difficulties may be encountered in using any simple weighted average to obtain a unidimensional ranking:

(a) The variables aggregated to obtain a ranking need to be commensurable (that is, numerically comparable); otherwise, those variables with larger variances dominate the construction of any final unidimensional scale.

The most common mechanism for ensuring commensurability is through a z-score transformation so each variable has a mean of zero and a standard deviation (and variance) of one.

The *US News* approach to commensurability is not through z-scores but by a transform of each variable to a point scale from 1 to 100 (with Harvard always getting 100 points).

Fixed Weighting Schemes

b) Any multivariate analysis course places great emphasis on linear combinations of variables, and the formulas introduced at the outset are useful tools for evaluating ranking systems.

These typically are expressions for means, variances, and covariances (correlations) of arbitrary linear combinations of sets of variables, both for an assumed collection of random variables characterizing a population and for an observed data matrix obtained on these random variables.

Thus, there are general mechanisms available for studying the types of weighting systems represented by the *US News* aggregation system.

For example, the effects of changing weighting systems can be studied through how the constructed scales would correlate;

also, several other statistical tools can be used: scatterplots, measures of nonlinear and rank correlation, consistency of generated rankings, changes in the variances of the generated scores, and sensitivity of rankings to the changes in aggregation weights.

As done now, the *US News* ranking is a "take it or leave it" proposition.

Not only is the particular choice of variables a *fait accompli*, the numerical aggregation system is a fixed entity as well.

It is statistically (and ethically) questionable for such a closed system to have the putatively large influence it does on the United States system of higher education.

Lack of a Criterion Variable

(c) The one glaring omission in the types of ranking systems used by *US News* is the lack of a defensible criterion variable.

All we have are input measures; there are no viable output (or criterion) measures.

If a criterion was available, the use of an arbitrary aggregation of input variables could be partially mitigated by the adoption of a defensible weighting mechanism through an optimization process.

For example, multiple regression provides that linear combination of the predictor variables correlating the highest with the criterion measure.

Principal Components to Define a Weighting Scheme

(d) The absence of a viable criterion measure does not preclude considering other optimization mechanisms for aggregating sets of variables.

The principal components of a correlation matrix, for example, provide a collection of possible weighting schemes having several nice properties.

If there are *p* original variables available, then there are *p* principal components, where each component provides a set of weights;

that is, each principal component defines a linear combination of the original variables; the components as a group "repackage" the information present in the original variables.

The First Principal Component

(e) There are other advantages in considering the principal components for the given set of variables.

First, the scores on the first component immediately define a ranking based on numerical values obtained by a process of maximizing the variability of the aggregate;

this might provide a defensible weighting strategy satisfying the statistical *literati*.

How the variables were chosen in the first place may still be questionable, but once this is decided, the method of weighting is constructed through the transparent optimizing of a variance criterion.

A principal component analysis itself can offer another justification for a unidimensional ranking, or conversely, be used to argue that more dimensions are needed.

If the proportion of variance explained by the first component is high (for example, 80% or more), the remaining components may not have much more to offer.

One could argue that most of what we know about our variables can be explained by a single unidimensional scale generated by the first principal component.

On the other hand, if several components are needed to attain a sufficient total "variance-accounted-for," the wisdom of relying on a single scale is called into question.

Graphical Presentation

Graphical and other visual methods of data analysis are central to an ability to tell what the 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.

For example, you should 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.

But be careful; don’t overuse unnecessary graphics:

Tufte has lamented the poor use of graphics that relies on “chart junk” for questionable visual effect, or gratuitous color or three-dimensions in bar graphs that do not represent anything real at all.

Heat Maps

In providing data in the form of matrices, such as subject by variable, we might consider the use of "heat maps," where numerical values, assumed commensurable over variables, are mapped into color spectra reflecting magnitude.

The further imposing of 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).

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 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 many beginning statistics classes: you cannot legitimately test a hypothesis on the same data that first suggested it.

An article from *ScienceNews* by Laura Sanders ("Trawling the Brain", October 19, 2009) provides a cautionary lesson for anyone involved with the interpretation of fMRI research (we read this in Week 2).

A dead salmon's brain can display much of the same beautiful red-hot areas of activity in response to emotional scenes flashed to the (dead) salmon that would be expected for (alive) human subjects.

