Chapter 6
Statistical Inference

In the last chapter, we developed probability theory and introduced the concept of mass and density functions, which can be used to handle large and/or complex sample spaces in which events have unequal probabilities of occurrence. These algebraic functions involve parameters and events (random variables), and if you know the parameters, you can deduce (compute) the probabilities for particular events—values of the random variable. Statistical inference involves reversing this process. When we take a sample from the population, we have a collection of events, but we usually do not know the values of the parameters that produced the observed sample data. The Central Limit Theorem plays an important role in helping us use sample statistics, like the sample mean and variance, to estimate population parameters, and especially to quantify uncertainty in our estimates.

6.1 The Central Limit Theorem and Inferential Statistics

In the previous chapter, we discussed how, if you assume the distribution for a variable in the population is normal, then we can use a z table to compute probabilities of obtaining a sample member with a value on the variable within any given range. The first step in the process of conducting statistical inference is to extend this idea of determining probabilities of obtaining a single sample member within a particular range of values on the variable to determining probabilities of obtaining an entire *sample* of a given size with a given value of a *statistic* like the mean from the population. For example, if we know the population mean, μ, is 10, we might want to know the probability of obtaining a sample of size $n = 100$ that has a mean, $\bar{x}$, of 8 or less.

In order to understand this process, we must discuss properties of statistics (like the mean) that can be drawn from a population. One of the most important theorems in statistics, the Central Limit Theorem (CLT), states that, as sample sizes increase, regardless of the distribution of a random variable in the population, sample means ($\bar{x}$'s) that can be drawn from that population distribution become normally

S.M. Lynch, *Using Statistics in Social Research: A Concise Approach*,
DOI 10.1007/978-1-4614-8573-5_6, © Springer Science+Business Media New York 2013

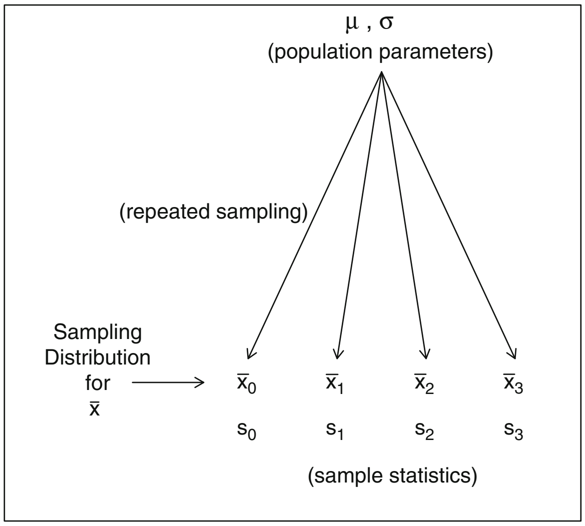

Fig. 6.1. Depiction of repeated sampling from the population and the concept of a sampling distribution.

distributed with a mean equal to the population mean (μ) and a variance equal to the variance in the population divided by the sample size used to compute the mean (σ^2/n). Equivalently, the standard deviation of the distribution of sample means is equal to $\sigma/\sqrt{n}$. The theorem's key result can be expressed in an abbreviated form as:

$$\sqrt{n}(\bar{x} - \mu) \xrightarrow{d} N(0, \sigma^2). \tag{6.1}$$

That is, the difference between sample means and the population mean, multiplied by the square root of the sample size, tends in distribution toward a normal distribution with a mean of 0 and a variance of σ^2. Another way to express the same idea is:

$$\bar{x} \overset{asy}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right). \tag{6.2}$$

In this form, "asy" means asymptotically. That is, the distribution of sample means becomes more normally distributed as the sample size increases.

This theorem is conceptually difficult but forms the basis for statistical inference, and so some extended explanation is warranted. When you take a sample from a population, instead of thinking of this sample as a collection of individuals, consider that you are selecting a *sample mean* from the population. Figure 6.1 shows this process graphically. The population parameters μ and σ (or σ^2) govern the samples that we can draw from the population. If we take repeated samples (of a given size n) from the population and compute sample statistics, like $\bar{x}$, for each one, the CLT

