Fundamentals of Clinical Research for Radiologists

Lawrence Joseph[1,2]
Caroline Reinhold[3,4]

Statistical Inference for Continuous Variables

Received November 5, 2004; accepted after revision November 11, 2004.

Series editors: Nancy Obuchowski, C. Craig Blackmore, Steven Karlik, and Caroline Reinhold.

This is the 15th in the series designed by the American College of Radiology (ACR), the Canadian Association of Radiologists, and the *American Journal of Roentgenology*. The series, which will ultimately comprise 22 articles, is designed to progressively educate radiologists in the methodologies of rigorous clinical research, from the most basic principles to a level of considerable sophistication. The articles are intended to complement interactive software that permits the user to work with what he or she has learned, which is available on the ACR Web site (www.acr.org).

Project coordinator: G. Scott Gazelle, Chair, ACR Commission on Research and Technology Assessment.

Staff coordinator: Jonathan H. Sunshine, Senior Director for Research, ACR.

[1]Department of Medicine, Division of Clinical Epidemiology, Montreal General Hospital, 1650 Cedar Ave., Montreal, QC H3G 1A4, Canada. Address correspondence to L. Joseph (Lawrence.Joseph@mcgill.ca).

[2]Department of Epidemiology and Biostatistics, 1020 Pine Ave. W, McGill University, Montreal, QC H3A 1A2, Canada.

[3]Department of Diagnostic Radiology, Montreal General Hospital, McGill University Health Centre, 1650 Cedar Ave., Montreal, QC H3G 1A4, Canada.

[4]Department of Oncology, Synarc, 575 Market St., San Francisco, CA 94105.

AJR 2005;184:1047–1056

0361–803X/05/1844–1047

Consider the following statements from an abstract reporting results from a study of CT in large cell neuroendocrine carcinoma of the lung [1]:

> In the 38 patients, six central tumors and 32 peripheral tumors, with diameters ranging from 12 to 92 mm (mean ± SD, 32 ± 19 mm), were identified. None of the tumors had air bronchograms or calcification in the mass or nodule… On contrast-enhanced CT scans, inhomogeneously enhanced tumors appeared to be larger (51 ± 18 mm) than homogeneously enhanced tumors (25 ± 10 mm; $p < 0.001$).

Proper interpretation of the above results, and of similar reports from much of the modern clinical literature, depends in large part on the understanding of statistical terms. In this case, terms such as "SD" were used for descriptive purposes, and p values were given to support evidence of between-group differences in tumor size. In other reports, one may see terms such as "confidence intervals," "t tests," "type 1 and type 2 errors," and so on. Clearly, radiologists who wish to keep pace with new technologies must at least have a basic understanding of statistical language. This is true not only if they desire to plan and perform their own research, but also if they simply want to read the medical literature with a keen critical eye or to make informed decisions about which new treatments or diagnostic techniques they may wish to use to treat their own patients.

Descriptive terms such as "means," "medians," and "SDs" have been covered in a previous article in this series [2]. Before reading this article, reviewing the previous modules on descriptive statistics [2] and probability and sampling [3] may be a good idea. In this module, we introduce the basic notions of inferential statistics—that is, we discuss how to draw inferences about one or more populations' characteristics using data from samples from these populations. We focus on continuous variables, including inferences for means and simple nonparametric methods. Rather than simply providing a catalogue of which formulas to use in which situation, we explain the logic behind each technique. In this way, informed choices and decisions can be made on the basis of a deeper understanding of exactly what information each type of statistical inference provides.

Recall from the discussion in a previous module [3] that there are two main schools of statistical inference: the frequentist school and the Bayesian school. These are each based on a different definition of probability, the frequentist school based on a long-run frequency definition and the Bayesian school based on a more subjective view of probability. We discuss these paradigms for statistical inference.

In the Statistical Inferences for Means section, the classical or frequentist school of statistical inferences for means is covered, and in the Nonparametric Inference section, we present a brief introduction to nonparametric inferences. In these sections, we explain exactly what is meant by ubiquitous statistical statements such as "$p < 0.05$"—which may not mean what many medical journal readers believe it to mean—and examine confidence intervals as an attractive alternative to p values. The problem of choosing an appropriate sample size for a given experiment is discussed in the Sample Size Calculations section. Increasingly important Bayesian alternatives to the

classical statistical techniques are presented in the Bayesian Inference section.

Statistical Inferences for Means

In this section, we consider how to draw inferences about populations by statistically analyzing samples of data using standard frequentist methods. We first consider inferences for a single mean when the variance in the population is known. We also initially assume that the data follow a normal distribution, so we are estimating the mean of this normal distribution. Once the basic concepts are understood in this simple case, we indicate how to extend the same ideas to cases in which the variance is unknown or more than one mean is of interest and to cases in which the normal distribution is not assumed.

In addition to the two different schools of inference (i.e., frequentist or Bayesian), statistical inferences can be divided into procedures that test a hypothesis and those that estimate parameters. We begin with hypothesis testing procedures that lead to *p* values, and then compare the information they provide to that provided by parameter estimation via confidence intervals.

Standard Frequentist Hypothesis Testing

Suppose we wish to test the hypothesis that a new accelerated radiation schedule for patients with brain cancer leads to smaller mean tumor diameters compared with the standard schedule versus a null hypothesis that the tumor diameters are the same regardless of schedule. Suppose further that it is known that patients on the standard schedule have a tumor diameter of 3.5 cm, on average, after completing their radiation therapy. Although it is somewhat unrealistic to assume this perfect knowledge of past tumor diameters, this example approximates the situation in which a large case series (e.g., a historical control series) of tumor diameters is available, so that most uncertainty arises from the data from the new schedule. Formally, we can state the hypotheses as:

$$H_0 \text{ (null hypothesis): } \mu = 3.5$$

$$H_A \text{ (alternative hypothesis): } \mu < 3.5$$

where μ represents the unknown true average tumor diameter of the accelerated radiation schedule.

There are four possible results when considering hypothesis testing, depending on the true state of nature, which is typically unknown, and the statistical test result, which depends on the data collected. The four possibilities are shown in Table 1.

