Fundamental concepts in statistics: elucidation and illustration

DOUGLAS CURRAN-EVERETT, SUE TAYLOR, AND KAREN KAFADAR
Departments of Pediatrics and of Preventive Medicine and Biometrics, School of Medicine, University of Colorado Health Sciences Center, Denver, 80262; and Department of Mathematics, University of Colorado at Denver, Denver, Colorado 80217-3364

Curran-Everett, Douglas, Sue Taylor, and Karen Kafadar. Fundamental concepts in statistics: elucidation and illustration. J. Appl. Physiol. 85(3): 775–786, 1998.—Fundamental concepts in statistics form the cornerstone of scientific inquiry. If we fail to understand fully these fundamental concepts, then the scientific conclusions we reach are more likely to be wrong. This is more than supposition: for 60 years, statisticians have warned that the scientific literature harbors misunderstandings about basic statistical concepts. Original articles published in 1996 by the American Physiological Society’s journals fared no better in their handling of basic statistical concepts. In this review, we summarize the two main scientific uses of statistics: hypothesis testing and estimation. Most scientists use statistics solely for hypothesis testing; often, however, estimation is more useful. We also illustrate the concepts of variability and uncertainty, and we demonstrate the essential distinction between statistical significance and scientific importance. An understanding of concepts such as variability, uncertainty, and significance is necessary, but it is not sufficient; we show also that the numerical results of statistical analyses have limitations.

confidence interval; estimation; tolerance interval; uncertainty; variability

There are very few things which we know, which are not capable of being reduc’d to a Mathematical Reasoning, . . . and where a Mathematical Reasoning can be had, it’s as great folly to make use of any other, as to grope for a thing in the dark when you have a Candle standing by you.

John Arbuthnot (1692)

Statistics is one kind of mathematical reasoning. Its concepts and principles are ubiquitous in science: as researchers, we use them to design experiments, analyze data, report results, and interpret the published findings of others. Indeed, it is from this foundation of statistical concepts and principles that scientific knowledge is accumulated. If we fail to understand fully these fundamental statistical concepts and principles—if our statistical reasoning is faulty—then we are more likely to reach wrong scientific conclusions. Wrong conclusions based on faulty reasoning is shoddy science; it is also unethical (1, 21, 30).

Regrettably, faulty reasoning in statistics rears its head in the practice of science: for 60 years, statisticians have documented statistical errors in the scientific literature (3, 4, 17, 33, 50). In part, these errors exist because many introductory textbooks of statistics paradoxically hinder literacy in statistics: they emphasize methods rather than concepts, they contain glaring errors, or they perpetuate misconceptions (4, 11, 12).

In his editorial prelude to a series of statistical papers, Yates (51) wrote that the papers were designed to raise statistical consciousness and thereby reduce statistical errors in journals published by the American Physiological Society. Rather than reinforce concepts, these papers reviewed methods: analysis of variance (20), linear regression (37, 46), mathematical modeling (22, 29, 40), risk assessment (36), and statistical packages (34). The proper use of any statistical technique, however, requires an understanding of the fundamental statistical concepts behind the technique.

How well do physiologists understand fundamental concepts in statistics? One way to answer this question is to examine the empirical incidence of basic statistical quantities such as standard deviations, standard errors, and confidence intervals. These quantities characterize different statistical features: standard deviations characterize variability in the population, whereas standard errors and confidence intervals characterize uncertainty about the estimated values of population parameters, e.g., means. Of the original articles published in 1996 by the American Physiological Society, the overwhelming majority (69–93%, range) report standard errors, apparently not as estimates of uncer-
partnership that typically reaps great rewards.

In this review, we summarize the primary scientific uses of statistics. Then, we illustrate several fundamental concepts: variability, uncertainty, and significance. Last, we illustrate that although an understanding of concepts such as variability, uncertainty, and significance is necessary, it is not sufficient; it is essential to realize also that the numerical results of statistical analyses have limitations.

### Glossary

- $\alpha$: Critical significance level
- Ave: Average of the quantity q
- $\mu$: Population mean
- $\nu$: Degrees of freedom
- N: Normal (Gaussian) distribution with mean $\mu$ and variance $\sigma^2$
- P: Achieved significance level
- Pr: Probability of event A
- $\sigma$: Population standard deviation
- $\sigma_r$: Standard deviation of the sampling distribution of the sample mean
- $s$: Sample standard deviation
- $\sigma^2$: Population variance
- $s^2$: Sample variance
- SE: Standard error of the quantity q
- Var: Variance of the quantity q
- Y: Random variable
- $y_i$: Sample observation i, where $i = 1, 2, \ldots, n$
- $\bar{y}$: Sample mean

### SCIENTIFIC USES OF STATISTICS

In science, there are two main uses of statistics: hypothesis testing and estimation. Most researchers use statistics solely for hypothesis testing. In many situations, statisticians play down hypothesis testing and prefer estimation instead.

Hypothesis testing. To test a scientific hypothesis, a researcher must formulate the hypothesis before any data are collected, then design and execute an experiment that is relevant to it. Because the hypothesis is most often one of no difference, the hypothesis is called, by tradition, the null hypothesis. Using data from the experiment, the researcher must next compute the observed value $T$ of a test statistic. Finally, the researcher must compare the observed value $T$ with some critical value $T^*$, chosen from the distribution of the test statistic that is based on the null hypothesis. If $T$ is more extreme than $T^*$, then that is a surprising result if the null hypothesis is true, and the researcher is entitled, on statistical grounds, to become skeptical about the scientific validity of the null hypothesis.