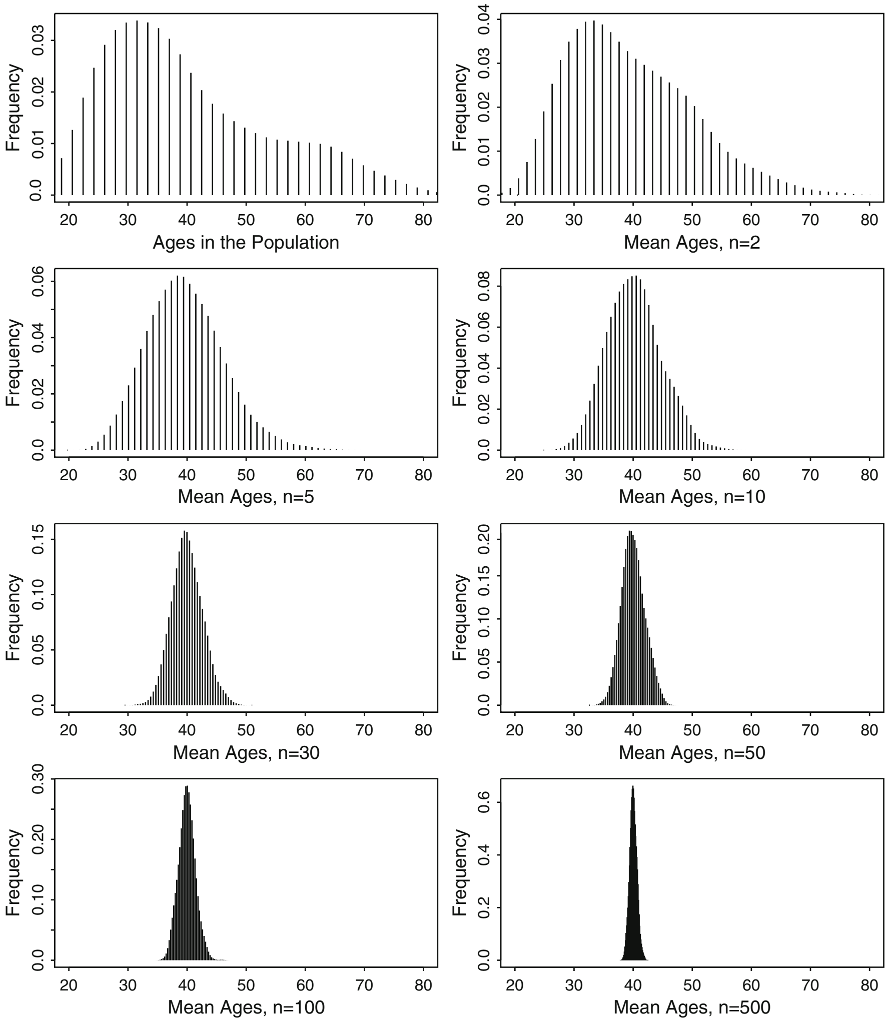

Fig. 6.2. Histograms of sampling distributions of the mean for various sample sizes

says that the distribution of these sample means—called the sampling distribution of the mean—will be normal with a mean of μ and a standard deviation of $\sigma/\sqrt{n}$, so long as n is "large."

In order to make these ideas concrete, I treated data on age from the 2004 General Social Survey as if it were a population distribution of ages, and from it I drew 1,000 random samples each of sizes n = 2, n = 5, n = 10, n = 30, n = 50, n = 100, and n = 500. For each of the samples, I computed the mean. So, for samples of size n = 2, I computed 1,000 means, for samples of size n = 5, I computed 1,000 means, etc. Figure 6.2 shows histograms of these collections of means by sample size.

Sample Size	$\mu_{\bar{x}}$	Observed s.e. ($\sigma_{\bar{x}}$)	Theoretical s.e. ($\sigma/\sqrt{n}$)
(Actual Population)	$\boxed{40.0}$	14.0 (σ)	14.0
2	39.4	10.2	9.9
5	39.7	6.3	6.3
10	40.2	4.5	4.4
30	$\boxed{40.0}$	2.5	2.6
50	40.0	1.9	2.0
100	40.0	$\boxed{1.4}$	$\boxed{1.4}$
500	40.0	.6	.6

Table 6.1. Results of simulation examining distributions of sample means

The upper left plot is the histogram of the original distribution of age. The upper right plot is the histogram of 1,000 sample means—called the sampling distribution—for the samples of size $n = 2$. The sampling distribution for the mean when $n = 2$ looks much like the original population distribution: it is skewed strongly to the right. However, as the figure shows, as the sample size increases, the sampling distributions become much more symmetric (normal), and their variance decreases. Ultimately, when the sample size is $n = 500$, the sampling distribution is very narrow.

Table 6.1 shows the means of the sampling distributions ($\mu_{\bar{x}}$), as well as the observed standard deviations of these sampling distributions ($\sigma_{\bar{x}}$; called the "standard error"—abbreviated s.e.). The far right column shows the theoretical standard error based on the CLT ($\sigma/\sqrt{n}$). Notice that the means of the sampling distributions are very close to the mean of age in the population, and that, when the sample size is 30, the mean of the sampling distribution is consistently within rounding of the true mean. Also notice that the observed s.e. consistently matches the theoretical s.e. ($\sigma/\sqrt{n}$) when the sample size is 100 or larger.

Why do the means of the sampling distribution become normal as the sample size increases, and why does the variance of the sampling distribution decrease? If you recall from the last chapter, the joint probability of two independent events is the product of their respective probabilities. When we take a simple random sample from the population, we are taking a collection of independent draws from the population distribution. Thus, their joint probability can be computed as the product of each of their probabilities. When the sample size is small, it may not be uncommon to draw a few extremely rare observations—observations that are far from the population mean—and thus obtain a sample mean that is far from the population mean. For example, the probability of obtaining two individuals whose values are rare enough that their probability of occurrence is .1 each is $.1 \times .1 = .01$. This is a small probability, but not incredibly small. However, it is extremely unlikely, if we draw a large sample, that we would draw a series of very rare values. For example, the probability of drawing five rare people like the ones we just discussed is $.1 \times .1 \times .1 \times .1 \times .1 = .00001$. This is an incredibly small probability. The implication is that it will be unlikely in a large sample to draw a large number of very rare individuals, and so our sample mean will tend to be close to the true mean.

The implication of the CLT for making inference about population means using sample means is that, if the sample size is large enough, we can very accurately and precisely estimate the population mean, and we can quantify our uncertainty in our estimate—that is, we can state how far away from the true population mean (μ) our estimate ($\bar{x}$) is likely to be.

