Ten categories of statistical errors: a guide for research in endocrinology and metabolism

Tyson H. Holmes
Division of Biostatistics, Department of Health Research and Policy, Stanford University School of Medicine, Stanford, California 94305-5405

Holmes, Tyson H. Ten categories of statistical errors: a guide for research in endocrinology and metabolism. Am J Physiol Endocrinol Metab 286: E495–E501, 2004;10.1152/ajpendo.00484.2003.—A simple framework is introduced that defines ten categories of statistical errors on the basis of type of error, bias or imprecision, and source: sampling, measurement, estimation, hypothesis testing, and reporting. Each of these ten categories is illustrated with examples pertinent to research and publication in the disciplines of endocrinology and metabolism. Some suggested remedies are discussed, where appropriate. A review of recent issues of American Journal of Physiology: Endocrinology and Metabolism finds that very small sample sizes may be the most prevalent cause of statistical error in this literature.

Statistics; bias; precision; sampling; hypothesis testing

STATISTICS OFFERS a widening range of powerful tools to help medical researchers attain full and accurate understanding of biological structure within their data. As research methodologies continue to advance in sophistication, these data are becoming increasingly rich and complex, thereby requiring increasingly thoughtful analysis. In response to these developments, funding sources, regulatory agencies, and journal editors are tightening their scrutiny of studies’ statistical designs, analyses, and reporting. Compared with the past, statistical errors of any magnitude carry greater weight today in the competition among scientists for research grants and in the publication of research findings.

This essay presents a framework to help researchers and reviewers identify ten categories of statistical errors. The framework has two axes. The first axis recognizes two canonical types of statistical error: bias and imprecision. The second axis distinguishes five fundamental sources of statistical error: sampling, measurement, estimation, hypothesis testing, and reporting.

Bias is error of consistent tendency in direction. For example, an assay that consistently tends to underestimate concentrations of a metabolite is a biased assay. In contrast, imprecision is nondirectional noise. An imprecise assay may give true readings on average, but those readings vary in value among repeats.

These two types of error arise from at least five sources. Sampling error is error introduced by the process of selecting units (e.g., human subjects, mice, and so forth) from a population of interest. Once a unit has been drawn into the sample, error can arise in measurements made on that unit. The researcher often wishes to use these measured data to estimate parameters (e.g., mean, median) or test hypotheses pertaining to the population from which units were drawn. In doing so, statistical methods should be chosen to minimize statistical error in these estimates and tests. Finally, with the study completed, results need to be reported in a manner that ensures that readers are as fully and accurately informed as possible. These five sources are conceptually distinguishable but interrelated in practice. As will be illustrated, discussion of one source often has implications for others.

The objective of this essay is to illustrate each of these ten categories of statistical error and also provide some constructive suggestions as to how these errors can be minimized or corrected. This essay is not intended to provide detailed training in statistical methods; rather, its purpose is to alert researchers to potential statistical pitfalls. Details on statistical methods that are pertinent to a researcher’s particular application can be found in texts (some of which are cited in this essay) or from consultation with a biostatistician. In this essay, formal mathematical presentations have been kept to a minimum. Illustrative examples are drawn from review of a variety of original-research articles published in this journal, American Journal of Physiology: Endocrinology and Metabolism, and in Endocrinology. Specific articles in the literature are not used as examples, because the purpose of this essay is not to critique particular authors but instead to identify where improvement may be possible in this literature as a whole.

CATEGORY I: SAMPLING BIAS

Sampling begins by identifying a specific population of interest for study. Key to minimizing sampling bias is specificity, because declaring interest in a broad study population (e.g., humans with type 2 diabetes) essentially guarantees sampling bias. To illustrate, if one truly wished to characterize average peripheral-blood concentrations of some metabolite in humans with type 2 diabetes, then all such humans would need to be available for sampling and each assigned a known probability of being selected for study. Not only is this impractical, because such a list could never be compiled, but it is also unethical, because human subjects must be allowed to volunteer for potential entry into research rather than being
randomly selected. In fact, because voluntary participation is the foundation of ethical research in human subjects, research on human subjects invariably contains self-selection bias. Although self-selection bias may not be eliminated, it can be minimized. During study planning, the researcher simply declares interest in subjects of that type who are likely to volunteer and meet entry criteria; and, during study implementation, the researcher keeps dropout (and noncompliance) to a minimum. In contrast to research in human subjects, with animal models samples are often drawn from bred populations, so that care must be taken, and apparently often is in the endocrinology and metabolism (E and M) literature, to designate during study planning the specific breeding line from which a sampling will be obtained.