According to Table 1, if the accelerated schedule in fact leads to smaller tumor diameters than the standard and we reject the null hypothesis, then we have made a correct decision, as also happens if the null hypothesis is in fact correct and we do not reject it. On the other hand, if we reject the null hypothesis as false when it is in fact true, we make a so-called type 1 error, which occurs with probability α, and if we fail to reject the null hypothesis when it is in fact false, we make a type 2 error, which occurs with probability β. The power of a study is defined as the probability of rejecting the null hypothesis when the alternative hypothesis is in fact true, so that the power is equal to $1 - \beta$. To summarize, we have equations 1–4:

$$\alpha = Pr\{\text{ rejecting } H_0 | H_0 \text{ is true}\} = \text{Type I error} \quad (1)$$

$$1 - \alpha = Pr\{\text{not rejecting } H_0 | H_0 \text{ is true}\} \quad (2)$$

$$\beta = Pr\{\text{not rejecting } H_0 | H_A \text{ is true}\} = \text{Type II error, and} \quad (3)$$

$$1 - \beta = Pr\{\text{ rejecting } H_0 | H_A \text{ is true}\} = \text{Power} \quad (4)$$

Recall from a previous module in this series [3] that probabilities written in the form of Pr{A | B} are called "conditional probabilities," and the notation is read as the probability that the event A occurs, given that the event B is known to have occurred. Thus, all of the quantities are conditional on knowing whether the null or alternative hypotheses are in fact true. Of course, we generally do not know whether the null hypothesis is true or not, so these conditional statements are at best of indirect interest. Once we obtain our data, we would ideally like to know the probability that the null hypothesis is true—not assume the null hypothesis is true. We will discuss this point further in the Bayesian Inference section.

Although it is important to understand the types of errors that can be made when hypothesis testing, the results of a hypothesis test are usually reported as a *p* value, which we now define: The *p* value is the probability of obtaining a result as extreme as or more extreme than that observed assuming that the null hypothesis is in fact true.

TABLE 1 Results of Hypothesis Testing

Test	True State of Nature	
	H_A	H_0
Reject H_0	$1 - \beta$	α
Do not reject H_0	β	$1 - \alpha$

It is important to note that the *p* value is not the probability that the null hypothesis is true after having seen the data, even though many clinicians often falsely interpret it this way. The *p* value does not directly or indirectly provide this probability and in fact can be orders of magnitude different from it. In other words, it is possible to have a *p* value equal to 0.05, when the probability of the null hypothesis is 0.5, different from the *p* value by a factor of 10 (see the Bayesian Inference section for how to calculate a more easily interpreted hypothesis test from a Bayesian viewpoint).

Given the definition of a *p* value, how would we calculate it? Suppose that we perform tumor measurements on 10 patients under the accelerated schedule and that these tumors have a mean diameter of $\bar{x} = 3.0$ cm, with a known SD of $\sigma = 1.5$ cm. The definition implies that we need to calculate the probability of obtaining mean tumor diameters of 3.0 cm or less (i.e., as extreme as or more extreme than what was observed), given that the true mean tumor diameter under the standard treatment schedule is exactly 3.5 cm (i.e., given the null hypothesis is true). Now, recall from a previous article in this series [3] that the probability density of our mean, $\bar{x}$, is usually considered as normal. Because for purposes of calculating a *p* value the null hypothesis is considered as exactly correct, the mean of our normal distribution is assumed to be 3.5 cm. The SD of our mean (known as the SE) is given as the SD in the population (assumed here to be $\sigma = 1.5$ cm) divided by the square root of the sample size [3]. Thus here, our SE is given by $1.5 / \sqrt{10} = 0.474$.

Therefore, we calculate equations 5 and 6:

$$p = Pr\{\text{of obtaining data as or more extreme than observed} | H_0 : \mu = 3.5\} \quad (5)$$

$$= Pr\{\bar{x} < 3.0 | \bar{x} \sim N(3.5, 0.474)\}. \quad (6)$$

This probability can be calculated from tables of the normal distribution, as explained in Joseph and Reinhold [3]. Normalizing, we find $Z = [(3.0 - 3.5) / 0.474] = -1.05$, and looking up −1.05 on standard normal tables, we find $p = 0.147$. Thus, there is about a 14.7% chance of obtaining results as extreme as or more extreme than the 3.0 cm observed, if the true mean tumor diameter for the new schedule is exactly 3.5 cm. Therefore, the observed result

is not unusual (i.e., it is compatible with the null hypothesis), so we cannot reject H_0.

Notice that if we had observed the same mean tumor diameter but with a larger sample size of 100, say, the p value would have been 0.0004. With a sample size of 100, the event of the observed data or data more extreme occurring would be a rare event if the null hypothesis were true, so the null hypothesis could be rejected. Therefore, p values depend not only on the observed mean tumor diameter, but also on the sample size.

The test described earlier was one-sided—that is, we a priori believed (perhaps from preliminary data or theoretic considerations) that the accelerated schedule would lead to equal or better results and not larger tumor sizes. To generalize, to perform a one-sided test of the null hypothesis that a single mean μ has value μ_0, calculate the statistic in equation 7:

$$z^* = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \tag{7}$$

and determine the p value from normal distribution tables as in equation 8:

$$p = Pr\{z > z^* | z \sim N(0,1)\} \tag{8}$$

On the other hand, often we may not want to specify the direction ahead of time. In this case, the alternative hypothesis is two-sided (i.e., the alternative hypothesis is H_A: $\mu \neq \mu_0$ rather than the one-sided H_A: $\mu < \mu_0$), and one performs the calculation in equation 9:

$$z^* = \left| \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \right| \tag{9}$$

where $|\, a \,|$ indicates the absolute value of a, and one determines the p value from normal distribution tables as in equation 10:

$$p = 2 \times Pr\{z > z^* | z \sim N(0,1)\} \tag{10}$$

In the one-sided case, we reject the null hypothesis only if we observe an extreme result in the direction specified by the alternative hypothesis. In the two-sided case, we reject if we observe an extreme result in either direction (larger or smaller tumor sizes). This results in a doubling of the p value, so for a two-sided alternative hypothesis (H_A: $\mu \neq 3.5$ in this case), we find $p = 2 \times 0.147 = 0.294$. The doubling results from adding the areas under the normal curve both below −1.05 (as in the one-sided case) and above 1.05.

Similar methods are available for tests involving comparisons between two means. For example, to test the null hypothesis that means in two different groups are equal to each other versus a two-sided alternative hypothesis, calculate as in equation 11:

$$z^* = \left| \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right| \tag{11}$$

For example, suppose we wish to again look at the difference in mean tumor diameter between two groups of patients with brain cancer, but this time in a clinical trial setting, with subjects randomized into accelerated and standard schedule groups (this would, of course, be a better design because concurrent groups are compared, minimizing potential confounding). Suppose we observe a mean tumor diameter of $\bar{x}_1 = 3.0$ cm ($\sigma_1 = 1.5$ cm) in 200 subjects under the new schedule, and a mean tumor diameter of $\bar{x}_2 = 3.7$ cm ($\sigma_2 = 1.4$ cm) in 200 subjects under the standard schedule. Plugging into the above formula, we get equation 12:

$$z^* = \left| \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right| = \left| \frac{3.0 - 3.7}{\sqrt{\frac{1.5^2}{200} + \frac{1.4^2}{200}}} \right| = 4.82 \tag{12}$$

Looking up 4.82 on normal tables gives a p value of $2 \times (0.0000007) = 0.0000014$. Because this indicates a very rare event under H_0, we can reject the null hypothesis that the two means are equal.

These formulas can be extended in a variety of directions, which we describe in the subsequent sections.