Bonferroni Corrections

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 event A_k is defined as the incorrect rejection of H_k (that is, rejection when it is true), the Bonferroni inequality gives

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

Noting that the event $(A_1 \text{ or } \cdots \text{ or } A_K)$ can be verbally restated as one of "rejecting incorrectly *one or more* of the hypotheses," the experiment-wise (or overall) error rate is bounded by the sum of the K α values set for each hypothesis.

Typically, we let $\alpha_1 = \cdots = \alpha_K = \alpha$, and the bound is then $K\alpha$.

Thus, the usual rule for controlling the overall error rate through the Bonferroni correction sets the individual αs at some small value such as $.05/K$;

the overall error rate is then guaranteed to be no larger than .05.

Odd Behaviors in Multiple Testing

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

(b) Scheffé'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 legitimately look at your data and then 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 (such as those defined by orthogonal comparisons), the problem of inflating the overall error rate remains;

similarly, in performing a multifactor analysis of variance (ANOVA) or testing multiple regression coefficients, all of the tests carried out should have some type of control imposed on the overall error rate;

(f) It makes little sense to perform a multivariate analysis of variance before you go on to evaluate each of the component variables.

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 accept 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 uninterpretable overall MANOVA test completely.

Issues in Repeated-Measures Analyses

The analysis of repeated measures generally needs special treatment in that the usual models are not very trustworthy and can lead to erroneous conclusions.

The starting place is commonly a Mixed Model III ANOVA with a fixed treatment factor and a subject factor considered random.

To model repeated observations justifying the usual F-ratio test statistic of Mean-Square Treatments to Mean-Square Interaction, an assumption is made that the observations within a subject are correlated.

The use of the usual F-ratio, however, requires that all these correlations be the same irrespective of which pair of treatments is considered (an assumption of "compound symmetry").

The compound symmetry assumption may be reasonable when the treatment times are randomly assigned, but if the responses are obtained sequentially, then possibly not.

Treatments further apart in time are typically less correlated because of fatigue, boredom, familiarity, and so on.

Unfortunately, there is strong evidence of nonrobustness in the use of the equicorrelation assumption when it is not true, with too many false rejections of the null hypothesis of no treatment differences.

Degree-of-Freedom Corrections

A way around this nonrobustness is implemented in many software packages.

If we knew the structure of all the variances and covariances among the treatments, we could obtain a parameter, say, θ, that would give an appropriate correction for the degrees of freedom of the F-distribution against which to compare the calculated F-ratio;

that is, we would use $F_{\theta(B-1),\theta(A-1)(B-1)}$, where there are B treatments and A subjects.

Although θ is unknown, there are two possible strategies to follow:

estimate θ with Huynh–Feldt procedures (as is done, for example, in SYSTAT);

or use the greatest reduction possible with the discounting bound of $1/(B-1)$ (that is, the Geisser–Greenhouse result of $1/(B-1) \leq \theta$, also as done when an analysis is carried out using SYSTAT).

So, if a rejection occurs with the Geisser–Greenhouse method, it would also occur for the Huynh–Feldt estimation, or if you knew and used the actual value of θ.

Profile Analysis

Another approach to repeated measures, called "profile analysis," uses Hotelling's T^2 statistic and/or MANOVA on difference scores.

In fact, the only good use of a usually noninformative MANOVA may be in a repeated-measures analysis.

Three types of questions are commonly asked in a profile analysis:

Are the profiles parallel to each other?

Are the profiles coincident?

And, are the profiles horizontal?

When done well, a profile analysis can give an informative interpretation of repeated-measures information with an associated graphical presentation.

Two Issues in Repeated Measures Analysis

Two possible issues with repeated measures should be noted.

First, it is assumed that the responses from our subjects are commensurable over the variables measured.

If not, an artificial transformation could be considered such as to z-scores, but by so doing, the test for horizontal profiles is not meaningful because the associated test statistic is identically zero.

Second, the number of subjects versus the number of measurement times may prevent carrying out a Hotelling T^2 comparison in a profile analysis (but not, say, a correction based on a Huynh–Feldt estimated θ).

Generally, if there are more time points than subjects, one of the degrees of freedom in the F-distribution used for the T^2 comparison is negative, and thus, the test is meaningless.