Sampling need not be from one population. A researcher may wish to study several populations simultaneously. This requires planning to avoid errors in parameter estimation and hypothesis testing. In particular, if more than one population is of interest, sampling should be stratified by design (13). Here we consider two potentially problematic types of mixed-population sampling that are found in the E and M literature: sampling from extremes and changes in protocol.

The E and M literature contains examples of deliberate sampling from the extremes. For example, a study in mice may compare the 10% lightest and the 10% heaviest in a batch. The limitation of this form of stratified sampling is that results obtained and conclusions drawn apply only to those extremes. Any interpolation may be biased if the extremes behave differently from intermediate levels, which can arise if, say, lightest mice possess medical complications that are uncharacteristic over the remainder of the weight range. A better basic design would be to sample at equal intervals over the entire weight range. Atop this, one may concentrate sampling around any suspected within-range change points (e.g., a weight at which the slope on weight changes sharply or, say, average response shifts abruptly up or down).

Mixed-population samples may also result from in-course changes in study protocol. For example, an injection schedule may be changed between subjects enrolled early during a study and those enrolled later. Post hoc statistical analysis (e.g., regression analysis using covariates) may be attempted to disentangle the effects of the protocol change from any treatment differences, time trends, and the like that were of primary interest in the original design of the study. However, no guarantee can be made that statistical “correction” for in-course changes in protocol will be fully effective, especially when in-course protocol changes are strongly confounded with any effects of interest. The best strategy is to avoid in-course protocol changes altogether. Any adjustments in protocol should be made before the conduct of the full study, perhaps on the basis of results of preliminary studies.

Sampling bias can also arise in experimental studies in the form of an “inadequate control.” This is an example of bias due to sampling from the wrong population. An ideal control possesses all characteristics of the treatment condition that impact response except for the putative mechanism under study. Such ideals can be very difficult to obtain in practice (e.g., with assays in the presence of cross-reactivity or in attempts to find “matching” controls in nonrandomized studies). One important safeguard against an inadequate control, which is not always followed in the E and M literature, is to observe treatment and control samples contemporaneously, not in sequence. For example, in a two-arm randomized trial, both arms should be running between the same start and end dates, rather than running one arm first and the other arm later. Contemporaneous implementation can control for a number of time-varying factors that may impact the outcome of interest, such as turnover in lab technicians and drift in instrument calibration.

Inadequate randomizations are a form of sampling bias, because the goal of randomization is to generate two (or more) samples at baseline that are, on average, as homogenous as possible on all factors that may influence outcome. Simple randomizations alone do not guarantee this homogenization, especially when sample sizes are small, as is common in the E and M literature (see Category II, for example). The best corrective action is avoidance through use of more sophisticated randomization methods, such as stratified or adaptive designs (14). These include “permuted block” designs, in which the order of treatment assignments is randomly permuted within small to moderately sized blocks of subjects of shared baseline characteristics. A much less desirable correction, because it is remedial and not preventive, is to “statistically adjust” for baseline differences during parameter estimation by using, for example, appropriate multiple-regression methods. The formulation of such regression models should be specified in advance of any data review to avoid “data snooping” (Category VII). Despite their ease of use, these mathematical adjustments to reduce estimation bias can never fully substitute for starting all arms of an experiment at approximately equivalent compositions.

Sampling error of consistent tendency in direction not only can lead to biased estimates of “location” parameters (e.g., means) but also can result in bias in parameters of any type, including parameters of “dispersion” such as the variance. Sampling bias in dispersion can mislead understanding of how heterogeneous a population truly is. Sampling bias in the variance from a pilot study can yield sample-size estimates that are too large or small for the confirmatory study. Overestimates of variances result in unnecessary losses in power, whereas underestimates increase chances that one will falsely conclude that a treatment is actually effective, that a mean response truly increases over time, and the like. A fundamental tenet to minimizing bias in variance estimates is to sample across those particular units to which conclusions are to apply. For instance, suppose one wished to test for the response of patients to a possible medical therapy that acts on pancreatic cells. One could conduct this test on multiple pancreatic cells drawn from a single donor, but the resulting variance estimate would be an estimate of within-subject variance (wrong parameter) from a single subject (potentially idiosyncratic), so that one could not use these results to draw conclusions of general use to therapists. Instead, one would need to draw a sampling of cells from each of several donors, calculate mean cell response for each donor, and then estimate the among-subject variance from these subjects’ means.