The statistical test of a null hypothesis is useful because it assesses the strength of the evidence: it helps guard against an unwarranted conclusion, or it helps argue for a real experimental effect (19, 48). Nevertheless, a null hypothesis is often an artificial construct: before any data are recorded, the investigator knows—at least, suspects—that the null hypothesis is not exactly true. Moreover, the only question a hypothesis test can answer is a trivial one: is there anything other than random variation here?

Statisticians have emphasized repeatedly the limited value of hypothesis testing (2, 4, 9, 18, 24, 28, 31, 38, 50). In fact, the P values that result from hypothesis tests have been described as "absurdly academic" (25) and as having a "strictly limited role" (19) in data analysis. Within the scientific community, unwarranted

---

1. The adjective "null" can be misleading: this hypothesis need not be one of no difference. The use of null persists because of historical inertia.
2. Kruskal (38) reviews other drawbacks to hypothesis testing.
3. Sir Ronald Fisher, the author of this phrase, developed many statistical procedures, including the analysis of variance.
A parameter is a numerical constant: for example, the population mean.
ESTIMATING VARIABILITY IN THE POPULATION

The preceding sample observations, \(-33, -15, \ldots, -7\), differ because the population from which they were drawn is distributed over a range of possible values. This intrinsic variability is more than a distraction; it is an integral part of statistics, and the careful study of variability may reveal something about underlying scientific processes (25). The most common measure of the variability among sample observations is the sample standard deviation \(s\), the square root of the sample variance \(s^2\):

\[
\sigma^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2
\]

(See also Refs. 2, 9, 42, and 48.) The sample standard deviation characterizes the typical distance of an observation from the distribution center; in other words, it reflects the dispersion of individual sample observations about the sample mean. The sample standard deviation \(s\) also estimates the population standard deviation \(\sigma\): the standard deviation of the sample observations \(-33, -15, \ldots, -7\) is \(s = 15.2\), which estimates \(\sigma = 20\).

Most journals would publish the preceding sample mean and standard deviation as

\[
-8.2 \text{ mmHg} \pm 15.2
\]

The \(\pm\) symbol, however, is superfluous: the standard deviation is a single positive number. A standard deviation can be reported clearly with notation of this form

\[
-8.2 \text{ mmHg (SD 15.2)}
\]

In a table, the symbol SD can be omitted without loss of clarity as long as the table legend identifies the parenthetical value as a standard deviation.

The standard deviation is often a useful index of variability, but in many experimental situations it may be a deceptive one: even subtle departures from a normal distribution can render useless the standard deviation as an index of variability (43); often, the distribution of a biological variable differs grossly from a normal distribution. As one example, the distribution of values for plasma creatinine (26) resembles the skewed distribution depicted in Fig. 2. When the tails of a distribution are elongated, as is the right tail of this skewed distribution, the sample standard deviation will be an inflated measure of variability in the population (43, 48). There are two remedies to this misrepresentation of variability by the standard deviation: use another measure of variability, or transform the data.

Alternative measures of variability. Two measures of variability that are useful with a variety of distributions are the mean absolute deviation and the interquartile range. The mean absolute deviation (\(\text{Ave} |\text{dev}|\)) is the average distance of the sample observations from the sample mean

\[
\text{Ave} |\text{dev}| = \frac{1}{n} \sum_{i=1}^{n} |y_i - \bar{y}|
\]

The interquartile range (often designated as IQR) encompasses the middle 50% of a distribution and is the difference between the 75th and 25th percentiles. For \(0 < \varphi < 1\), the 100\(\varphi\)th percentile is the value below which 100\(\varphi\)% of the distribution is found.

Data transformation. When the sample observations happen to be drawn from a population that has a skewed distribution (e.g., a constituent of blood or the growth rate of a tumor), a transformation may change the shape of their distribution so that the distribution of the transformed observations is more symmetric (14, 23, 26, 32, 48). Common transformations include the logarithmic, inverse, square root, and arc sine transformations. The appendix reviews a useful family of data transformations.
In the next section, we revisit the unknown discrepancy between the sample estimate of a population parameter and the population parameter itself.

**ESTIMATING UNCERTAINTY ABOUT A POPULATION PARAMETER**

In the sampling exercise from *USING SAMPLES TO LEARN ABOUT POPULATIONS*, the sample mean \( \bar{y} = 8.2 \) (Eq. 2) estimated the population mean \( \mu_1 = 15 \). If we had calculated this sample mean from experimental observations, then we would be uncertain about the magnitude of the discrepancy between the sample estimate \( \bar{y} \) and the population parameter \( \mu_1 \). The ability to estimate the level of uncertainty about the value of a population parameter by using the sample estimate of that parameter is a powerful aspect of statistics (47).