The Variance of Mean Differences

Anyone analyzing repeated measures needs to remember that the variance of the difference between two means, say $\bar{X}$ and $\bar{Y}$, is not the same when $\bar{X}$ and $\bar{Y}$ are based on independent samples.

In particular, suppose $\bar{X}$ is obtained for the observations $X_1, \ldots, X_N$, and $\bar{Y}$ for $Y_1, \ldots, Y_N$.

When the samples are independent, the variance of the difference $\bar{X} - \bar{Y}$, $S^2_{\bar{X}-\bar{Y}}$, can be estimated as $S^2_{\bar{X}} + S^2_{\bar{Y}}$, where $S^2_{\bar{X}} \equiv S^2_X/N$, $S^2_{\bar{Y}} \equiv S^2_Y/N$, and S^2_X and S^2_Y are the sample variances for $X_1, \ldots, X_N$ and $Y_1, \ldots, Y_N$, respectively.

In the repeated-measures context, the variance of $\bar{X} - \bar{Y}$ can be estimated as $S^2_{\bar{X}} + S^2_{\bar{Y}} - 2(S_{XY}/N)$, where S_{XY} is the sample covariance between the observations $X_1, \ldots, X_N$ and $Y_1, \ldots, Y_N$.

Thus, we have a difference in the term, $-2(S_{XY}/N)$, which in most instances will be a negative correction when the X and Y observations are positively related.

In other words, the variance of the difference, $\bar{X} - \bar{Y}$, will generally be less in the context of repeated measures compared to independent samples.

Neuroimaging and Repeated Measures

In some areas of neuroimaging, the repeated-measures nature of the data is just ignored;

we have pixels (or voxels) that are spatially arranged (and subject to various types of spatial autocorrelation) that move through time (and subject again to various types of temporal autocorrelation).

In these frameworks where the repeated measures are both spatial and temporal, it is not sufficient to just use the various multivariate general linear model extensions that assume all error terms are independent and identically distributed (as suggested by some best-selling fMRI handbooks; for example, see Huettel, Song, & McCarthy, 2004, pp. 336–348).

Time Series Analyses

A related repeated-measures topic is in the time-series domain, where some variable is observed temporally.

Substantial modeling efforts have involved the Box–Jenkins approach of using ARIMA (autoregressive-integrated-moving-average) models.

A more subtle question in this context is to assess the effects of an intervention on the progress of such a time series.

In the case of single-subject designs, where a subject serves as his or her own control, the issue of evaluating interventions is central (see Kazdin, 1982).

A particularly elegant approach to this problem has been developed by Edgington (see Edgington & Onghena, 2007, Chapter 11: "N-of-1 Designs"), where intervention times are chosen randomly.

The same logic of analysis is possible as in a Fisherian approach to analyzing an experiment where the various units have been assigned at random to the conditions.

Matching and Blocking

One of the main decision points in constructing an experimental design is whether to block or match subjects, and then within blocks randomly assign subjects to treatments.

Alternatively, subjects could be randomly assigned to conditions without blocking.

As discussed earlier, it is best to control for initial differences beforehand.

Intact groups can't be equated legitimately after the fact through methods such as analysis of covariance or post-hoc matching.

But the question here is whether blocking makes sense over the use of a completely randomized design.

This choice can be phrased more formally by comparing the test statistics appropriate for a two-independent or a two-dependent samples *t*-test.

The principle derived from this specific comparison generalizes to more complicated designs.

Suppose we have two equal-sized samples of size N, $X_1, \ldots, X_N$ and $Y_1, \ldots, Y_N$.

When the two samples are independent, the two-independent samples *t*-statistic has the form

$$\frac{\bar{X} - \bar{Y}}{\sqrt{(S_X^2 + S_Y^2)/(N-1)}} ,$$

where S_X^2 and S_Y^2 are the sample variances;

this statistic is compared to a *t*-distribution with $2(N-1)$ degrees of freedom.

When the samples are dependent and X_i and Y_i are repeat observations on the ith subject, the paired t-statistic has the form

$$\frac{\bar{X} - \bar{Y}}{\sqrt{S_D^2/(N-1)}} \ ,$$

where S_D^2 is the sample variance of the difference scores.