CATEGORY II: SAMPLING IMPRECISION

Compared with other medical and scientific disciplines, many E and M research studies tend to use very small sample sizes, with sample sizes <10 being commonplace. Doubtless
practitioners use small samples because observations are difficult and/or expensive to obtain. Some may also argue that physiological characteristics are less variable than, say, clinical or epidemiological characteristics and that, therefore, smaller sample sizes are permissible in most laboratory E and M studies. Although these arguments carry some weight, they cannot justify the extremely small sample sizes that are so common in this literature. Just how extremely small these samples are can be illustrated as follows.

Suppose we randomly draw a sample of size \( n \) from a population and calculate the sample arithmetic mean. Then we repeat this process many times. Our estimate of the sample mean will vary to some degree among our different samples. Traditionally, one way to characterize this sampling imprecision in the mean is to estimate the standard deviation (i.e., standard error) of the samples’ means. Fortunately, one can estimate the standard error from a single sample as \( s/\sqrt{n} \), where \( s \) is the sample standard deviation. Notice how the standard error depends inversely on sample size \( n \), which makes sense intuitively, as one would expect precision of the sample mean to increase as sample size increases. However, this relationship is not linear, in that the standard error is proportional to the inverse of the square root of sample size. As illustrated in Fig. 1, sample sizes <10 tend to have large relative imprecision (1/\( \sqrt{n} \)) and thus yield unquestionably imprecise estimates of the mean.

The degree of sampling imprecision in the E and M literature could see major reductions if samples of size 15–30 were more widely employed or required. For example, Fig. 1 shows that a sample size of 5 has relative imprecision of ~0.45, whereas a sample size of 20 has relative imprecision of ~0.22, which is a reduction in imprecision of >50%.

Imprecision can also result if sampling is from more than one source population but not stratified by design (13). Mixed-population sampling can arise during recruitment in a hemodynamic study in which, for instance, some subjects present with arterial hypertension and others do not. Because hypertension may affect hemodynamic response, an efficient design would deliberately stratify sampling on hypertension status. If stratified sampling were not designed into the study, some type of poststratification could be performed after data collection; however, this remedial approach risks large disparities in sample size among strata, so that some strata may have small sample sizes, which reduces precision and statistical power. Instead, in advance of implementation of a study, the researcher could identify any factors that may have a significant impact on outcome, form a parsimonious quantity of strata from these factors, and then sample from each stratum. If the researcher wishes to test statistically for differences among these strata, then sample size could be controlled by study administrators to be nearly equal across all strata.

**CATEGORY III: MEASUREMENT BIAS**

In contrast to sampling errors, E and M research is clearly devoted to minimizing errors of measurement, as evidenced by the large proportion of published METHODS sections that address issues of measurement (e.g., specimen handling and storage, preparation of solutions and cells, measurement of fluxes, and the like). Seemingly, enough detail is typically given to permit readers to reproduce all or most all of the reported measurement methods.

Nevertheless, measurement bias can creep into E and M research data in subtle ways, often through some alteration that takes place over the course of the study. Changes in reagent manufacturers and turnover in device operators are just a couple of possible examples. All such alterations should be identified and tested to see if they introduce bias into measured outcomes. For instance, a switch to a new source of reagent should never be sequential; rather, if a switch is identified in advance, a series of test assays should be run simultaneously on both reagents, and results compared. Key to testing an alteration is to recognize that one’s intent is to prove that the two conditions (e.g., reagents, operators) yield equivalent, not different, results. As such, instead of standard two-sample testing (e.g., with Student’s \( t \)-test), bioequivalency testing should be performed in consultation with a statistician. In these discussions, the researcher should come prepared to provide the statistician with quantitative upper and lower bounds that define the range of biological equivalence (e.g., \( \pm 1\% \) difference between means of old and new reagents).