Suppose we measure the same response variable, the change in systolic blood pressure, in a second sample of 10 independent observations drawn from the same population. We know beforehand that because of random sampling the mean of the second sample, \( \bar{y}_2 \), will differ from the mean of the first sample, \( \bar{y}_1 = \mu_1 = 8.2 \). If we measure the change in systolic blood pressure in 100 samples of 10 independent observations, then we expect 100 different estimates of the population mean \( \mu_1 \); for example

\[
\bar{y}_1 = 8.2, \quad \bar{y}_2 = 8.1, \quad \cdots, \quad \bar{y}_{100} = 22.5
\]

If we treat these 100 observed sample means as 100 observations, then we can calculate their mean and standard deviation, designated as \( \text{Ave} \bar{y} \) and \( \text{SD} \bar{y} \)

\[
\text{Ave} \bar{y} = 14.5 \quad \text{and} \quad \text{SD} \bar{y} = 6.07
\]

We can generalize from this empirical distribution of sample means to a theoretical distribution of the sample mean for a sample of size \( n \). Consider a random variable \( Y \) that is distributed normally with mean \( \mu \) and variance \( \sigma^2 \), which are known; the notation for this normal distribution is \( Y \sim N(\mu, \sigma^2) \). If an infinite number of samples, each with \( n \) independent observations, is drawn from this normal distribution, then the sample means \( \bar{y}_1, \bar{y}_2, \ldots, \bar{y}_n \) will also be distributed normally.\(^8\) The average of the sample means, \( \text{Ave} \bar{y} \), is the population mean \( \mu \), but the variance of the sample means (\( \text{Var} \bar{y} \)) is smaller than the population variance \( \sigma^2 \) by a factor of \( 1/n \)

\[
\text{Ave} \bar{y} = \mu \quad \text{and} \quad \text{Var} \bar{y} = \sigma^2 / n
\]

(The appendix derives these expressions. Figure 3 develops these expressions using empirical examples.)

Therefore, the standard deviation of the theoretical distribution of the sample mean, \( \sigma_{\bar{y}} \), is

\[
\sigma_{\bar{y}} = \sigma / \sqrt{n}
\]

If the sample size \( n \) increases, then the standard deviation \( \sigma_{\bar{y}} \) will decrease; that is, the more sample observations we have, the more certain we will be that the point estimate \( \bar{y} \) is near the actual population mean \( \mu \).

The standard deviation of the theoretical distribution of the sample mean is known also as the standard error of the sample mean, \( \text{SE} \bar{y} \), that is

\[
\text{SE} \bar{y} = \sigma / \sqrt{n}
\]

In estimation, the standard error of the mean has no particular value; instead, it is useful because of its role

\(^8\)The Central Limit Theorem states that the theoretical distribution of the sample mean will be approximately normal, regardless of the distribution of the original observations (35, 42). If the distribution of the original observations happens to be normal, then the theoretical distribution of the sample mean will be exactly normal.
in the calculation of a confidence interval for the population mean \( \mu \).

Confidence intervals. When we construct a confidence interval for the population mean, we assign numerical bounds to the expected discrepancy between the sample mean \( \bar{y} \) and the population mean \( \mu \). In essence, a confidence interval is a range that we expect, with some level of confidence, to include the actual value of the population mean. Below, we use the theoretical distribution of the sample mean to derive the confidence interval for the population mean \( \mu \).

In the theoretical distribution of the sample mean, 100\((1 - \alpha)\)% of the possible sample means is included in the interval

\[
[ \mu - a, \mu + a ]
\]

where the allowance \( a \) is

\[
a = z_{\alpha/2} \cdot \text{SE} | \bar{y} |
\]

In Eq. 5, \( z_{\alpha/2} \) is the 100\((1 - (\alpha/2))\)th percentile from the standard normal distribution, i.e., a normal distribution with mean 0 and variance 1, and \( \text{SE} | \bar{y} | \) is defined by Eq. 3. Therefore, when the population standard deviation \( \sigma \) is known, 95% of the possible sample means are within 1.96 \( \text{SE} | \bar{y} | \) of the population mean \( \mu \).

The interval in Eq. 4 can be written as the probability expression

\[
\Pr[ \mu - a \leq \bar{y} \leq \mu + a ] = 1 - \alpha
\]

which declares that the probability is 1 - \( \alpha \) that a sample mean lies within the interval \([ \mu - a, \mu + a ]\). After algebraic rearrangement, this expression can be written

\[
\Pr[ \bar{y} - a \leq \mu \leq \bar{y} + a ] = 1 - \alpha
\]

but note that the randomness resides in the parameter estimate \( \bar{y} \), not in the actual parameter \( \mu \). In this form, the interval

\[
[ \bar{y} - a, \bar{y} + a ]
\]

is called the 100\((1 - \alpha)\)% confidence interval for the population mean \( \mu \).

In practice, the sample standard deviation \( s \) estimates the population standard deviation \( \sigma \), which means that \( s/\sqrt{n} \) estimates the standard error of the mean (Eq. 3). In calculating a 100\((1 - \alpha)\)% confidence interval for the mean \( \mu \), this uncertainty about the actual value of \( \sigma \) is handled by replacing \( z_{\alpha/2} \) in Eq. 5 with \( t_{\alpha/2, n} \), the 100\((1 - (\alpha/2))\)th percentile from a Student's distribution with \( v = n - 1 \) degrees of freedom. Therefore, the allowance applied to the sample mean to obtain the 100\((1 - \alpha)\)% confidence interval for the population mean (Eq. 6) is

\[
a = t_{\alpha/2, n} \cdot \text{SE} | \bar{y} |
\]

where \( \text{SE} | \bar{y} | = s/\sqrt{n} \). Note that this allowance exceeds the allowance in Eq. 5; there is greater uncertainty about the value of the population mean \( \mu \). This happens because if \( \nu < \infty \), then \( t_{\alpha/2, \nu} > z_{\alpha/2} \) for all values of \( \alpha \).