6.2 Hypothesis Testing Using the z Distribution

Given that we know that the sampling distribution for $\bar{x}$ is normal, and we know its standard deviation (more on whether we "know" this later), we can construct a standardized score to determine the probability of obtaining a sample mean in some range from a sample of size n, given a particular true population mean:

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}. \tag{6.3}$$

Thus, if we knew the population mean were, say, 50, and the population standard deviation were 20, we could determine the probability of obtaining a sample of size 100 with a mean of $\bar{x} = 54$ or greater. The z-score would be:

$$z = \frac{54 - 50}{20/\sqrt{100}} = 2. \tag{6.4}$$

The corresponding probability of obtaining a sample with a mean of 54 or greater, then, would be roughly .025. In other words, it would be somewhat unlikely to obtain such a sample.

We can construct a similar z score if we are interested in computing the probability of obtaining a particular range for a proportion. For example, suppose we know that the proportion of persons in the population who support a particular policy is .6 (60 %), What is the probability of obtaining a sample of 500 persons in which less than 50 % of the respondents support the policy? In this situation, we simply replace μ with p_0 (the population proportion), $\bar{x}$ with $\hat{p}$ (the sample proportion), and σ with $\sqrt{p_0(1 - p_0)}$ (the standard deviation of a proportion):

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} \tag{6.5}$$

$$= \frac{.5 - .6}{\sqrt{.6(.4)/500}} \tag{6.6}$$

$$= -4.56 \tag{6.7}$$

Thus, the probability of obtaining a sample of 500 persons with a proportion of .5 or smaller when the population proportion is .6 is $p(z < -4.56) \approx 0$.

Formal classical hypothesis testing utilizes this approach, but instead of using a known value for μ (which, in practice, we never have!), we substitute H_0 for μ into the calculation, where H_0 is called the "null hypothesis" we are interested in testing—it is a hypothesized value for the true population mean, μ. We then compute the probability of obtaining a sample mean as extreme as, or more than, our observed sample mean under this hypothesis. This probability is called a "p-value." If a sample mean is highly improbable to occur (i.e., p is small) under H_0, we reject H_0 and conclude that H_0 is probably not the true value of μ. It is common to say, when rejecting a null hypothesis, that the difference between the (null) hypothesized value for μ and the one potentially implied by the sample data, is "statistically significant" at the value at which p is considered to be small enough to reject the null. It is essential to recognize that the p-value *does not tell us the probability that H_0 is true.* It tells us the probability of obtaining our sample mean under the assumption that H_0 is true. In other words, hypothesis testing follows a "modus tollens" structure:

1. If H_0 is true, then $\bar{x}$ will be close to H_0 (its probability of occurrence will be high).
2. $\bar{x}$ is not close to H_0 (p is small).
3. Therefore, H_0 is not true (rejected).

A few questions/comments about this process are in order. First, how improbable must a sample mean be under the null hypothesis for us to reject the hypothesis? Second, what does this process tell us about the true value of the population mean?

Regarding the first question, there are two types of errors that can be made when following this hypothesis testing strategy: Type I and Type II errors. A Type I error is committed when we reject a null hypothesis that happens to be true. A Type II error is committed when we fail to reject a null hypothesis that is false. In practice, scientists are generally conservative and therefore most concerned with rejecting null hypotheses that are in fact true. Why? Suppose we were examining whether some experimental drug treatment is efficacious. The null hypothesis would be that there is no difference between the treatment and control groups on the outcome of interest. Rejecting this null would lend support to the view that the drug works, and we would want to be extremely confident that the treatment actually had an effect before marketing the drug, especially if it had negative side effects. Similarly, in social science, we want to be certain that socioeconomic, racial, sex, or some other difference in some outcome of interest in fact exists before we begin making policy recommendations that cost taxpayers money or cause some conflict.

The probability of making a Type I error is represented as α, and we generally set $\alpha = .05$ (called the "critical alpha") as our acceptable probability for making such an error in social research. This means that, when we conduct our hypothesis test, we want there to be a probability of less than .05 that we would observe a sample mean as extreme as (or more so than) ours if the null hypothesis were true (i.e., we want $p < \alpha = .05$). We generally split this probability across the two tails of the normal distribution, and so we typically will not reject a null hypothesis unless the z score that produces the p-value is greater than 1.96 (or less than -1.96). Thus, in the example above, if 50 were our hypothesized population mean, we determined that the probability of obtaining a sample with a mean at least as extreme as we had was .05. Extremeness is measured on *both* sides of the distribution, and so, $p(z > 1.96) + p(z < -1.96) < .05$; alternatively: $p(|z| > 1.96) < .05$. We call such an approach a "two-tailed test." We would therefore reject the null hypothesis and conclude that the population mean is probably not 50. The data simply aren't very consistent with that null hypothesis.