**CATEGORY IV: MEASUREMENT IMPRECISION**

Measurement imprecision is straightforward to characterize through use of “technical repeats.” Technical repeats result from taking multiples of the same measurement (e.g., an assay) on the same specimen at the same time. Technical error, sometimes termed “intra-assay variation,” is not commonly reported in the E and M literature. Whether technical repeats are rarely performed or this simply represents a failure to report technical error is unclear. The coefficient of variation (CV) provides a unitless measure of technical error. Ideally, one hopes to see technical-error CVs of <5–10% at most. Typically, the CV is estimated by the ratio of the sample standard deviation to the sample mean. This estimator is biased, and in small samples this bias is large enough that a corrected formulation should be employed (12).

![Fig. 1. Relative precision 1/\( \sqrt{n} \) as a function of sample size \( n \). Dashed vertical line denotes a sample size of \( n = 10 \).](http://ajpendo.org)
E498 STATISTICAL ERRORS

CATEGORY V: ESTIMATION BIAS

Estimation bias is error of consistent tendency in direction for a sample-based estimate of the value of a parameter, where a parameter is a characteristic of the population of interest (e.g., mean or variance). Estimation bias is distinct from sampling bias and measurement bias in that we are not concerned with bias arising from the collection of data per se. Rather our focus is limited to bias that arises in the computational process by which we formulate an estimate of a parameter from data that have already been collected. Formally, any such formulation is termed an “estimator” of a parameter.

This is not to say that parameter estimation and sampling bias are wholly separable. A clear example of their overlap is with missing data. Missing data do occur in the E and M literature, perhaps more often than can be detected because of spotty reporting of sample sizes in figures and tables. Possible signs of missing data are when an article reports that all “available” data were analyzed or when sample sizes differ between methods and results. In addition, the cause of missing data is often not reported. This cause should not only be given but detailed, as details are all-important to understanding the scale and character of any effect of missing data on estimation bias. Roughly speaking, missing data consist of two types: those which are missing because of the value they would take if observed (“informative”), and those that are missing for reasons unrelated to the value that they would take if observed (“noninformative”) (7). A specimen that is accidentally dropped is an example of noninformative missing data. In contrast, informative missing data arise when specimens of a particular type of response (e.g., high readings) are more likely to be lost. Informative missing data have the potential for making unbiased estimation difficult to impossible, particularly when the precise reason for why data are missing is unknown or unclear. Unbiased estimation in the presence of missing data is most complex when data consist of repeated measurements on each subject and some subjects are missing some of their repeated measurements. Of course, the quantity of missing data should be minimized whenever possible, even if this requires preparation of more reagent, buffer, collection tubes, and the like. A complete analysis data set includes information on the precise cause for each missing datum. For example, separate codes could be used for measurements not obtained due to 1) specimen lost, 2) reading above instrument range, 3) reading below instrument range, and so forth. Suppose a research project generated data of which one-fourth measured above instrument range, one-fourth measured below instrument range, and the remainder were nonmissing. These data could be salvaged by transformation to an ordinal scale (below range, in-range, above range) and analyzed, albeit with a potential loss of power compared with an analysis on the original interval or ratio scale with all data nonmissing.

Another situation in which estimation bias can arise is when estimation fails to recognize the presence of an “incomplete design.” Suppose the effect of compound Y on intracellular pH has been previously studied, and primary interest is determining how compound X affects intracellular pH. An experiment is designed in which cell specimens are randomized to being cultured in the absence of X and Y, with Y alone, and with X and Y together (X + Y). This design does not permit unbiased estimation of the effects of X alone, because the difference in performance between Y alone and X + Y provides an estimate of the added effect of X in the presence of Y. A “complete” design would also include randomization to X alone, so that this response could be compared with the response obtained in the absence of X and Y, allowing a separate effect of X to be estimated.

A special category of designs known as “crossover designs” appears on occasion in the E and M literature but may not be identified as such, so that estimation of means and variances may be biased. In the simplest of these designs, all treatment conditions are applied sequentially, and in some random order, to each subject. For example, one-half of the subjects may be randomly assigned to receive an infusion of compound A first and compound B second, with the remainder randomly assigned to receive the two infusions in the order of B and then A. Crossover designs are rife with opportunities to introduce estimation bias into estimates of treatment effects and experimental error. The possibilities of carryover effects from one treatment application to the next within a subject and for change in outcome over time within subjects, regardless of treatment ordering, are among the factors that should be considered in estimation. For more information see Ref. 8.