Suppose we want to calculate a confidence interval for the population mean \( \mu_1 = -15 \) by using the observations \( -33, -15, \ldots, -7 \) of the first sample. The mean and standard deviation of these 10 observations are \( \bar{y} = -8.2 \) and \( s = 15.2 \). Therefore, the estimated standard error of the mean is

\[
\text{SE} | \bar{y} | = s/\sqrt{n} = 15.2/\sqrt{10} = 4.81
\]

Because \( n = 10 \), there are \( v = n - 1 = 9 \) degrees of freedom. If we want a 95% confidence interval, then \( \alpha = 0.05 \), \( t_{0.025, 9} = 2.26 \), and the allowance \( a = 2.26 \times 4.81 = 10.9 \). Therefore, the 95% confidence interval is

\[
[-19.1, +2.7]
\]

In other words, we can declare, with 95% confidence, that the population mean is included in the interval \([-19.1, +2.7]\).

Bear in mind that a single confidence interval either does or does not include the value of the population

---

9 References 2, 9, 42, and 48 discuss the calculation of confidence intervals for other population parameters.

10 Moses (Ref. 42, p. 113–117) illustrates further the concept of a confidence interval by using empirical examples.
parameter; in experimental situations, we are uncertain about which of these outcomes has occurred. Instead, the level of confidence in a confidence interval is based on the concept of drawing a large number of samples, each with n observations, from the population. When we measured the change in systolic blood pressure in 100 random samples, we obtained 100 different sample means and 100 different sample standard deviations. As a consequence, we will calculate 100 different 100(1 − α)% confidence intervals; we expect, 100(1 − α)% of these observed confidence intervals to include the actual value of the population mean (see Fig. 4).

A confidence interval characterizes the uncertainty about the estimated value of a population parameter. Sometimes, an investigator may be interested less in the value of the population parameter and more in the distribution of individual observations. A tolerance interval characterizes the uncertainty about the estimated distribution of those individual observations (see APPENDIX).

Next, we illustrate the distinction between statistical significance and scientific importance. Last, we show that the numerical results of statistical analyses have limitations.

**STATISTICAL AND SCIENTIFIC SIGNIFICANCE DIFFER**

Hypothesis testing, as the primary scientific use of statistics, has a drawback: the result of a hypothesis test conveys mere statistical significance. In contrast, estimation conveys scientific significance.11 This distinction is obvious if we use the results of a recent clinical trial. In this trial, the Systolic Hypertension in the Elderly Program (SHEP) Cooperative Research Group (45) evaluated the impact of antihypertensive drugs on the incidence of stroke in persons with isolated systolic hypertension. When compared with placebo, these drugs reduced by 36% (P = 0.0003) the incidence of stroke. Associated with this reduced incidence of stroke was a greater decrease in systolic blood pressure.

To appreciate the distinction between statistical significance and scientific importance, consider two populations that represent the theoretical distributions of the decreases in systolic blood pressure for the two groups. Let the decrease in systolic blood pressure of the placebo group be designated Y_1 and that of the drug treatment group be designated Y_2. Assume that Y_1 and Y_2 are distributed normally

\[ Y_1 \sim N(\mu_1, \sigma_1^2) \quad \text{and} \quad Y_2 \sim N(\mu_2, \sigma_2^2) \]

The normal probability density function (Eq. 1), in which approximate values for the observed sample means and variances from the SHEP trial, \( \bar{y}_1 \) and \( s_1^2 \), are substituted for the population means and variances, generates the population distributions depicted in Fig. 5

\[ \bar{y}_1 = -15 \Rightarrow \mu_1, \quad s_1^2 = 400 \Rightarrow \sigma_1^2 \]

\[ \bar{y}_2 = -25 \Rightarrow \mu_2, \quad s_2^2 = 400 \Rightarrow \sigma_2^2 \]

Suppose our objective is to estimate the difference between population means

\[ \mu_2 - \mu_1 = -25 - (-15) = -10 \text{ mmHg} \]

The SHEP group established convincingly that the difference \( \mu_2 - \mu_1 \), which represents the greater decrease in systolic blood pressure after drug therapy, was important. To estimate \( \mu_2 - \mu_1 \), we would sample at random from each population: the difference between sample means, \( \bar{y}_2 - \bar{y}_1 \), estimates the difference between population means, \( \mu_2 - \mu_1 \).

---

11 The word “significance,” when used to refer to scientific consequence, is ambiguous. Hereafter, we use the word “importance.”
Fig. 5. Statistical and scientific significance differ: placebo (black) and drug-treatment (gray) populations. The populations represent theoretical distributions of changes in systolic blood pressure during year 5 of the Systolic Hypertension in the Elderly Program clinical trial (see Ref. 45). The distributions are described by the normal probability density function (Eq. 1) in which the sample means and variances, $\bar{x}_1$ and $\bar{x}_2$, are substituted for the population means and variances. To generate samples of size $n$ from each population, observations (Obs) were drawn at random from the placebo population; corresponding observations from the drug-treatment population were obtained by subtracting 10 from each placebo observation. The sampling procedure is illustrated for $n = 2$.

By drawing samples of 2–128 observations from each population (Table 2) and by forcing $\bar{x}_2 - \bar{x}_1 = -10$ (see Fig. 5), the distinction between statistical significance and scientific importance becomes clear. As sample size $n$ grows, the statistical significance increases, from $P = 0.71$ for $n = 2$ to $P < 0.001$ for $n = 128$. Regardless of sample size, one aspect of scientific importance, that reflected by the difference $\bar{x}_2 - \bar{x}_1$, remains constant. As sample size increases, uncertainty about the actual difference $\bar{x}_2 - \bar{x}_1$, another aspect of scientific importance characterized by the numerical bounds of the confidence interval, decreases.