If we decided to set α much lower, it is possible that we might fail to reject a null hypothesis that would be rejected if α had been .05. In that case, the result would not be statistically significant at the chosen α level. Thus, we must keep in mind that statistical significance is somewhat subjective—it depends on the value chosen for α as much as it depends on the data. Furthermore, the fact that some result is declared to be statistically significant does not necessarily indicate that the result is substantively important. Consider the denominator of the z calculation: it depends on n. The larger the sample size, the larger z will be, even if the difference between $\bar{x}$ and H_0 remains constant in the numerator. As the CLT simulation showed, with a large enough sample size, it becomes extremely difficult to obtain a sample mean very different from the true population mean. Thus, with large n, even substantively trivial differences between $\bar{x}$ and H_0 may be declared statistically significant. For example, suppose our null hypothesis is that 80 % of people in the world like their statistics class (i.e., $p_0 = .8$). We obtain a sample of $n = 5,000$ persons who have taken statistics courses and find the 82 % liked their course. What would we conclude about the null hypothesis?

$$z = \frac{.82 - .8}{\sqrt{.8(.2)/5000}} = 3.53. \tag{6.8}$$

$p(|z| > 3.53) \approx 0$, and so we would reject the null hypothesis. But is the difference between our observed 82 % and our hypothesized 80 % worth discussing? Clearly, the vast majority of students like their statistics class.

This brings us to the second question: What does the process of hypothesis testing tell us about the true population mean? Unfortunately, the answer is: by itself, not much. It only tells us what the true population mean probably is not; it does not give us any indication what it is. The sample mean itself gives us an indication of this, however, and we will discuss this issue later in the chapter in the context of confidence intervals.

6.3 Hypothesis Testing When σ Is Unknown: The t Distribution

The discussion of hypothesis testing in the previous section assumed that the true population standard deviation (σ) is known. However, it is unreasonable to expect, if we don't and can't know μ, that we could possibly know σ. Fortunately, just as the CLT justifies our belief that $\bar{x}$ in large samples is very close to μ, the CLT also justifies the belief that s (the sample standard deviation) is close to σ in large samples. Thus, when σ is unknown, we may consider using s as our estimate of σ in making our standardized score calculations. However, given that there is uncertainty in the representation of σ with s (i.e., we don't in fact know how close s is to σ), the standardized score is no longer normally distributed. Instead, it is t distributed, and the associated score is thus no longer called a z score—it is called a t score:

$$t = \frac{\bar{x} - H_0}{s/\sqrt{n}} \tag{6.9}$$

The t distribution looks similar to the normal distribution. It is symmetric around its mean, median, and mode and is bell-shaped. However, it has fatter tails that reflect our uncertainty in using s as an estimate of σ. The probability density function for the t distribution is more complicated than the normal distribution, and it is unnecessary to discuss it. However, it is important to know what parameters are associated with the distribution. The t distribution has a mean and variance, just like the normal distribution, but it also has a "degrees of freedom" parameter (d.f.). For our purposes, the degrees of freedom associated with the distribution is $n-1$, where n is the sample size. As $n - 1$ gets larger, the distribution becomes more and more normal in appearance, reflecting the fact that a large sample size reduces uncertainty associated with using s to approximate σ. When d.f. > 120, the t and standard normal distributions are virtually indistinguishable. Figure 6.3 shows the t distribution with different degrees of freedom. Notice that, as the degrees of freedom increases, the tails of the t distribution flatten, and the distribution becomes more peaked in the center like the normal distribution. At 120 degrees of freedom, the figure shows that the t distribution cannot be differentiated from the z distribution.

6.4 Confidence Intervals

The process of hypothesis testing allows us to decide whether we think a hypothesized value for the population mean is reasonable, but it does not allow us to directly make inference about the true value of a population mean. In other words, the logic of hypothesis testing is that we computed the probability of observing the sample mean we did under some hypothesized value for the population mean. If

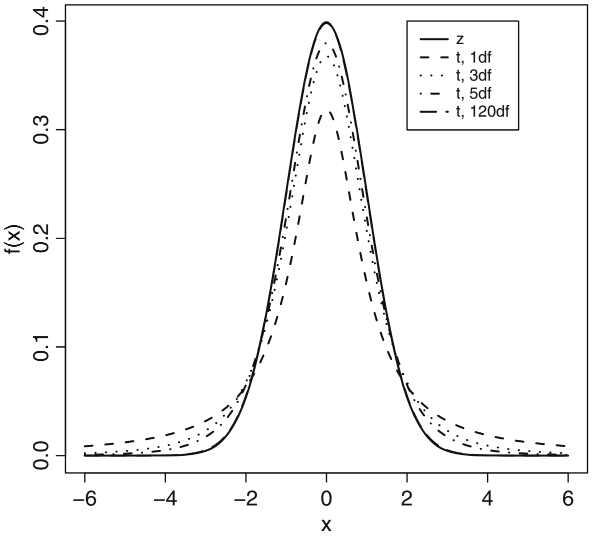

Fig. 6.3. Some t distributions

that probability (the p-value) is low, then we reject the hypothesized value. This approach does nothing to help us know, if the hypothesized value is rejected, what the true value of the population mean is. Confidence interval construction can, however, help us make inference about the population mean.

Confidence intervals are reported regularly in the popular media in discussions of both science and politics. During election seasons in particular, news agencies routinely report poll results describing preferences for particular candidates. For example, we might hear something like: "56 % of respondents favor candidate A over candidate B, with a margin of error of ±3 %". In this example, the polling agency is claiming that the true proportion of the population that favors candidate A is between 53 and 59 %. This result is a basic confidence interval.

The logic of confidence interval construction is quite simple. If we let the sample mean be our best guess for the population mean, and we let the sample standard deviation be our best guess for the population standard deviation, then, based on the CLT, we can construct an interval around our sample mean using the sample standard deviation and make some statement about where the true population mean is expected to fall.