Paired versus unpaired tests.—In comparing the two mean tumor diameters, we have assumed that the design of this study was unpaired, meaning that the data were composed of two independent samples, one from each treatment group. In some experiments, for example, if one wishes to compare quality of life before and after any medical procedure is performed, a paired design is appropriate because the patient is being compared with him- or herself—that is, the patient serves as his or her own control. Here, one would subtract the value measured on an appropriate quality-of-life scale before the procedure from that measured on the same scale after the procedure to create a single set of before-to-after differences. Once this subtraction has been done for each patient, one in fact has reduced the two measures on each patient (i.e., before and after) to a single set of numbers representing the differences. Therefore, paired data can be analyzed using the same formulas as used for single-sample analyses. Paired designs are often more efficient than unpaired designs, as between-group variability is reduced by the pairing.

Assumptions behind the Z tests.—For ease of exposition, we have presented all of the test formulas using percentiles that came from the normal distribution, but in practice there are two assumptions behind this use of the normal distribution. The first assumption is that the data arise either from a normal distribution or the sample size is large enough for the central limit theorem [3] to apply. The second assumption is that the variance or variances involved in the calculations are known exactly.

The first of these assumptions is often satisfied at least approximately in practice, but the second assumption almost never holds in real applications. We usually have to use estimates s^2, s_1^2, and s_2^2 in the above formulas rather than the exact values σ^2, σ_1^2, and σ_2^2, respectively, because the variances would usually be estimated from the data rather than being known exactly. To account for the extra uncertainty due to the fact that the variance is estimated rather than known, the distribution of our test statistic changes. We thus use t distribution tables rather than normal distribution tables. In calculations, this means that the z values used in all of the formulas need to be switched to the corresponding values from t tables. This requires knowledge of the degrees of freedom (*df*), which for single-mean problems is simply the sample size minus 1. This of course applies to paired designs as well, because they reduce to single-sample problems. For two sample unpaired problems, a conservative number for the *df* is the minimum of the two sample sizes minus 1 ($n - 1$, where n is the sample size) [4]. These tests are called t tests.

Equal or unequal variances.—The tests described earlier assume that the variances in the two groups are unequal. Slightly more efficient formulas can be derived if the variances are the same, as a single pooled estimate of the variance can be derived from combining the information in both samples together. We do not discuss pooled variances further here, in part because in practice the difference in analyses done with pooled or unpooled variances is usually quite small and

TABLE 2 Tests and Confidence Intervals Required for One and Two Sample Problems for Continues Variables

1 or 2 sample	$\sigma_1 = \sigma_2$?	σ's known?	σ estimate	test	CI
1	N/A	Yes	N/A	$z = \frac{\overline{x}-x_0}{\sigma/\sqrt{n}}$	$\overline{x} \pm z\sigma/\sqrt{n}$
1	N/A	No	$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}}$	$t = \frac{\overline{x}-x_0}{s\sqrt{n}}$	$\overline{x} \pm ts/\sqrt{n}$
2	Yes	Yes	N/A	$z = \frac{(\overline{x}-\overline{y})-(x_0-y_0)}{\sqrt{\sigma^2(\frac{1}{n_1}+\frac{1}{n_2})}}$	$\overline{x}-\overline{y} \pm z\sqrt{\sigma^2(\frac{1}{n_1}+\frac{1}{n_2})}$
2	Yes	No	$s_1 = \sqrt{\frac{\sum_{i=1}^{n_1}(x_i-\overline{x})^2}{n_1-1}}$, $s_2 = \sqrt{\frac{\sum_{i=1}^{n_2}(y_i-\overline{y})^2}{n_2-1}}$ $s = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$	$t = \frac{(\overline{x}-\overline{y})-(x_0-y_0)}{\sqrt{s^2(\frac{1}{n_1}+\frac{1}{n_2})}}$	$\overline{x}-\overline{y} \pm t\sqrt{s^2(\frac{1}{n_1}+\frac{1}{n_2})}$
2	No	Yes	N/A	$\frac{(\overline{x}-\overline{y})-(x_0-y_0)}{\sqrt{(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2})}}$	$\overline{x}-\overline{y} \pm z\sqrt{(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2})}$
2	No	No	$s_1 = \sqrt{\frac{\sum_{i=1}^{n_1}(x_i-\overline{x})^2}{n_1-1}}$, $s_2 = \sqrt{\frac{\sum_{i=1}^{n_2}(y_i-\overline{y})^2}{n_2-1}}$	$t = \frac{(\overline{x}-\overline{y})-(x_0-y_0)}{\sqrt{(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2})}}$	$\overline{x}-\overline{y} \pm t\sqrt{(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2})}$

Note.—In all cases, the data are assumed to be normally distributed or the sample size large enough for the central limit theorem to apply. The data are assumed to be represented by x_i, $i = 1,\ldots, n$ for a single-sample problem or by x_i, $i = 1,\ldots, n_1$ and y_i, $i = 1,\ldots, n_2$ for a two-sample problem. Sample sizes are n for a single-sample problem and n_1 and n_2 for the two-sample problem. The z indicates a normal table is used, t indicates a t table is required. When a t table is required, the degrees of freedom are equal to $n-1$ for a single-sample problem, while the degrees of freedom are $n_1 + n_2 - 2$ for a two-sample problem with equal variances, and $\min(n_1-1, n_2-1)$ for unequal variances (conservative value). The x_0 and y_0 indicate null values under the null hypothesis (usually but not always equal to zero). For paired two-sample problems, form the within-individual differences, and use the formulas for the one-sample case. N/A = not applicable.

in part because it is rarely appropriate to pool the variances, because the variability is usually not exactly the same in both groups.

Analysis of variance: more than two means.—We have seen tests for one or two means, but sometimes one wishes to test the equality of three or more means. Although this topic is not covered here, readers should be aware that analysis of variance is a technique that extends our two-sample procedure to three or more means. See, for example, Armitage and Berry [5] or Rosner [6] for details.

Table 2 provides the test statistics used for all possible cases with one or two means, as discussed earlier.

How Useful Are p *Values for Medical Decision Making?*

Although p values are still often found in the literature, there are several major problems associated with their use. First, as mentioned earlier, they are often misinterpreted as the probability of the null hypothesis given the data, when in fact they are calculated assuming the null hypothesis to be true. Second, clinicians often use them to dichotomize results into important or unimportant depending on whether $p < 0.05$ or $p > 0.05$, respectively. However, there is not much difference between p values of 0.049 and 0.051, so the cutoff of 0.05 is arbitrary. Third, p values concentrate attention away from the magnitude of treatment differences. For example, one could have a p value that is very small but is associated with a clinically unimportant difference. This is especially prone to occur in cases in which the sample size is large. Conversely, results of potentially great clinical interest are not necessarily ruled out if $p > 0.05$, especially in studies with small sample sizes. Therefore, one should not confuse statistical significance (i.e., $p < 0.05$) with practical or clinical importance. Fourth, the null hypothesis is almost never exactly true. In the example described, does one seriously think that the mean tumor diameter of the patients on the standard treatment schedule could be exactly 3.5 cm (rather than, say, 3.50001 cm)? Because one knows the null hypothesis is almost surely false to begin with, it makes little sense to test it. Instead, one should concern oneself with the question, By how much are the two treatments different?