Practical considerations. In experimental situations, the distinction between statistical significance and scientific importance can be maintained by routinely addressing two questions: how likely is it that the experimental effect is real, and is the experimental effect large enough to be relevant? The first question can be answered simply: compare the $P$ value, obtained in the hypothesis test, with the critical significance level $\alpha$, chosen before any data are collected; if $P < \alpha$, then the experimental effect is likely to be real. The second question can be answered in two steps: calculate a confidence interval for the population parameter, and then assess the numerical bounds of that confidence interval for scientific importance; if either bound of the confidence interval is important from a scientific perspective, then the experimental effect may be large enough to be relevant.

Consider the results when 15 sample observations were drawn from the placebo and drug treatment populations: when compared with placebo, the greater decrease in systolic blood pressure after drug therapy was unconvincing from a statistical perspective ($P = 0.18$). Because the 95% confidence interval was $[-25, +5]$, uncertainty about the actual impact of drug treatment on systolic blood pressure is relatively large. Note, however, that the additional decrease in systolic blood pressure gained by drug treatment may have been as pronounced as 25 mmHg. From a scientific perspective, further studies, designed with greater statistical power, are warranted.

To illustrate that a significant statistical result may have little scientific importance, imagine that systolic blood pressure had been measured in mmHgO rather than in mmHg. Consider the results when 128 sample observations were drawn from the two populations: the greater decrease in systolic blood pressure after drug therapy was compelling from a statistical perspective ($P < 0.001$). If the confidence interval $[-15, +5]$ is expressed in mmHgO (by dividing each bound by 13.6), the investigator can declare, with 95% confidence, that the magnitude of the greater decrease in systolic blood pressure was 0.4–1.1 mmHgO. In this example, the

<table>
<thead>
<tr>
<th>$\bar{x}_2 - \bar{x}_1$</th>
<th>SE ($\bar{x}_2 - \bar{x}_1$)</th>
<th>95% Confidence Interval*</th>
<th>$t$</th>
<th>$P$ (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>-10</td>
<td>23.1</td>
<td>-110 to +90</td>
<td>0.43</td>
</tr>
<tr>
<td>4</td>
<td>-10</td>
<td>22.6</td>
<td>-65 to +45</td>
<td>0.44</td>
</tr>
<tr>
<td>8</td>
<td>-10</td>
<td>12.3</td>
<td>-36 to +16</td>
<td>0.81</td>
</tr>
<tr>
<td>10</td>
<td>-10</td>
<td>10.1</td>
<td>-31 to +11</td>
<td>0.99</td>
</tr>
<tr>
<td>15</td>
<td>-10</td>
<td>7.3</td>
<td>-25 to +5</td>
<td>-1.38</td>
</tr>
<tr>
<td>20</td>
<td>-10</td>
<td>5.8</td>
<td>-22 to +2</td>
<td>1.74</td>
</tr>
<tr>
<td>25</td>
<td>-10</td>
<td>5.3</td>
<td>-21 to +1</td>
<td>1.88</td>
</tr>
<tr>
<td>32</td>
<td>-10</td>
<td>4.7</td>
<td>-19 to -1</td>
<td>2.13</td>
</tr>
<tr>
<td>64</td>
<td>-10</td>
<td>3.7</td>
<td>-17 to -3</td>
<td>2.70</td>
</tr>
<tr>
<td>128</td>
<td>-10</td>
<td>2.4</td>
<td>-15 to -5</td>
<td>4.25</td>
</tr>
</tbody>
</table>

n, Sample size drawn from placebo (population 1) and drug treatment (population 2) populations (see Fig. 5). *Confidence interval for the difference between population means, $\mu_2 - \mu_1$ (see Eq. A2). $t$ Test statistic used to evaluate statistical significance of the difference $\bar{x}_2 - \bar{x}_1$ (see Eq. A3). $P$ Probability (2-tailed) that $\mu_2 - \mu_1 = 0$. The difference $\bar{x}_2 - \bar{x}_1$ and the 95% confidence interval for the difference $\mu_2 - \mu_1$ reflect the magnitude and uncertainty of the experimental results. The test statistic and its associated $P$ value reflect statistical significance. An increase in the no. of observations drawn from each population increases SE ($\bar{x}_2 - \bar{x}_1$); as a consequence, the statistical significance increases (irregularly, because of random sampling), but the estimated difference between population means remains constant at $\bar{x}_2 - \bar{x}_1 = -10$. The APPENDIX details the statistical equations required to perform this sampling exercise.
investigator can be quite certain of a trivial experimental effect.

Whatever the statistical result of a hypothesis test, assessment of the corresponding confidence interval incorporates the scientific importance of the experimental result.

LIMITATIONS OF STATISTICS

Although the process of scientific discovery requires an understanding of fundamental concepts in statistics, the use of statistics does have limitations. For example, not many of us would accept, solely on the basis of a close temporal relationship, that solar radiation governs stock market prices (Fig. 6). The limitations of statistics are more subtle if an association is plausible.