Constructing confidence intervals follows the following steps:

1. Decide on a "confidence level." Just as in hypothesis testing, we generally establish $\alpha = .05$, which corresponds to a confidence level of .95.
2. Find $t_{\alpha/2}$. That is, find the value of t that leaves half of α in each tail of the t distribution.
3. Compute the standard error of the mean, which is $\frac{s}{\sqrt{n}}$
4. Construct the interval estimate as: $\bar{x} \pm (t_{\alpha/2} \times s/\sqrt{n})$

In the first step, we decide how confident we wish to be with respect to whether our estimated interval captures the population mean. Here we have to make a trade-off between precision and accuracy: the more precise we are, the less confident, and hence the less likely to be accurate, our estimate will be. If we increase our confidence (make $(1-\alpha)$ large), we will necessarily have to increase our interval width. Conversely, if we decrease our confidence, we can narrow the width of the interval (it will then be less likely to capture the mean). Put in the extreme: we can be 100 % confident that the population mean falls in the interval $(-\infty\ ,\ \infty)$, or we can be 0 % confident that the population mean is exactly some number M. As a matter of convention, it is common in social science to choose $\alpha = .05$, which corresponds to a confidence level of 95 %.

In the second step, we find the appropriate t value that corresponds to the two-tailed probability $\alpha/2$ we are interested in. For example, if we want to be 95 % confident, then $\alpha = .05$, and we need to find $t_{.025}$—the value of t for which .025 % of the mass of the t distribution falls beyond it. Recall that the t distribution has a degrees of freedom parameter associated with it: if we have d.f. > 120, then the appropriate t value for this example would be 1.96 (same as the z). Note that, if σ is known, t can be replaced with z.

In the third step, we compute the standard error of the mean. This value, $s/\sqrt{n}$, derives from the CLT as we discussed before. It is a measure of the extent to which sample means derived from samples of a specific size (n) can be expected to vary from the true population mean. The factor $(t_{\alpha/2} \times s/\sqrt{n})$ is the margin of error.

Finally, we construct the interval estimate as indicated. One way to understand this interval estimate, and to relate it back to hypothesis testing, is to view it as a rearrangement, in a sense, of the t test we developed in the last section:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}. \tag{6.10}$$

If we multiply both sides by the denominator, we get:

$$t \times \frac{s}{\sqrt{n}} = \bar{x} - \mu. \tag{6.11}$$

If we then isolate μ:

$$\mu = \bar{x} - t \times \frac{s}{\sqrt{n}}. \tag{6.12}$$

This result looks remarkably like the interval estimate presented in step 4 above, with a couple of differences. First, the t statistic is computed under the hypothesis testing approach, but under the interval estimate approach it is fixed at a given value reflecting the confidence level (hence t is replaced with $t_{\alpha/2}$). Second, the interval estimate approach constructs an interval around $\bar{x}$, rather than simply a one-sided estimate (hence the "–" is replaced by "±"). Recall, however, that, when we obtain

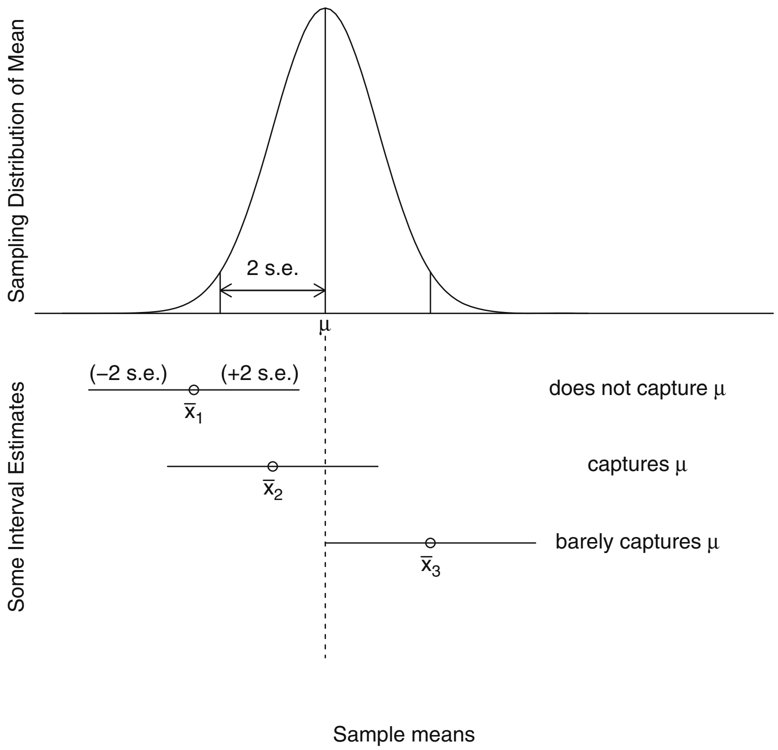

Fig. 6.4. Illustration of concept of "confidence." Plot shows that 95 % of the sample means that can be drawn from their sampling distribution will contain μ within an interval ranging 2 standard errors around $\bar{x}$.

a p-value (or choose an α and find a critical t), we explicitly assume that our t statistic could be positive or negative, which implicitly corresponds to using ± in the numerator of the equation for t.