Here, the paired t-statistic is compared to a t-distribution with $N - 1$ degrees of freedom.

We note the relation $S_D^2 = S_X^2 + S_Y^2 - 2S_{XY}$, where S_{XY} is the sample covariance between X and Y.

In the initial design of an experiment, there may be a choice: match subjects and assign members within a pair to the treatments, or just assign all subjects randomly to the two treatments without matching.

Generally, if the matching variable is not very important in that the sample covariance is not that large (and positive), to compensate for the halving of the degrees of freedom in going from $2(N-1)$ to $(N-1)$, it only hurts to match.

To compensate for the loss of degrees of freedom and make the paired t-statistic sufficiently larger than the independent sample t-statistic, the variance of the differences, S_D^2, in the denominator of the paired t-statistic must be sufficiently smaller compared to $S_X^2 + S_Y^2$ in the denominator of the independent samples t-statistic.

Unfortunately, unless one has some estimate of the covariance of X and Y, the choice of design must be based on a guess.

A dictum, however, may still be gleaned: don't block or match on variables that have no possible (positive and relatively strong) relation to the type of responses being measured.

Randomization and Permutation Tests

An important benefit from designing an experiment with random assignment of subjects to conditions, possibly with blocking in various ways, is that the method of analysis through randomization tests is automatically provided.

As might be expected, the original philosophy behind this approach is due to R. A. Fisher (1971), but it also has been developed and generalized extensively by others (see Edgington & Onghena, 2007).

In Fisher's time, and although randomization methods may have been the preferred strategy, approximations were developed based on the usual normal theory assumptions to serve as computationally feasible alternatives.

But with this view, our standard methods are just approximations to what the preferred analyses should be.

A short quotation from Fisher's *The Design of Experiments* (1971) makes this point well:

In these discussions it seems to have escaped recognition that the physical act of randomisation, which, as has been shown, is necessary for the validity of any test of significance, affords the means, in respect of any particular body of data, of examining the wider hypothesis in which no normality of distribution is implied. The arithmetical procedure of such an examination is tedious, and we shall only give the results of its application . . . to show the possibility of an independent check on the more expeditious methods in common use.

A randomization (or permutation) test uses the given data to generate an exact null distribution for a chosen test statistic.

The observed test statistic for the way the data actually arose is compared to this null distribution to obtain a *p*-value, defined as the probability (if the null distribution were true) of an observed test statistic being as or more extreme than what it actually was.

Three situations lead to the most common randomization tests: K-dependent samples, K-independent samples, and correlation.

When ranks are used instead of the original data, all of the common nonparametric tests arise.

In practice, null randomization distributions are obtained either by complete enumeration, sampling (a Monte Carlo strategy), or through various kinds of large sample approximations.

Sample Reuse Methods Generally

The idea of repeatedly using the sample itself to evaluate a hypothesis or to generate an estimate of the precision of a statistic, can be placed within the broader category of resampling statistics or sample reuse.

Such methods include the bootstrap, jackknife, randomization and permutation tests, and exact tests (for example, Fisher's exact test for 2×2 contingency tables).

Given the incorporation of these techniques into conveniently available software, such as R, there are now many options for gauging the stability of the results of one's data analysis.

Pitfalls of Software Implementation

Most of our statistical analyses are now done through the use of packages such as SYSTAT, SPSS, or SAS.

Because these systems are blind to the data being analyzed and the questions asked, it is up to the user to know some of the pitfalls to avoid.

For example, the fact that an analysis of covariance is easy to do does not mean that is should be done or that it is possible to legitimately equate intact groups statistically.

Even though output may be provided, this 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, and nonrobust tests for variances.

More General Software Traps

(a) In the construction of items or variables, the numbers assigned may at times be open to arbitrary coding. 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 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 or reverse code” a subset of the items in the hope of obtaining an approximate “positive manifold” for the correlation matrix, characterized by few if any negative correlations that can’t be attributed to sampling error.

(b) Certain methods of analysis (for example, most forms of multidimensional scaling, K-means and mixture model cluster analysis, and some strategies involving optimal scaling) are prone to local optima where results are presented but not the best possible ones according to the goodness-of-fit measure being optimized.

The strategies used in the optimization cannot guarantee global optimality because of the structure of the functions being optimized.