Imagine this scenario: a neurological syndrome results from impaired production of some neurotransmitter. Drugs A and B, derivatives of the same parent compound, both stimulate production of this neurotransmitter. Just one of the drugs, however, continues to increase neurotransmitter production over its entire therapeutic range. At higher doses, the second drug becomes less effective at boosting neurotransmitter production and causes neurotoxicity. For each drug, Table 3 lists administered drug concentrations and measured increases in neurotransmitter production. If you rely on only the regression statistics in Table 3, which drug is which? If you are unfortunate and happen to have this hypothetical syndrome, then your choice assumes added importance.

From the regression statistics alone, it is impossible to differentiate the drugs. Their identities are plain, however, when the data are plotted (Fig. 7): drug A increases neurotransmitter production over the entire range of drug concentrations; the increase in neurotransmitter production begins to fall at higher concentrations of drug B.

Practical considerations. Data graphics are essential also if the requisite assumptions behind a particular statistical technique are to be verified. For examples in regression, see chap. 3 in Ref. 23.

SUMMARY

It is depressing to find how much good biological work is in danger of being wasted through incompetent and misleading analysis . . .

Frank Yates and Michael J. R. Healy (1964)

This scathing remark, written almost 35 years ago (50) but relevant even now (4), reflects the frustrations felt by statisticians over the statistical misconceptions held by scientists. These misconceptions exist in large part because of shortcomings in the cursory statistics education we received in graduate or medical school (4, 11, 12). The major defect in most introductory courses in statistics is that fundamental concepts in statistics, the cornerstone of scientific inquiry (47), are neglected (5) but relevant even now (4), reflects the frustrations felt by statisticians over the statistical misconceptions held by scientists. These misconceptions exist in large part because of shortcomings in the cursory statistics education we received in graduate or medical school (4, 11, 12). The major defect in most introductory courses in statistics is that fundamental concepts in statistics, the cornerstone of scientific inquiry (47), are neglected rather than emphasized (4, 7, 17, 44, 50). Statisticians share responsibility with other faculty for ensuring

Table 3. Limitations of statistics: raw data and regression statistics

<table>
<thead>
<tr>
<th>Drug A</th>
<th>Drug B</th>
</tr>
</thead>
<tbody>
<tr>
<td>x y x y</td>
<td>x y x y</td>
</tr>
<tr>
<td>10 8.04 10 9.14 No. of observations (n) = 11</td>
<td>No. of observations (n) = 11</td>
</tr>
<tr>
<td>8 9.14 8 8.14 Average of x values (X) = 9.0</td>
<td>Average of x values (X) = 8.8</td>
</tr>
<tr>
<td>13 7.58 13 8.74 Average of y values (Y) = 7.5</td>
<td>Average of y values (Y) = 8.77</td>
</tr>
<tr>
<td>9 8.81 9 8.77 Equation of regression line: ( \hat{y} = 3 + 0.5x )</td>
<td>Equation of regression line: ( \hat{y} = 3 + 0.5x )</td>
</tr>
<tr>
<td>11 8.33 11 9.26 Standard error of estimate of slope ( [SE(b_1)] = 0.118 )</td>
<td>Standard error of estimate of slope ( [SE(b_1)] = 0.118 )</td>
</tr>
<tr>
<td>14 9.96 14 8.10 t(13): slope ( (b_1) = 4.24 )</td>
<td>t(13): slope ( (b_1) = 4.24 )</td>
</tr>
<tr>
<td>6 7.24 6 6.13 Sum of squares of x values ( S(x - \mu)^2 = 110.0 )</td>
<td>Sum of squares of x values ( S(x - \mu)^2 = 110.0 )</td>
</tr>
<tr>
<td>4.26 4 3.10 Regression sum of squares = 27.50</td>
<td>Regression sum of squares = 27.50</td>
</tr>
<tr>
<td>12 10.84 12 9.13 Residual sum of squares = 13.75</td>
<td>Residual sum of squares = 13.75</td>
</tr>
<tr>
<td>7 4.82 7 7.26 Correlation coefficient ( (r) = 0.82 )</td>
<td>Correlation coefficient ( (r) = 0.82 )</td>
</tr>
<tr>
<td>5 5.68 5 4.74 %Total sum of squares explained by regression ( (R^2) = 67% )</td>
<td>%Total sum of squares explained by regression ( (R^2) = 67% )</td>
</tr>
</tbody>
</table>

For drug A and B, values are administered drug concentration \( x \), measured increase in neurotransmitter production \( y \), and statistics from regression analysis of the first-order model \( Y = \beta_0 + \beta_1X + \epsilon \), where \( \epsilon \) is error. Additional regression analyses (23) reveal that this model is inappropriate for drug B (see Fig. 7). Data are from Anscombe (6).

Fig. 6. Limitations of statistics: solar radiation and New York stock market prices during 1929 (after Ref. 27). In general, increases in stock prices were associated with decreases in solar radiation. This nonsensical association illustrates the phenomenon of spurious correlation.

Fig. 7. Limitations of statistics: scatterplots of drug concentration \( x \) and increase in neurotransmitter production \( y \). For each drug, the fitted first-order model \( y = 3 + 0.5x \) and corresponding regression statistics are identical (see Table 3). For only drug A, however, is this first-order relationship plausible. For drug B, a second-order model of the form \( Y = \beta_0 + \beta_1X + \beta_2X^2 + \epsilon \) is required.
that introductory courses in statistics are relevant and sound (7, 44, 50).

In this review, we have reiterated the primary role of statistics within science to be one of estimation: estimation of a population parameter or estimation of the uncertainty about the value of that parameter. Moreover, we have demonstrated the essential distinction between statistical significance and scientific importance; of the two, scientific importance merits more consideration. We have shown also that without data graphics, data analysis is a game of chance. And last, that this review was written by a physiologist and two statisticians embodies one of the most basic notions in all science: collaboration.