The first difference between the hypothesis testing and interval construction approaches highlights the fundamental difference between hypothesis tests and intervals. Under hypothesis testing, we determine how probable a particular sample mean is under a hypothesized value of μ. Under the interval estimate approach, we predetermine the probability that intervals of a given width will contain the true value of μ.

How do we interpret confidence intervals? Ultimately, what we can say about our interval estimate is that, if we took repeated samples from the population and constructed $(1-\alpha)\%$ intervals around all of our sample estimates, $(1-\alpha)\%$ of such intervals would contain the true value of the parameter. Figure 6.4 illustrates this idea. The top half of the figure shows the sampling distribution for means $\bar{x}$ that can be drawn from a population with mean μ. Under the CLT, we know that this sampling distribution is normal, and so 95 % of all sample means (of a given size,

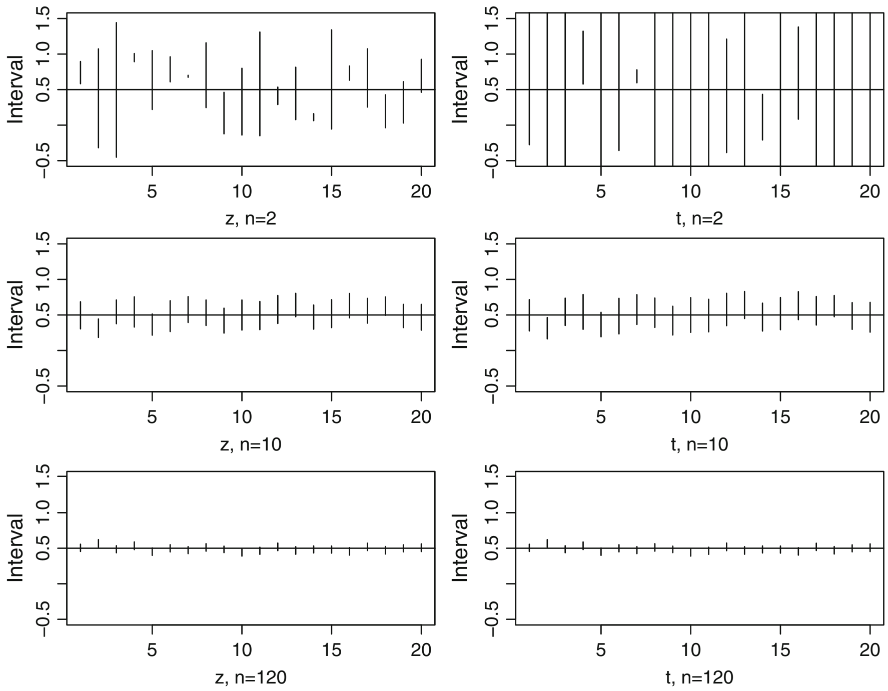

Fig. 6.5. Some confidence intervals from samples of different sizes (based on the z and t distributions)

n) that could be drawn from the population will fall within two standard errors of μ. Thus, if we take sample means and attach (add and subtract) two standard errors to them, 95 % of such intervals will contain μ. The bottom half of the figure shows three such intervals. In the first one, $\bar{x}_1$ is more than two standard errors away from μ, and so its interval does not contain μ. In the latter two, the sample means are within two standard errors of μ, and so, when we attach two standard errors to the sample estimates, the intervals contain μ.

Note that we cannot say, under this approach to constructing intervals, that μ falls within the interval with probability $1-\alpha$. We can only say that $(1-\alpha)\%$ of all intervals of a given width constructed in this fashion will contain μ. μ is considered fixed; only the intervals are random; they are the result of random sampling from the population. In order to demonstrate this idea, Fig. 6.5 shows some confidence intervals that were constructed around sample means which were drawn from a population for which the true mean was .5. In constructing the intervals, I used both the z distribution (critical z of 1.96) and the t distribution to show how uncertainty in using s as an estimate of σ diminishes as n increases. The plots in the left column of the figure are based on the z distribution; the plots in the right column are based on the t distribution.

Sample Size	z intervals	t intervals
n = 2	.67	.94
n = 10	.93	.95
n = 120	.94	.94

Table 6.2. Proportion of "95 %" confidence interval estimates capturing the true mean by type of interval (z/t) and sample size.

Notice that (1) not all of the intervals actually capture the true mean, (2) the proportion of intervals that do capture the true mean appears to increase as the sample sizes increase (ultimately, the proportion is limited to 95 %, because they are intervals based on $\alpha = .05$), and (3) the widths of the intervals in the smallest sized samples fluctuate drastically. This latter issue exemplifies why z-based intervals are not appropriate when the population standard deviation is unknown and the sample size is small: a large proportion of these intervals do not contain the population mean. However, when the intervals are based on the t distribution, they are wider and more capture μ.

Table 6.2 presents the proportions of confidence intervals (out of 1,000) that capture the true population mean for both z and t based samples of different sizes. Notice how the 95 % intervals based on the t distribution are almost always truly 95 % intervals; that is, that 95 % of them tend to capture the true mean, regardless of the sample size.