There are so many problems associated with p values that most statisticians now recommend against their use, in favor of confidence intervals or Bayesian methods. In fact, some prominent journals have virtually banished p values from publication [7], others strongly discourage their use [8], and many others have published articles and editorials encouraging the use of Bayesian methodology [9, 10]. We cover these more informative techniques for drawing statistical inferences, starting with confidence intervals.

Frequentist Confidence Intervals

Although the p value provides some information concerning the rarity of events as extreme as or more extreme than that observed assuming the null hypothesis to be exactly true, it provides no information about what the true parameter values might be. In the two-mean example described earlier, we observed a tumor diameter difference of 0.7 cm, which was shown to be "statistically significant," with a p value of approximately 0.000001. Although we observed a difference of 0.7 cm, we know that our data are from a random sample of patients to whom this procedure could be applied, so the true mean difference could in fact be higher or lower than our observed difference. How likely is it that the true mean difference in tumor diameter is clinically important?

One way to answer this question is with a confidence interval. The formula in equation 13 provides 95% confidence interval limits for means (the value 1.96 could be changed to other values if intervals with coverage other than 95% are of interest) [3]:

$$\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right) \quad (13)$$

where $\bar{x}$ is the sample mean and σ is the known SD from a sample of size n. As before, if σ is not known, it is replaced by its estimate from the data, s, and 1.96 is increased somewhat, as a percentile from the t distribution replaces the normal percentile.

Applying this formula to the single-mean example we first discussed, where $\bar{x} = 3.0$, $n = 10$, and $\sigma = 1.5$, we obtain a 95% confidence interval of (2.1–3.9 cm). We cannot conclude very much from this interval because we have not ruled out mean tumor diameters as small as 2.1 cm, which is clinically superior to the 3.5 cm from the old schedule; however, on the other hand, diameters as large as 3.9 cm have also not been ruled out, which is even worse than the tumor diameter in the standard group. Thus, further data would need to be collected before any conclusions could be drawn about this new schedule.

Our two-group clinical trial example had larger sizes, so it will presumably provide a more accurate estimate. We can calculate a 95% confidence interval for the difference in means for the two groups using the formula in equation 14,

$$\left(\bar{x}_1 - \bar{x}_2 - 1.96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},\ \bar{x}_1 - \bar{x}_2 + 1.96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) \quad (14)$$

where the same comment regarding unknown variances again applies. Plugging in the values we obtained from our clinical trial example given earlier, we find a confidence interval of –0.46 to –0.94 cm. Thus, roughly speaking, it is likely that the true tumor diameter difference between our two schedules is between approximately 0.5 cm less under the new schedule (–0.46 cm) and up to almost a 1-cm reduction (–0.94 cm). Although our *p* value for this same data set was small, which enabled us to reject the null hypothesis, we can see that the confidence interval provides more clinically useful information about the magnitude of the difference. We can also see that, in contrast to what may be believed after seeing the *p* value, we are still uncertain about the clinical utility of the new schedule, because values near the lower limit of the confidence interval would not be interesting clinically—it would represent less than a 30% change from the mean baseline tumor size—while differences near 1 cm may be clinically interesting. Therefore, our conclusions from the confidence interval are more detailed than those from the *p* value. This is true in general, as we now discuss.

Interpreting Confidence Intervals

Confidence intervals are derived from procedures that are set up to "work" 95% of the time (if a 95% confidence interval is used). The two confidence interval equations discussed earlier provide procedures that, when used repeatedly across different problems, will capture the true value of the mean (or difference in means) 95% of the time and fail to capture the true value 5% of the time. In this sense, we have confidence that the procedure works well in the long run, although in any single application, of course, the interval either does or does not contain the true mean. Note that we are careful not to say that our confidence interval has a 95% probability of containing the true parameter value. For example, we did not say that the true difference in mean tumor diameter is in the interval –0.49 to –0.94 cm with 95% probability. This is because the confidence limits and the true mean tumor diameters are both fixed numbers, and it makes no more sense to say that the true mean is in this interval than it does to say that the number 2 is inside the interval (1, 6) with probability 95%. Of course, 2 is inside this interval, just like the number 8 is outside of the interval (1, 6). However, the procedure used to calculate confidence intervals provides random upper and lower limits that depend on the data collected; in repeated uses of this formula across a range of problems, we expect the random limits to capture the true value 95% of the time and exclude the true mean 5% of the time. Refer to Figure 1. If we look at the set of confidence intervals as a whole, we see that about 95% of them include the true parameter value. However, if we pick out a single trial, it either contains the true value (≈ 95% of the time) or excludes this value (≈ 5% of the time).

Despite their somewhat unnatural interpretation, confidence intervals are generally preferred to *p* values. This is because confidence intervals focus attention on the range of values compatible with the data on a scale of direct clinical interest. Given a confidence interval, one can assess the clinical meaningfulness of the result, as can be seen in Figure 2.

Depending on where the upper and lower confidence interval limits fall in relation to the upper and lower limits of the region of clinical equivalence, different conclusions should be drawn. The region of clinical equivalence, sometimes called the region of clinical indifference, is the region inside of which two treatments, say, would be considered to be the same for all practical purposes. The point 0, indicating no difference in results between the two treatments, is usually included in the region of clinical equivalence, but values above and below 0 are usually also included. How wide this region is depends on each individual clinical situation. For example, if one treatment schedule is much more expensive than another, one may want at least a 50% reduction in tumor diameter to consider it the preferred treatment.

There are five different conclusions that can be made after a confidence interval has been calculated, as illustrated by the five hypothetic intervals displayed in Figure 2. The first conclusion (interval 1) is that the confidence interval includes zero and that both upper and lower confidence interval limits, if they were the true values, would not be clinically interesting. Therefore, this variable has been shown to have no important effect.

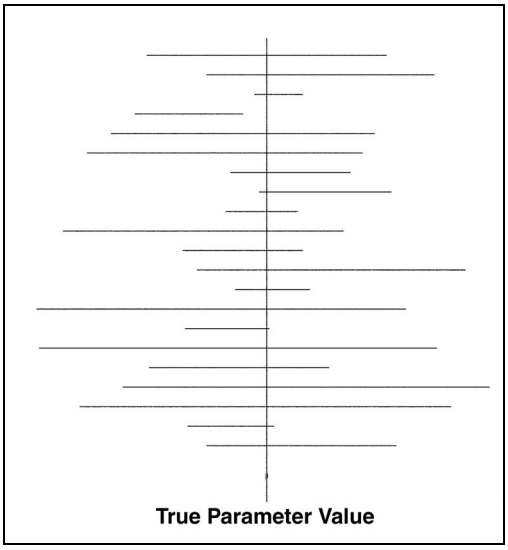

Fig. 1.—Drawing shows series of 95% confidence intervals for unknown parameter.