**APPENDIX**

This APPENDIX reviews the lognormal distribution (a distribution that reveals limitations of the standard deviation as an estimate of variability), a versatile family of data transformations, the theoretical distribution of the sample mean, tolerance intervals, the statistical equations required to perform the significance sampling exercise, and the confidence interval for the difference between two population means.

Lognormal distribution. The lognormal distribution is a common probability distribution model for skewed data. The random variable Y is distributed lognormally if the logarithm of Y is distributed normally with mean \( \mu \) and variance \( \sigma^2 \), or \( \ln Y \sim N(\mu, \sigma^2) \). Formally, the lognormal probability density function \( g(y) \) is

\[
g(y) = \frac{1}{y \sqrt{2\pi} \sigma} \exp\left[-\frac{(\ln y - \mu)^2}{2\sigma^2}\right], \quad y > 0 \tag{A1}
\]

The mean \( \mu_y \) and variance \( \sigma_y^2 \) of the lognormal distribution specified by Eq. A1 are

\[
\mu_y = e^{\mu + \frac{\sigma^2}{2}} \quad \text{and} \quad \sigma_y^2 = e^{2\mu + \sigma^2} \left(e^\sigma - 1\right)
\]

For the distribution in Fig. 2, \( \tau = 1.803 \) and \( \xi^2 = 1 \); therefore, \( \mu_y = 10 \) and \( \sigma_y^2 = 172 \).

A family of data transformations. Box and Cox (14) have described a family of power transformations in which an observed variable \( y \) is transformed into the variable \( w \) by using the parameter \( \lambda \)

\[
w = \begin{cases} 
(y^\lambda - 1)/\lambda & \text{for } \lambda \neq 0, \\
\ln y & \text{for } \lambda = 0
\end{cases}
\]

The inverse (\( \lambda = -1 \)) and square root transformations (\( \lambda = 0.5 \)) are members of this family. Draper and Smith (Ref. 23, p. 225–226) summarize the steps required to estimate the parameter \( \lambda \) so that the distribution of \( w \) is as normal (Gaussian) as possible.

Theoretical distribution of the sample mean. Suppose some random variable \( X \) is distributed normally with mean \( \mu \) and variance \( \sigma^2 \): that is, \( X \sim N(\mu, \sigma^2) \). When a sample of \( n \) independent observations, \( x_1, x_2, \ldots, x_n \), is drawn repeatedly from this distribution, the observed sample means can be treated as observations. These sample means will be distributed normally with mean \( \mu \) and variance \( \sigma^2/n \), or

\[
\text{Ave } \overline{X} = \mu \quad \text{and} \quad \text{Var } \overline{X} = \sigma^2/n
\]

As you might expect, there is a mathematical foundation to these relationships.

Consider the linear function \( L \)

\[
L = k_1x_1 + k_2x_2 + \cdots + k_m x_m
\]

For \( i = 1, 2, \ldots, m \), each \( k_i \) is a real constant, and each \( x_i \sim N(\mu_i, \sigma_i^2) \). The mean of \( L \), \( \text{Ave } L \), is

\[
\text{Ave } L = k_1\mu_1 + k_2\mu_2 + \cdots + k_m \mu_m = \sum_{i=1}^{m} k_i \mu_i
\]

If \( x_1, x_2, \ldots, x_n \), are mutually independent, then the variance of \( L \), \( \text{Var } L \), is

\[
\text{Var } L = k_1^2\sigma_1^2 + k_2^2\sigma_2^2 + \cdots + k_m^2\sigma_m^2 = \sum_{i=1}^{m} k_i^2\sigma_i^2
\]

If the function \( L \) is \( x \), the mean of the sample observations \( x_1, x_2, \ldots, x_n \), then \( m = n \), and furthermore, for \( i = 1, 2, \ldots, n \)

\[
k_i = 1/n \quad \text{and} \quad x_i \sim N(\mu, \sigma^2)
\]

Therefore

\[
\text{Ave } L = \sum_{i=1}^{n} k_i \mu_i = \sum_{i=1}^{n} \mu/n = n(\mu/n) = \mu = \text{Ave } x
\]

and

\[
\text{Var } L = \sum_{i=1}^{n} k_i^2\sigma_i^2 = \sum_{i=1}^{n} \sigma_i^2/n^2 = n(\sigma^2/n^2) = \sigma^2/n = \text{Var } x
\]

Tolerance intervals. A tolerance interval identifies the bounds that are expected to contain some percentage of a population, not just a single population parameter such as the mean (41). If a normal distribution has mean \( \mu \) and variance \( \sigma^2 \), which are known, then the 100\( \% \) tolerance interval is

\[
[\mu - \{z_{1-\alpha/2} \cdot \sigma\}, \mu + \{z_{1-\alpha/2} \cdot \sigma\}]
\]

where \( z_{1-\alpha/2} \) is the \( 100(1 - (1 - \alpha)/2) \)th percentile from the standard normal distribution, i.e., \( N(0, 1) \). This tolerance interval covers exactly 100\( \% \) of the distribution. If \( \alpha = 0.95 \), then \( z_{1-0.025} = 1.96 \). For the population that represented the change in systolic blood pressure after some intervention (see USING SAMPLES TO LEARN ABOUT POPULATIONS), \( \mu = -15 \) and \( \sigma = 20 \); therefore, the exact 95\( \% \) tolerance interval is

\[
[-54, 24]
\]