One standard method of local optimality exploration is to start (repeatedly and randomly) a specific analysis method, observe how severe the local optima problem is for a given dataset, and then choose the best analysis found for reporting a final result.

Unfortunately, none of the current packages offer these random-start options for all the methods that may be prone to local optima.

c) Methods of analysis involving optimization often use iterative algorithms that converge to a solution even though it may be only a local optimum.

Generally, various convergence criteria are set by default in the software; for example, maximum number of iterations allowed, minimal change in the stepsize used by the algorithm, or minimal change in the value of the loss criterion being optimized.

In any event, it is up to the user to know when a premature termination of an algorithm has occurred, and to be able to change the default convergence criteria values to insure that a "real" solution has been achieved.

Again, attempts to explain prematurely terminated results show that you really don't know what you are doing.

In an era when computer time was expensive, one had to be very stingy about setting the defaults to ensure that only a limited amount of computational effort could be expended.

Now, running our machines for a few (more) hours has no real cost attached. So, if you see a message such as "maximum number of iterations exceeded (or reached)," don't ignore it.

Instead, change the default limits until the message disappears for the solution achieved.

Open Versus Closed Statistical Systems

Closed statistical systems allow analyses to be done with little or no understanding of what the "point-and-clicks" are really giving.

At times, this may be more of an impediment to clear reasoning than assistance. The user does not need to know much before being swamped with copious amounts of output, with little help on how to wade through the results or to engage in further exploration (for example, in investigating local minima or carrying out alternative analyses).

One of the main reasons for now employing one of the newer statistical environments (such as R or MATLAB) is that they do not rely on pull-down menus.

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.

The R environment has become the *lingua franca* for framing cutting-edge statistical development and analysis, and is becoming a 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.

The Unfortunate Case of Excel

The documented inadequacies of Excel are legion, as is Microsoft's inability to correct flaws when they are pointed out.

A good place to learn about Excel's statistical failings is a special section on Microsoft Excel 2007 in the journal *Computational Statistics & Data Analysis* (2008, *52*, 4568-4606), beginning with a scathing editorial by McCullough (2008b, pp. 4568–4569).

Several of the points raised in this collection of articles will be noted in the readings, usually with direct quotations from the articles themselves.

One overarching conclusion would be that relying solely on Excel's statistical routines needlessly puts people at risk.

Sample Size Selection

At times the issue of picking sample size is moot, such as when some number of subject hours are allocated from the subject pool in a beginning psychology course and no more.

In other instances where there is the possibility of recruiting paid subjects, it may be incumbent upon the experimenter to have a defensible rationale for how many subjects are necessary for the study to have a reasonable likelihood of success.

A formal process for determining sample size becomes particularly crucial when obtaining data from even one subject is very expensive (for example, in neuroimaging), or when subjects are paid from a grant and allocation of funds for that purpose must be justified, or when competing clinical trials require differential allocation from a common pool of patients.

The choice of sample size almost always involves a trade-off between the size of the effect the experimenter wishes to detect with high probability, and the maximum number of subjects available.

It may be that the limited number of possible subjects makes even carrying out the experiment unnecessary because the size of effect likely to be present will not be detectable with the limited number of subjects available.

It is also important to remember that measure fallibility reduces power and the ability to detect effects of interest.

Other things being equal, measures having low reliability require more subjects to detect an effect of a given size.

Are Large Clinical Trials in Rapidly Lethal Diseases Usually Unethical?

The choice of sample size appears to be fairly mechanical once an effect size is specified that one wishes to detect at a given probability level.

However, when delving further into the ethics behind sample size selection in medical trials, the issues become murkier.

As a poignant illustration, we give a “personal paper” written by David Horrobin, having the title given to this subsection, that appeared in the *Lancet* the same year he died (2003).

As a medical scientist, Horrobin was a controversial figure for his advocacy of evening primrose oil for a variety of medical purposes.

As one indication of this controversy, a negative Horrobin obituary that ran in the *British Medical Journal* in 2003 generated the most responses (both pro and con) for any obituary in the history of the journal.

For our purposes, however, there is no need to pursue this part of Horrobin's career to appreciate the emotionally moving perspective he brings to sample size selection in clinical trials for lethal diseases, his included.