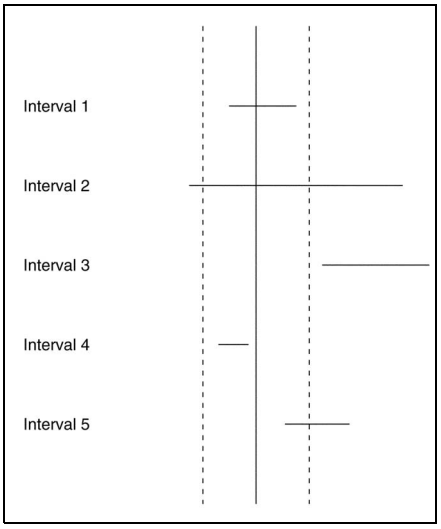

Fig. 2.—Drawing shows how to interpret confidence intervals. Depending on where confidence interval lies in relation to region of clinical equivalence, different conclusions can be drawn.

The second conclusion (interval 2) is that the confidence interval includes zero but that one or both of the upper or lower confidence interval limits, if they were the true values, would be interesting clinically. Therefore, the results of this variable in this study are inconclusive, and further evidence needs to be collected.

The third conclusion (interval 3) is that the confidence interval does not include zero and that all values inside the upper and lower confidence interval limits, if they were the true values, would be clinically interesting. Therefore, this study shows this variable to be important.

The fourth conclusion (interval 4) is that the confidence interval does not include zero but that all values inside the upper and lower confidence interval limits, if they were the true values, would not be clinically interesting. Therefore, this study shows this variable, although having some small effect, is not clinically important.

The fifth conclusion (interval 5) is that the confidence interval does not include zero but that only some of the values inside the upper and lower confidence interval limits, if they were the true values, would be clinically interesting. Therefore, this study shows this variable has at least a small effect and may be clinically important. Further study is required to better estimate the magnitude of this effect.

Revisiting the two confidence intervals discussed earlier in light of Figure 2, we see that the interval based on our single-sample experiment, which ranged from 2.1 to 3.9 cm, is clearly of type 2 and the interval based on the two-group clinical trial is of type 5. Once again, note that these confidence intervals provide much more detailed conclusions than the information contained in a *p* value.

The *p* values group together intervals 1 and 2 as "nonsignificant" and intervals 3, 4, and 5 as "significant." This can lead to misleading conclusions from a clinical viewpoint. For example, similar clinical conclusions should be drawn from intervals 1 and 4, even though one is "significant" and the other is not. It should now be clear why many journals discourage reporting results in terms of *p* values and encourage confidence intervals.

Summary of Frequentist Statistical Inference

The main tools for statistical inference from the frequentist point of view are *p* values and confidence intervals. The *p* values have fallen out of favor among statisticians, and although they continue to appear in medical journal articles, their use is likely to greatly diminish in the coming years. Confidence intervals provide more clinically useful information than *p* values, so confidence intervals are to be preferred in practice. Confidence intervals still do not allow the formal incorporation of preexisting knowledge into any final conclusions. For example, in some cases there may be compelling medical reasons why a new technique may be better than a standard technique, so if faced with an inconclusive confidence interval, a radiologist may still wish to switch to the new technique, at least until more data become available. On what basis could this decision be justified? We return to this question in the Bayesian Inference section, which appears later in this article.

Nonparametric Inference

Thus far, statistical inferences on populations have been made by assuming a mathematic model for the population (e.g., a normal distribution) and estimating parameters from that distribution based on a sample. Once the parameters have been estimated—for example, the mean or variance for a normal distribution—the distribution is fully specified. This is known as parametric inference.

Sometimes we may be unwilling to specify the general shape of the distribution in advance and prefer to base the inference only on the data, without a parametric model. In this case, we have distribution-free or nonparametric methods.

For example, consider the following data, which represent the tumor diameters of the marker liver metastases for two different chemotherapy regimens in patients with colorectal carcinoma: conventional treatment, 21, 12, 11, 28, 3, 10, 9, 5, 7, 10, 6; new treatment, 4, 3, 4, 5, 20, 22, 5, 12, 15, 5, 1, 14, 13.

Because we are making nonparametric inferences, we no longer refer to tests of similarity of group means. Rather, the null and alternative hypotheses here are defined as follows: For the null hypothesis (H_0), there is no treatment effect—that is, conventional treatment tends to give rise to tumor sizes similar to those from the new treatment. For the alternative hypothesis (H_A), the new treatment tends to give rise to different values for tumor sizes compared with those from the conventional treatment group.

The first step to nonparametrically test these hypotheses is to order and rank the data from lowest to highest values, keeping track of which data points belong to each treatment group, as shown in Table 3.

Thus, in ranking the data, we simply sort the data from the smallest to the largest value regardless of group membership and assign a rank to each data point depending on where its value lies in relation to other values in the data set. Hence, the lowest value receives a rank of 1, the second lowest a rank of 2, and so on. Because there are many "ties" in this data set, we need to rank the data accounting for the ties, which we do by grouping all tied values together and distributing the sum of the available ranks evenly among the tied values. For example, the second and third lowest values in this data set are both 3, and there is a total of five ranks (2 + 3) to be divided among them. Hence, each of these values receives a rank of 2.5 (5 / 2). Similarly, the sixth through ninth values are all tied at 5. There are 30 total ranks (6 + 7 + 8 + 9) to divide up among four tied values, so each receives a value of 7.5 (30 / 4), and so on.

The next step is to sum the ranks for the values belonging to the conventional treatment group, which yields a total of 147.5 (2.5 + 7.5 + 10 + 11 + 12 + 13.5 + 13.5 + 15 + 16.5 + 22 + 24).

We now reason as follows: There is a total of 300 ranks (1 + 2 + 3 + …+ 23 + 24) that can be distributed among the conventional and new treatment groups. If the sample sizes were equal, therefore, and if the null hypothesis were exactly true, we would expect that these ranks should divide equally among the two groups, so each would have a sum of ranks of 150. Now, the sample sizes are not

TABLE 3 First Step to Nonparametrically Test Null and Alternative Hypotheses: Order and Rank the Data

Treatment Group	N	N	C	N	N	N	N	N	C	C	C	C	C	C	C	C	N	N	N	N	N	C	N	C
Data	1	3	3	4	4	5	5	5	5	6	7	9	10	10	11	12	12	13	14	15	20	21	22	28
Ranks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Ranks with ties	1	2.5	2.5	4.5	4.5	7.5	7.5	7.5	7.5	10	11	12	13.5	13.5	15	16.5	16.5	18	19	20	21	22	23	24

Note.—N = new treatment, C = conventional treatment.

quite equal, so here we expect 300 × (11 / 24) = 137.5 of the ranks to go to the conventional group, and 300 × (13 / 24) = 162.5 of the ranks to go to the new treatment group. Note that 137.5 + 162.5 = 300, which is the total sum of ranks available. We have in fact observed a sum of ranks of 147.5 in the conventional group, which is higher than expected. Is it high enough that we can reject the null hypothesis? To answer this question, we must refer to computer programs that will calculate the probability of obtaining a sum of ranks of 147.5 or greater given that the null hypothesis of no treatment difference is true (remember the definition of the *p* value discussed earlier). Most statistical computer packages will perform this calculation, which in this case gives $p = 0.58$. Hence, the null hypothesis cannot be rejected, because our result and those more extreme are not rare under the null hypothesis.