In practice, the sample statistics \( \bar{y} \) and \( s \) are used to estimate the population parameters \( \mu \) and \( \sigma \). This element of uncertainty about the values of \( \mu \) and \( \sigma \) is handled by replacing \( z_{1-\alpha/2} \) with the confidence coefficient \( k \), where \( k \) depends on \( \alpha \) as well as the sample size \( n \). Therefore, the estimated 100\( \% \) tolerance interval is

\[
[y - ks, y + ks]
\]

If \( \alpha = 0.95 \) and \( n = \infty \), then \( k = z_{1-0.025} = 1.96 \) as above, when \( \mu \) and \( \sigma \) were known.] The coefficient \( k \) is chosen to enable the declaration, with 100\( (1 - \alpha) \% \) confidence, that the estimated tolerance interval covers 100\( \% \) of the distribution (see Table XIV in Ref. 41).

For the observations listed in USING SAMPLES TO LEARN ABOUT POPULATIONS, \( y = -8.2 \) and \( s = 15.2 \). Suppose we want to estimate with 95\% confidence a 90\% tolerance interval based on these results. When we use these percentages and the sample size of 10, the coefficient \( k \) is 2.839. Therefore, the
tolerance interval is

$$[-51, +35]$$

In other words, we can declare, with 95% confidence, that 90% of persons will have a change in systolic blood pressure of between −51 and +35 mmHg after the intervention. Note that this statement differs markedly from our previous assertion, made also with 95% confidence, that the population mean µ was included in the interval [−19.1, +2.7].

The tolerance intervals outlined above are appropriate only if the distribution of the underlying population is normal; other formulas exist to construct tolerance intervals when the population is distributed nonnormally.

Equations for the significance sampling exercise. For two samples of equal size n, the standard error of the difference between sample means, SE |Δy2 − Δy1|, is estimated as

$$\text{SE} |Δy2 − Δy1| = \frac{\sqrt{s_2^2 + s_1^2}}{\sqrt{n}}$$

where s² is sample variance.

The 100(1 − a)% confidence interval for µ2 − µ1, the difference between population means, is

$$[(\bar{y}_2 − \bar{y}_1) − a·(\text{SE} |Δy2 − Δy1| + a)]$$

(A2)

The allowance a applied to the difference Δy2 − Δy1 is

$$a = t_{n/2, \alpha/2}·\text{SE} |Δy2 − Δy1|$$

where t_{n/2, \alpha/2} is the 100(1 − (α/2))th percentile from a Student t distribution with ν = 2n − 2 degrees of freedom. In this sampling exercise, we use the t distribution because we assume the standard deviations of the populations are unknown (42).

The test statistic used to evaluate statistical significance of the difference Δy2 − Δy1 is

$$t \cdot H_0: \mu_2 − \mu_1 = 0 = \frac{(\bar{y}_2 − \bar{y}_1) − 0}{\text{SE} |Δy2 − Δy1|}$$

(A3)

Confidence interval for the difference between population means. In the significance sampling exercise (see statistical and scientific significance differ), we calculated a confidence interval for the difference between two population means. Rather than construct a confidence interval for this difference, a researcher could construct a confidence interval for each population mean: if the two confidence intervals fail to overlap, the researcher would conclude that the population means differ. This approach is conservative.

Consider the results when 32 sample observations were drawn from the placebo and drug treatment populations: when compared with placebo, drug therapy was associated with a greater decrease in systolic blood pressure (P = 0.04), and the 95% confidence interval for the difference between population means was [−19, −1]. That this confidence interval excludes 0 corroborates that the population means differ at the α = 0.05 level.

The observed sample means for the placebo and drug treatment groups were

Δy1 = −9.9 and Δy2 = −19.9

the standard errors of these sample means were

$$\text{SE} |Δy1| = \text{SE} |Δy2| = 3.32$$

Because n = 32, each sample has ν = n − 1 = 31 degrees of freedom.

If we want a 95% confidence interval for each population mean (Eq. 6), then α = 0.05, t_{n/2, \alpha/2} = 2.04, and the allowance

$$a = 2.04 \times 3.32 = 6.8.$$ Therefore, the 95% confidence interval for the mean of the placebo population is

$$[-17, −3]$$

the 95% confidence interval for the mean of the drug treatment population is

$$[-27, −13]$$

Because these individual confidence intervals overlap, we might conclude that there is insufficient evidence to declare that the two population means differ. In this example, 86% confidence intervals for the population means would just fail to overlap: that is, we could declare that the population means differ at the P = 0.14 level. Note that when we calculate a confidence interval for the difference between these population means, we are more confident that an actual difference exists.

We thank Brenda B. Rauner, Publications Manager and Executive Editor, APS Publications, for providing the information about research manuscripts published by the American Physiological Society. This review was supported in part by the Dept. of Pediatrics (M. Douglas Jones, Jr., Chair); by a Grant-in-Aid from the American Heart Association of Colorado and Wyoming (to D. Curran-Everett); and by the National Science Foundation Grant DMS 95-10435 (to K. Kafadar). Address for reprint requests: D. Curran-Everett, Dept. of Pediatrics, B-195, Univ. of Colorado Health Sciences Center, 4200 East 9th Ave., Denver, CO 80262 (E-mail: dcurrenever@castle.cudenver.edu).

REFERENCES