CATEGORY VI: ESTIMATION IMPRECISION

Like estimation bias, estimation imprecision is error in the estimation of the value of a parameter; however, unlike estimation bias, estimation imprecision lacks consistent tendency in direction and is therefore sometimes referred to as estimation “noise,” “instability,” or “unreliability.” Of course, small sample sizes are a common reason for imprecise estimates in the E and M literature (discussed in categories I and II). However, one can treat that source of imprecision as distinct from estimation imprecision, because it is introduced during the sampling process. Instead, in this section, we consider only imprecision that is introduced during estimation.

Estimation imprecision can potentially be reduced through careful choice of estimation methods. Statisticians refer to such reductions as improvements in estimation “efficiency,” because greater precision is obtained from a data set of a given sample size. Common situations in which reductions arise are when only a portion of available data is analyzed (e.g., the last 10 min of recordings), or some form of data reduction is performed before estimation. To illustrate the latter, suppose one is investigating the impacts of body weight on concentrations of a circulating hormone. Ten subjects are sampled within each of eight weight categories. Responses for subjects within each category are averaged, and then average response is regressed on average weight, so that the regression line is estimated with eight pairs of means. This type of averaging often can reduce estimation efficiency seriously, because the quantity of data used for regression analysis is smaller, sometimes much smaller (tenfold in our example) than the true total sample size, and an unnecessarily large standard error for estimates of regression coefficients results. In some instances, however, the loss in effective sample size can be compensated for by the variance reduction that comes with averaging, especially if variances in response are much larger within than among averaged groups. Because efficiency is an advanced statistical topic, the researcher is advised to consult with a statistician on these issues in his/her work.
Research results are never certain. Hypothesis testing recognizes statistical uncertainty by making probability statements about observed findings. In this section and the next, we focus on bias and imprecision pertaining to these probability statements, beginning in this section with a discussion of bias in Type I error. Recall that Type I error is the probability of incorrectly rejecting a null hypothesis. Useful pneumonics are to refer to Type I error as the false-positive or false-alarm rate. Typically studies set Type I error rates at, say, 5% to serve as the “nominal” level. In this section, we define bias as a difference of consistent tendency in direction between actual and nominal Type I error rates.

In the course of reviewing data from a study, a pattern may appear to the researcher which he/she thinks may be worthy of further study. However, within a study, perhaps the best rule to follow is that hypothesis formulation should dictate what data are collected and used for testing rather than allowing collected data to direct which hypotheses to test. A subtle, but at times decisive, form of bias in Type I error can arise when one chooses between one-tailed and two-tailed testing. When, say, one wishes to compare means between two groups, a one-tailed test corresponds to an alternative hypothesis, in which one population’s mean is strictly larger than the other’s. In contrast, a two-tailed test applies to an alternative hypothesis, which states that the two means differ, without specifying which is larger than the other. The advantage of a one-tailed test is that it offers greater statistical power (i.e., probability of correctly rejecting the null) in the specified direction. One-tailed tests are scientifically justifiable in two special circumstances: 1) a difference between two means in one of two directions is known to be impossible or 2) a difference in one direction is of no interest whatsoever in any circumstance. For example, suppose a researcher anticipates that a compound will elevate an average glucose-cycling rate but acknowledges that the compound may, for reasons not yet understood or foreseen, depress the average rate. This possibility must be allowed by employing a two-tailed test, unless the researcher would never be interested in recognizing such a counter-theoretical result. Because these two instances are rare in this literature, one-tailed testing should be rare, and if employed, given very strong justification, which is not always the practice in the E and M literature. When a one-tailed test is employed outside these constraints, the nominal Type I error is less than the actual.

Another arena in which actual Type I error rate may exceed the nominal rate is with multiple hypothesis testing. Multiple hypothesis testing is common in the E and M literature, especially where more than one type of measurement is collected on each subject (e.g., as may be reported in a table of data on free fatty acids, triglycerides, glycerol, glucose, and so on) and a separate test performed on each measurement. Each test conducted carries a particular Type I error rate, so that across multiple tests, the total Type I error rate compounds. For a provocative, accessible, and playful discussion of this topic see Ref. 1. When one does wish to control for compounded Type I error, a number of powerful methods are available, including those described in Refs. 6 and 11, with the latter suggested for the nonstatistician.