This nonparametric test is called the Wilcoxon's rank sum test. An exactly equivalent test can be based on counts rather than ranks, and it is called the Mann-Whitney test. The Mann-Whitney test always provides the same *p* value as the Wilcoxon's rank sum test, so either can be used. The analogous parametric test, the unpaired *t* test for the same data, also gives a *p* value of 0.58, so the same conclusion is reached.

Because the two tests do not always provide the same conclusions, which of these tests is to be preferred? The answer is situation-specific. Remember that the *t* test assumes either that the data are from a normal distribution—here, it would imply that the tumor diameters are approximately normally distributed—or that the sample size is large. A histogram would show that the data are skewed toward the right, so that normality is unlikely, and the sample sizes are 11 and 13, hardly large. Hence, in this example the nonparametric test is preferred because the assumptions behind the *t* test do not seem to hold. In general, if the assumptions required by a parametric test may not hold, a nonparametric test is to be preferred, whereas if the distributional assumptions do likely hold, a parametric test provides slightly increased power compared with a nonparametric test.

The Wilcoxon's rank sum test is appropriate for unpaired designs. A similar test exists for paired designs, called the Wilcoxon's signed rank test. Nonparametric confidence intervals are also available, as are tests for two or more groups, such as the Kruskal-Wallis test. See Sprent [11] for further details about these methods.

Sample Size Calculations

As previously discussed, there has been a strong trend away from hypothesis testing and *p* values toward the use of confidence intervals in the reporting of results from biomedical research. Because the design phase of a study should be in sync with the analysis that will eventually be performed, sample size calculations should be performed on the basis of ensuring adequate numbers for accurate estimation of important quantities that will be estimated in the study, rather than by power calculations. This distinction is important because it has been shown [12] that sample sizes calculated from a power viewpoint are often insufficient when viewed from a confidence interval viewpoint. In other words, although high power ensures rejection of the null hypothesis with high probability, it does not ensure than the confidence interval will be narrow enough to allow good clinical decision making. Therefore, in this section, we focus on sample size methods based on confidence interval width. For similar methods based on power, see the book by Lemeshow et al. [13].

The question of how accurate is "accurate enough" can be addressed by carefully considering the results you would expect to get (a bit of a catch-22 situation, because if you knew the results you will get, there would be no need to perform the experiment) and making sure your interval will be small enough to land in intervals numbered 1, 3, or 4 of Figure 2. The determination of an appropriate width is a nontrivial exercise, but a reasonable target confidence interval width can usually be found.

For estimating the sample size requirements in experiments involving population means, two different formulas are available, depending on whether there is a single sample or two samples. These are derived by solving for the sample size *n* in the formulas for the confidence intervals discussed.

Single Sample

Let μ be the mean that is to be estimated, and assume that we wish to estimate μ to an accuracy of a total confidence interval width of w (so that the confidence interval will be $\bar{x} \pm d$, where $2 \times d = w$). Let σ be the SD in the population.

Then the required sample size, n, is given by equation 15,

$$n = \frac{z^2\sigma^2}{d^2} = \frac{4 \times z^2\sigma^2}{w^2} \qquad (15)$$

where, as usual, z is replaced by the appropriate normal distribution quantile (z = 1.96, 1.64, or 2.58 for 95%, 90%, or 99% intervals, respectively).

For example, suppose that we would like to estimate average tumor size to an accuracy of $d = 2$ mm with a 95% confidence interval and that we expect the patient-to-patient variability will be $\sigma = 10$ mm. Then, from the previous formula, we need to perform the calculation in equation 16,

$$n = \frac{1.96^2 \times 10^2}{2^2} = 96 \qquad (16)$$

rounding up to the next highest integer. The most difficult problem in using this equation is to decide on a value for the SD σ, because it is usually unknown. A conservative approach would be to use the maximum value of σ that seems reasonably likely to occur in the experiment.

Two Samples

Let μ_1 and μ_2 be the means of two populations, and suppose that we would like an accurate estimate of $\mu_1 - \mu_2$. Again assume a total confidence interval width of w (so that again $2 \times d = w$). Let σ_1 and σ_2 be the SD in each population, respectively.

Then the sample size is given in equation 17,

$$n = \frac{z^2(\sigma_1^2 + \sigma_2^2)}{d^2} = \frac{4 \times z^2(\sigma_1^2 + \sigma_2^2)}{w^2} \quad (17)$$

where now n represents the required sample size for each group. As usual, z is chosen as we did earlier and is usually 1.96, corresponding to a 95% confidence interval.

Bayesian Inference

Consider again the single-sample tumor diameter problem introduced in the Statistical Inferences for Means section. Recall that in this example patients undergoing the standard radiation therapy schedule are assumed to have a mean of 3.5 cm, whereas the data collected so far for the new accelerated schedule indicate a mean of 3.0 cm, but are based on only 10 subjects. The frequentist confidence interval was wide, ranging from approximately 2.1 to 3.9 cm, so it has not been particularly helpful in making a decision about which technique to use for the next patient. At this point, with the data being relatively uninformative, the treating physician may decide to be conservative and remain with the standard schedule until more information becomes available about the new schedule or may go with their "gut feeling" as to the likelihood that the new schedule is truly better or not. If there have been data from animal experiments or strong theoretic reasons why the new schedule may be better, there may be temptation to try the new one. Can anything be done to aid in this decision-making process?

Bayesian analysis has several advantages over the standard or frequentist statistical analyses discussed in this article so far, including the ability to formally incorporate relevant information not directly contained in the current data set into any statistical analysis. We will see how this can help with the problem discussed earlier, but first we will cover some basics of Bayesian analysis.

Let us generically denote our parameter of interest by θ. Hence, θ can be a binomial parameter, the mean from a normal distribution, an odds ratio, a set of regression coefficients, and so on. Note in particular that θ can be two or more dimensional. The parameter of interest is sometimes usefully thought of as the "true state of nature."

The three basic elements of any Bayesian analysis are, first, the prior probability distribution, $f(\theta)$. This prior distribution summarizes what is known about θ before the experiment is performed. It is based on a "subjective" assessment of the available past information, so may vary from investigator to investigator.