6.5 Additional Hypothesis Testing Tools

So far we have only discussed hypothesis testing and confidence interval construction for a single sample mean. Although this type of hypothesis test and confidence interval is important, we generally are more interested in comparing sample means—to test hypotheses about differences between groups—in actual social research. For example, we may be interested in determining whether men and women earn comparable incomes on average. Or, we may want to know whether whites and nonwhites vote similarly in an election. Or, we may want to know whether persons with a high school diploma are more likely to be depressed than persons with a college diploma. Or, we may want to know whether the proportion of persons with a high school diploma who are unemployed is comparable to the proportion of persons with a college diploma who are unemployed. These cases require extensions of our basic one-sample t test, and a number of extensions are available.

Although the tests we will discuss appear different from the one-sample test, they all derive from the basic one-sample t test and the basic logic of hypothesis testing. Suppose we were interested in the first question—whether men and women earn comparable incomes on average. We could hypothesize some difference between men and women in the population and construct the following t test based on the observed difference in sample means:

$$t = \frac{(\bar{x}_{men} - \bar{x}_{women}) - H_0 : (\mu_{men} - \mu_{women})}{S.E.(\bar{x}_{men} - \bar{x}_{women})}. \tag{6.13}$$

The test is possible because, under the CLT, both $\bar{x}_{men}$ and $\bar{x}_{women}$ are normally distributed, and the difference between two normally distributed variables is also normally distributed. What is the standard error here? This is the standard error of the difference between men and women. Recall that before, the denominator of the t statistic was $s/\sqrt{n}$. Here, we need an alternate measure of the sample standard error, because we are examining the difference between two groups. There are actually at least two ways to construct this standard error,[1] but the most common, and usually the most appropriate, uses a basic rule of variance algebra: the variance of the difference of two random variables is equal to the sum of the variances of the two random variables minus twice the covariance of the two random variables:

$$Var(A \pm B) = Var(A) + Var(B) \pm 2 \times Cov(A, B). \tag{6.14}$$

We will discuss the covariance later, but for now, note that two variables that are independent do not covary (i.e., they are uncorrelated). Because our sample members are selected independently, the mean for men and the mean for women are independent, and so the latter term involving the covariance is 0. Thus, we can compute the standard error of the difference between the means for men and women as:

$$S.E.(\text{difference}) = \sqrt{\frac{s^2_{men}}{n_{men}} + \frac{s^2_{women}}{n_{women}}}. \tag{6.15}$$

If we also recognize that, if we hypothesize a difference to be 0, then the latter two terms in the numerator disappear ($\mu_{men} - \mu_{women} = 0$), so our entire test reduces to:

$$t = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}. \tag{6.16}$$

[1] A common and alternate calculation is $s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$, where $s_p^2 = \frac{\sum(x_1-\bar{x}_1)^2 + \sum(x_2-\bar{x}_2)^2}{n_1+n_2-2}$. This is called the pooled variance estimate. From a theoretical standpoint, s_p^2 is appropriate, if, members of the two groups come from the same distribution. However, we cannot really know this until *after* the test. Thus, I prefer the measure used in the text above, which assumes the samples are from different populations and makes the t-test more conservative (the non-pooled estimate is always at least as large as the pooled estimate). In the case the two groups are from the same overall population, the calculations are equivalent; if they aren't, the non-pooled estimate presented in the text is correct.

Here, I have simply replaced the "men" and "women" labels with 1 and 2 for generality. This t test is called the "independent samples t-test." The degrees of freedom associated with the test is tedious to compute and is as follows:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}} \tag{6.17}$$

In most social science settings, the subsample sizes will be well over 120, and so the z distribution can be used as an approximation and the degrees of freedom calculation is not needed. Furthermore, when the subsample sizes do not exceed 120, it is usually the case that $df \geq \min(n_1-1, n_2-1)$. Thus, we can set the degrees of freedom equal to the smaller of the subsample sizes -1 for a conservative result.

As a simple example, suppose the mean income for men in a sample was found to be \$53,735, while the mean income for women was \$45,624. The standard deviations were \$32,508 and \$31,887, respectively, and the subsample sizes were $n = 12{,}038$ men and $n = 14{,}190$ women. If we construct the t statistic, we find:

$$t = \frac{53735 - 45624}{\sqrt{\frac{(32508)^2}{12038} + \frac{(31887)^2}{14190}}} = 20.31. \tag{6.18}$$

Thus, under the null hypothesis that the difference in incomes between men and women is 0, the probability of obtaining a sample in which the means were this different is $p(|t| \geq 20.31) \equiv p(t > 20.31) \times 2 \approx 0$. In other words, if mean income for men and women were in fact equal, we would practically never see this type of difference in a sample as large as we have.

Sometimes we may be interested in comparing groups, but in terms of comparing differences in proportions, rather than a continuous variable like income. In that case, we can simply replace our sample means with our sample proportions, and replace our sample standard errors with the standard errors for proportions, and the test proceeds as before. As we saw in Chap. 5, the variance of a proportion, p is $p(1-p)$. Thus,

$$t = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}. \tag{6.19}$$

Here, I have used the symbol $\hat{p}$ to indicate that the sample proportion is an estimate of the population proportion p. I use this notation to emphasize that the variance of the proportions to be used is based on the sample proportions.

A final test that we may sometimes need is for examining pairs of individuals (or individuals' scores from two points in time), rather than comparing the means of independent samples. For example, suppose we want to determine whether some experimental treatment reduces depression, and suppose our data are individual depression measures taken before and after treatment. In this case, we do not have independent samples, because the two samples are really the same sample, just measured at different times. In that case, we simply construct the differences pairwise and compute the standard error of these differences. Thus, ultimately this test is a one-sample test on differences, rather than an independent samples test.