The researcher should also be aware that the results of a hypothesis test are only as good as the probability model on which they are based. Take, for example, the two-sample t-test, which is employed very widely in the E and M literature. Contrary to many practitioners’ beliefs, the two-sample t-test does not require that each group’s population be approximately normally distributed (although this helps). Rather, it assumes that the distribution of differences in sample means is approximately normally distributed. To illustrate, suppose that we sample n subjects from each group and calculate the difference in their sample means. Then we draw another sample of n subjects from each group and calculate the differences in these two new sample means. Repeating this process infinitely many times generates the full distribution of differences in sample means. It is this sampling distribution that we are assuming is normal in the two-sample t-test. Because of a statistical property described by the Central Limit Theorem, this distribution of sample means tends toward the normal as n increases in size. However, when n is small and the populations corresponding to the two groups are strongly skewed or possess multiple modes, then the sampling distribution of the differences in means may be nonnormal, and use of the t-test may lead to actual Type I error rates that are either smaller or larger than the nominal rate.

The E and M researcher is advised to know what assumptions underlie a specific method of hypothesis testing and to assess if his/her data meet these, at least approximately. Small sample sizes, common in the E and M literature, limit the analyst’s ability to check whether data meet a test’s assumptions. As a very rough rule, it is desirable to have ≥30 subjects per population sampled to permit adequate examination of the assumptions underlying many statistical tests. Even so-called nonparametric tests (which is a misnomer, for many of these tests are often used to test hypotheses regarding parameters) can be based on specific assumptions about the populations that have been sampled (see, for example, Ref. 2). The E and M literature also contains some examples of more sophisticated statistical modeling methods (e.g., repeated-measures ANOVA, which is common in the E and M literature and makes very strong assumptions about correlations among repeated measurements). With this added sophistication comes a complexity of assumptions that should be carefully examined. Statements such as “appropriate regression procedures were applied” are inadequate alone. When a statistical method’s assumptions are examined, these results should be reported. Choosing an appropriate method for fitting data and testing hypotheses requires striking a balance. Methods that rely on many strong assumptions can be quite powerful when data clearly meet those assumptions but can introduce appreciable bias in Type I error when those assumptions are violated.

Imprecision in hypothesis testing is measured by Type II error. Type II error is the probability of failing to accept the alternative hypothesis when true. To illustrate, one might pose a null hypothesis that two populations’ means do not differ and an alternative hypothesis that they do. If, in reality, the two populations’ means differ and we fail to detect this, we have
committed a Type II error. The complementary probability to Type II error is statistical power. That is, if we denote Type II error by \( \beta \), power is given by \( 1 - \beta \). Power is the probability of rejecting the null hypothesis when the alternative is true.

Type II error can grow from any source of imprecision that arises in the process that leads up to hypothesis testing. Thus Type II error increases with 1) smaller sample sizes (common in the E and M literature), 2) larger technical error (category IV), or 3) less efficient estimators (category VI).

Beyond these sources, which have already been discussed, the E and M literature contains other examples of lost opportunities to enhance statistical power. An example is the use of unequal sample sizes. In the E and M literature, one can find statements such as “5–25 specimens were measured per group” or “each group consisted of a minimum of 4 subjects.” Broadly speaking, when comparing groups’ means for a fixed total sample size, power is greatest when sample sizes are equal for those groups (e.g., 15).

As indicated in the previous section, multiple testing is common in the E and M literature, in part because studies typically take measurement on more than one parameter for each subject. As a result, most characterizations of subjects are multivariate, which makes sense scientifically given that most metabolic and endocrine processes involve multiple parameters. Despite this, instances of multivariate hypothesis testing, in which one conducts tests on multiple parameters simultaneously (e.g., Hotelling’s \( T^2 \) test), are rare to nonexistent in the E and M literature. Instead, several univariate (single-parameter) tests are conducted. The shortcomings of this wholly univariate approach is that univariate hypothesis testing of multivariate processes can sometimes result in a failure to detect patterns of scientific interest. Rencher (10) provides an introduction to multivariate hypothesis testing that is reasonably accessible to nonstatisticians.