The second basic element of Bayesian analysis is the likelihood function: $f(x \mid \theta)$. The likelihood function summarizes the information contained in the data, x. For instance, it may be created from a normal distribution for a mean. It is important to realize that Bayesians and frequentists can use the same likelihood function because both need to calculate the probability of data given various values for the parameter θ. The way the likelihood function is used, however, differs between the two paradigms.

The third basic element is the posterior distribution: $f(\theta \mid x)$. The posterior distribution summarizes the information in the data, x, together with the information in the prior distribution. Thus, it summarizes what is known about the parameter of interest θ after the data are collected.

Bayes' theorem, posthumously published by Thomas Bayes [14] in 1763, relates the three quantities: posterior distribution = [likelihood of the data × prior distribution] / a normalizing constant, or using our notation above in equation 18,

$$f(\theta|x) = \frac{f(x|\theta) \times f(\theta)}{\text{a normalizing constant}} \quad (18)$$

or, omitting the normalizing constant in equation 19,

$$f(\theta|x) \propto f(x|\theta) \times f(\theta) \quad (19)$$

where $\propto$ indicates "is proportional to."

Thus, we update the prior distribution to a posterior distribution after seeing the data via Bayes' theorem. The current posterior distribution can be used as a prior distribution for the next study, so Bayesian inference provides a natural way to represent the learning that occurs as science progresses.

Radiologists are already familiar with the Bayesian way of thinking, using it every day in the context of interpreting diagnostic tests. The prior probability used in Bayes' theorem is analogous to the background rate of a condition in the population, which is updated to a positive or negative predictive value (analogous to a posterior distribution) after seeing the results of a diagnostic test (analogous to seeing the data). It is thus just a short step from using predictive values in a clinical setting to using Bayes' theorem in a research setting.

The most contentious element in Bayesian analysis is the need to specify a prior distribution. Because there is no unique way to derive prior distributions, they are necessarily subjective, in the sense that one radiologist may derive a different prior distribution than another and, hence, arrive at a different posterior distribution. Several points can be made regarding this controversy.

First, Bayesians can use diffuse, flat, or reference prior distributions that, for all practical purposes, consider all values in the feasible range as equally likely. Hence, if little prior information exists or if a Bayesian wishes to see what information the data themselves provide, this choice of prior distribution can be used. In fact, in many situations, a Bayesian analysis using reference priors will result in similar interval estimates as those provided by frequentist confidence intervals, but with a more natural interpretation: Unlike confidence intervals, Bayesian intervals (often called credible intervals) can be directly interpreted as containing the true parameter value with the indicated probability. Thus, no references to long runs of other trials are necessary to properly interpret a credible interval.

Second, although many frequentists have been quick to criticize Bayesian analysis because of the difficulty in deriving prior distributions, frequentist analysis formally ignores this information, which can hardly be considered as a better solution.

Third, if different clinicians have a range of prior opinions and hence a range of prior distributions, there will also be a range of posterior distributions. Presenting several Bayesian analyses matching this range of prior opinions helps to raise the level of debate after the publication of results in medical journals, because it accurately reflects the range of clinical opinion that exists in the community. Furthermore, it can be shown that as more data accumulate, the posterior distributions from different priors tend to converge toward a single distribution, accurately mirroring the process of eventual consensus among clinicians as data accumulate. When viewed in this light, prior distributions can be seen as a great advantage. See Spiegelhalter et al. [15] or a more introductory level article [9] for more information on using a range of prior distributions when carrying out a Bayesian analysis.

Having discussed the basic elements, let us see how Bayesian analysis works in practice by again considering our example of tumor diameters after radiation for brain cancer. We will discuss the three elements that lead to the posterior

distribution calculated from Bayes' theorem, which are listed in the previous section.

Recall that in our data set we had $\bar{x} = 3.0$, $\sigma = 1.5$, and $n = 10$, so that our likelihood function is a normal distribution with mean 3.5 and SE of 0.474, the same as was used in the frequentist inferences discussed previously. In general, the choice of prior distribution is based on any information that is available at the time of the experiment. We will consider two different prior distributions. The first (prior distribution 1 in Fig. 3) will be a normal distribution with a mean of 3.5 cm and a very large variance, say, 10,000. This is a noninformative prior, because all values in the likely range have an approximately equal chance of being the true value, the curve being quite flat over a wide range. Note that an equal 50% chance is given to both the null and alternative hypotheses that the new schedule is superior to that of the old, because the distribution is centered at 3.5 cm. The second prior distribution (prior distribution 2 in Fig. 3) will be centered at 3.0, with an SD of 0.5 (variance of 0.25). This would represent the opinion of a radiologist who is enthusiastic about the new schedule, with a prior opinion that the new mean tumor diameter will be between about 2.0 and 4.0 cm, with 95% probability (as calculated from the range of the normal [μ = 3.0, $\tau^2 = 0.25$] distribution, where τ^2 is our prior variance). Do not be confused by the two distinct SDs that are used here: σ represents the variability of the tumor diameters among the patients, whereas τ represents how certain we are of our prior mean value.

We now wish to combine this prior density with the information in the data as represented by the likelihood function to derive the posterior distribution, using Bayes' theorem. After some algebra, the posterior distribution can be shown to be given by a normal distribution shown in equation 20,

$$N\left(A \times \mu + B \times \overline{x}, \frac{\tau^2\sigma^2}{n\tau^2 + \sigma^2}\right) \qquad (20)$$

where $A = [(\sigma^2 / n) / (\tau^2 + \sigma^2 / n)]$ and $B = [(\tau^2) / (\tau^2 + \sigma^2 / n)]$. Note that the posterior mean value depends on both the prior mean, μ, and the observed mean in the data set, $\bar{x}$. Plugging these values into the previous equation and using the first (very flat) prior distribution, we find that the posterior distribution for our mean tumor diameter is N ($A \times \mu + B \times \bar{x}$ = 3.0, $[(\tau^2\sigma^2) / (n\tau^2 + \sigma^2)] = 0.225$). For the second more informative prior, the corresponding posterior distribution is N (3.0, 0.118).

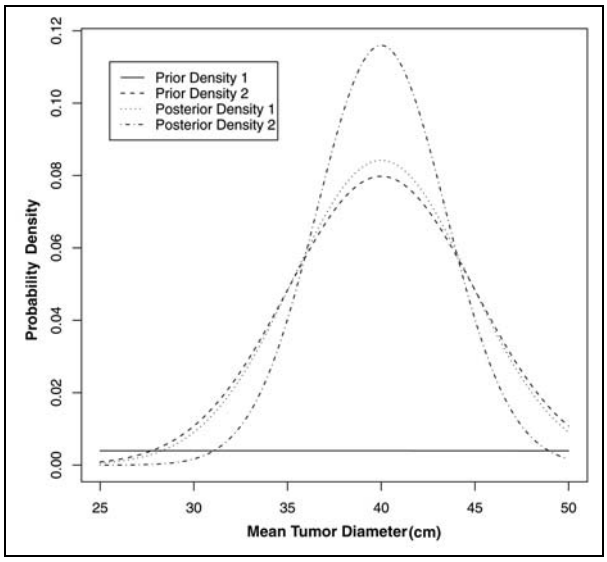

Fig. 3.—Graph shows two prior and corresponding posterior densities for tumor diameter example.

The two prior and two posterior densities are displayed in Figure 3. Note that the second posterior distribution is narrower, because a stronger prior distribution was used. These posterior distributions can be used to derive 95% credible intervals and to test hypothesis from a Bayesian viewpoint. These calculations can be done using normal tables. Because these posterior distributions directly represent the probability distribution for our unknown parameter, interpretation of these quantities is straightforward.

For example, a 95% credible interval from posterior distribution 1 is given by (2.1–3.9). In comparing this interval to the prior 95% confidence interval calculated in the Statistical Inferences for Means section, we see that they are numerically identical (at least to one decimal place). However, the interpretations of these two intervals are different because the Bayesian credible interval is directly interpreted as the probability that the true mean tumor diameter lies in the given interval, given the data and the prior information used. This is in contrast to the less direct interpretation of a confidence interval, discussed earlier. Many people misinterpret confidence intervals as if they were Bayesian intervals. This error is often not too serious, because if little prior information is available, the two intervals are numerically similar. Therefore, even though it is technically incorrect, one does not go too far wrong thinking of confidence intervals as approximate Bayesian intervals, when there is little prior information. A 95% credible interval from our second posterior distribution is given by (2.3–3.7), which is somewhat narrower than the first interval.

We can also perform Bayesian hypothesis tests, again just using the posterior distributions. For example, suppose we wish to test H_0 ($\mu \geq 3.5$) versus H_A: ($\mu < 3.5$). We can calculate $\Pr\{H_0 \mid \text{data}\} = \Pr\{\mu \geq 3.5 \mid \text{data}\}$, which is equal to 14.5% for posterior 1 and 7.3% for posterior 2. Thus, we are approximately 85.5% or 92.7% sure that the tumor diameter under the accelerated schedule is better than the standard schedule, depending on which prior we use. Based on this, each clinician can make a decision about which schedule to apply to the next patient. Note again the very direct statements available for Bayesian hypothesis tests, compared with the nonintuitive interpretation of a p value. This clarity, however, comes at the expense of having to specify a prior distribution.

Carrying out Bayesian analyses is made easier via the use of freely available customized software. The posterior distributions shown earlier were performed using the First Bayes package [16], and more complex Bayesian analyses can be done via specialized Monte Carlo numeric routines implemented in WinBUGS software [17] made freely available by the Medical Research Council of Great Britain [18]. An excellent introductory text on Bayesian analysis is one written by Gelman et al. [19].

Conclusions

This module has introduced some of the major ideas behind statistical inference, with em-

phasis on the simple methods for continuous variables. Rather than a simple catalogue listing of which tests to use for which types of data, we have tried to explain the logic behind the common statistical procedures seen in the medical literature, the correct way to interpret the results, and what their advantages and drawbacks may be. We have also introduced Bayesian inference as a strong alternative to standard frequentist statistical methods, both for its ability to incorporate the available prior information into the analysis and for its ability to address questions of direct clinical interest.

The next few modules in this series will cover techniques suitable for other types of data, including proportions and regression methods.

References

1. Oshiro Y, Kusumoto M, Matsuno Y, et al. CT findings of surgically resected large cell neuroendocrine carcinoma of the lung in 38 patients. *AJR* 2004;182:87–91
2. Karlik SJ. Visualizing radiologic data. *AJR* 2003;180:607–619
3. Joseph L, Reinhold C. Fundamentals of clinical research for radiologists: introduction to probability theory and sampling distributions. *AJR* 2003; 180:917–923
4. Moore D, McCabe G. *Introduction to the practice of statistics,* 3rd ed. New York, NY: Freeman and Company, 1988
5. Armitage P, Berry G. *Statistical methods in medical research,* 3rd ed. Oxford, England: Blackwell Scientific Publications, 1994
6. Rosner B. *Fundamentals of biostatistics*. Belmont, MA: Duxbury, 1995
7. Rothman K. Writing for epidemiology. *Epidemiology* 1998;9:333–337
8. Evans S, Mills P, Dawson J. The end of the p-value. *Br Heart J* 1988;60:177–180
9. Brophy J, Joseph L. Placing trials in context using Bayesian analysis: GUSTO revisited by Reverend Bayes. *JAMA* 1995;273:871–875
10. Lilford R, Braunholz D. The statistical basis of public policy: a paradigm shift is overdue. *BMJ* 1996;313:603–607
11. Sprent P. *Applied nonparametric statistical methods*. New York, NY: Chapman and Hall, 1989
12. Bristol D. Sample sizes for constructing confidence intervals and testing hypotheses. *Stat Med* 1989;8:803–811
13. Lemeshow S, Hosmer D, Klar J, Lwanga S. *Adequacy of sample size in health studies*. Chichester, England: Wiley, 1990
14. Bayes T. An essay towards solving a problem in the doctrine of chances: 1763. *Philos Trans R Soc* 1763;53:370–418
15. Spiegelhalter D, Freedman L, Parmar M. Bayesian approaches to randomized trials. *J R Stat Soc [Ser A]* 1994;157:387–416
16. O'Hagan A. First Bayes software. Available at: www.shef.ac.uk/st1ao/1b.html. Accessed December 25, 2003
17. Spiegelhalter D, Thomas A, Best N. *WinBUGS version 1.4 user manual.* Cambridge, England: MRC Biostatistics Unit, 2003
18. WinBUGS, version 1.4. Available at: www.mrc-bsu.cam.ac.uk/bugs/. Accessed December 25, 2003
19. Gelman A, Carlin J, Stern H, Rubin D. *Bayesian data analysis*, 2nd ed. London, England: Chapman and Hall, 2003

The reader's attention is directed to earlier articles in the Fundamentals of Clinical Research series:

1. Introduction, which appeared in February 2001
2. The Research Framework, April 2001
3. Protocol, June 2001
4. Data Collection, October 2001
5. Population and Sample, November 2001
6. Statistically Engineering the Study for Success, July 2002
7. Screening for Preclinical Disease: Test and Disease Characteristics, October 2002
8. Exploring and Summarizing Radiologic Data, January 2003
9. Visualizing Radiologic Data, March 2003
10. Introduction to Probability Theory and Sampling Distributions, April 2003
11. Observational Studies in Radiology, November 2004
12. Randomized Controlled Trials, December 2004
13. Clinical Evaluation of Diagnostic Tests, January 2005
14. ROC Analysis, February 2005