CATEGORY IX: REPORTING BIAS

A form of reporting bias that may be undetectable to even the most astute reader is the failure to report. Sometimes referred to as the “file drawer problem,” results go unpublished because an author or editor chooses not to publish “statistically insignificant” findings. This generates bias in the literature. Although it is true that underpowered studies are poor candidates for funding on ethical (4) and statistical grounds, one may not wish to strictly equate statistical insignificance with biological insignificance. For instance, an otherwise fully powered study may be worthy even if an underpowered portion lacks a statistically significant result, because that result may nevertheless provide an empirical suggestion, admittedly weak, for hypothesis generation. Such a hypothesis could be more rigorously tested in subsequent, fully powered studies. This is not to say, however, that the importance of findings from underpowered research should ever be overstated. On the flipside are studies designed with ample power that generate negative findings. These too can serve an important role in the literature by directing subsequent research away from apparently fruitless avenues.

Statistically insignificant results can also give rise to another form of reporting bias. Suppose in testing a null hypothesis of no difference between two populations against an alternative claiming that their means differ, we obtain an attained significance level of \( P = 0.13 \). Authors are advised to avoid stating in conclusive terms that no difference exists between the populations, which is an overstatement (and thus directional error). As an example, one may come across statements such as “incubation with the inhibitor failed to alter the rate of metabolic transport.” A more accurate conclusion would be that “no change in rate was detected statistically.” As mentioned above, if one is seeking statistical support for the assertion that two groups do not differ, then bioequivalence testing should be employed (see category III).

CATEGORY X: REPORTING IMPRECISION

Like the file drawer problem discussed above, reporting imprecision arises when useful information is withheld from the reader, but here any resulting misunderstanding is not directional. An example from the E and M literature is when sample means are reported “±” a second value, but it is not indicated if the second value is a sample standard deviation (estimate of the population standard deviation), standard error (estimate of the standard deviation of the sample means), or some other measure of dispersion. Identification of which specific dispersion parameter has been estimated is necessary if other researchers are to use these estimates to plan sample sizes in future research. Another example found in this literature is when a result is presented along with an attained significance level (\( P \) value), but no clear information is provided on what statistical method was employed to obtain that result. This leaves the reader with no means of assessing the value of the result of a hypothesis test.

RESULTS AND DISCUSSION

This essay has illustrated ten categories of statistical errors. It is intended to highlight potential statistical pitfalls for those who are designing, implementing, and reviewing E and M research. Some accessible references to the statistical literature have been given throughout to provide an entry for those who are interested in learning more about the topics discussed.

A review of the recent E and M literature finds that the most common potential cause of statistical error is small sample size \( (n \leq 10) \). Very small sample sizes 1) result in parameter estimates that are unnecessarily imprecise, 2) enhance potential for failed randomizations, and 3) yield hypothesis testing that is underpowered, or 4) yield hypothesis testing that is biased because assumptions underlying the applied statistical methods could not be examined adequately. Missing data exacerbate this problem by further reducing and unbalancing sample size and, when “informative,” missing data can introduce bias. E and M research could also see major improvements through more widespread use of stratified sampling designs, careful selection and implementation of experimental controls, use of adaptive and stratified randomization procedures, routine reporting of technical error, application of bioequivalence testing to studies designed to demonstrate equivalency, identification of statistically efficient means of data reduction through consultation with a statistician, more restrained use of one-tailed tests, greater controls on Type I and Type II errors for hypothesis tests across multiple parameters, fuller descriptions of statistical methods and methods of examining methodological assumptions, and more frequent reporting on amply powered but statistically nonsignificant results.
Despite these limitations, the E and M literature appears to contain fewer statistical errors, in kind and quantity, than other medical literatures examined by the author. In part, this may be for lack of opportunity, as the range of statistical methods employed in the E and M literature is comparatively narrow. This is a lost opportunity, as processes studied in E and M are typically high dimensional and time varying (e.g., electrolyte concentration vs. time since initiation of dialysis), which makes this discipline ripe for greater application of multivariate (5, 10) and more flexible longitudinal (3) statistical designs and analyses. For example, generalized estimating equations (3) offer a highly flexible and robust method of analyzing repeated-measures data, especially when no data are missing; dimensionality-reduction techniques, such as principal-components analysis and cluster analysis (5), are useful for forming a few strata from several baseline characteristics; and these are just a few possibilities. The capacity for expanding the utility and sophistication of multivariate statistics in application to E and M research is tremendous.

REFERENCES