Table 6.3 summarizes all of these various z and t tests and confidence intervals we have discussed, in addition to a few that we have not. For example, we did not discuss the one sample z test for proportions, but it is a direct extension of the original one sample z test discussed earlier, just applied to proportions. It is not a t test, because, if p_0 (the population proportion) is hypothesized or known, then the population variance is known.

6.6 Conclusions

In this chapter, we discussed the Central Limit Theorem in some detail and developed the process of hypothesis testing and confidence interval construction from it. Hypothesis testing reveals the probability of obtaining the observed data under some hypothesis, allowing us to reject (or fail to reject) scientific hypotheses. However, it does not allow us to make inference about the true population parameter of interest. Confidence intervals do. A number of z and t tests and confidence interval strategies follow from the CLT and our original z test, and we will develop still more as we move to later chapters and topics.

6.7 Items for Review

- Central Limit Theorem
- Sampling distribution
- Null hypothesis and hypothesis testing
- p-value
- Type I and Type II errors
- α
- Critical value (for t or α)
- t distribution
- Degrees of freedom
- Margin of error
- Confidence interval
- Pooled and unpooled variance estimates
- Various statistical tests as shown in Table 6.3

Name	Formula
One-sample z test	$\frac{\bar{x}-H_0}{\sigma/\sqrt{n}}$
One-sample t test	$\frac{\bar{x}-H_0}{s/\sqrt{n}}$ df=$n-1$
Independent samples t test	$\frac{(\bar{x}_1-\bar{x}_2)-H_{\mu_1-\mu_2}}{\sqrt{(s_1^2/n_1)+(s_2^2/n_2)}}$ df=min(n_1-1, n_2-1)
One sample z test for proportions	$\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}$
Independent samples t test for proportions	$\frac{(\hat{p}_1-\hat{p}_2)-H_{p_1-p_2}}{\sqrt{(\hat{p}_1(1-\hat{p}_1)/n_1)+(\hat{p}_2(1-\hat{p}_2)/n_2)}}$ df=min(n_1-1, n_2-1)
Paired Sample t test	$\frac{\bar{x}_{\text{diff}}-H_0}{s_{\text{diff}}/\sqrt{n}}$ df=$n-1$, where n is number of pairs
$(1-\alpha)$% Confidence interval for a mean	$\bar{x} \pm (t_{\alpha/2})\,(s/\sqrt{n})$ df=$n-1$
$(1-\alpha)$% Confidence interval for a proportion	$\hat{p} \pm (t_{\alpha/2})\left(\sqrt{\hat{p}(1-\hat{p})/n}\right)$ df=$n-1$
$(1-\alpha)$% Conf. Int. for difference in means	$(\bar{x}_1-\bar{x}_2) \pm (t_{\alpha/2})\left(\sqrt{s_1^2/n_1+s_2^2/n_2}\right)$ df=min(n_1-1, n_2-1)
$(1-\alpha)$% Conf. Int. for difference in proportions	$(\hat{p}_1-\hat{p}_2) \pm (t_{\alpha/2})\left(\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right)$ df=min(n_1-1, n_2-1)

Table 6.3. Various hypothesis testing and confidence interval formulas following from the CLT.

6.8 Homework

1. Suppose years of schooling in the population has a mean of 12 and a standard deviation of 3. What is the probability of obtaining a sample of 50 people with a mean of 11 or more?
2. What is the probability that I could obtain a sample of 100 people in which the mean years of schooling was less than 12.1?

3. What is the probability of obtaining a sample of 500 people with a mean less than 11.8 or greater than 12.2?
4. If 52 % of the population supports a particular political candidate, what is the probability of obtaining a sample of 200 people in which less than 50 % of respondents support him/her?
5. I believe that systolic blood pressure averages 120 mmHg in the population with a standard deviation of 10 mmHg. If I randomly sampled 50 people and found that mean systolic blood pressure was 130, what could/would I say about my assumption that the population mean is 120?
6. If a sample mean for years of schooling was 13.58 with a standard deviation of 2.66 years (n = 2,386), and you wanted to be 95 % confident in estimating the population mean education level, what would you say (i.e., what would your interval estimate be)?
7. I've been told that one-third of the population is in excellent health. In the GSS, 32.4 % of the sample ($n = 26{,}228$) claims excellent health. Is what I've been told reasonable?
8. Does income predict happiness? In a subsample from the GSS, mean happiness for those with higher than average income is 1.32 (s = .59, n = 11,366), while mean happiness for those with lower than average income is 1.11 (s = .64, n = 14,862).
9. Construct 95 % confidence intervals for happiness by income group using the data in the previous question.
10. Do whites and blacks experience different levels of life satisfaction? In the GSS, a mean satisfaction scale score for whites is 23.03 (s = 4.52, n = 10,791) and for blacks is 20.55 (s = 5.07, n = 1,269).
11. I've heard that real incomes have been stagnant since 1973. In the GSS, mean income in 1973 was $53,641 (s = $31,213, n = 1,106). In 2006 (the last data available, say), mean income was $54,291 (s = $40,511, n = 1,420). Are the data consistent with that claim?
12. In a sample of 10 men, 70 % claim to like country music. In a sample of 8 women, 65 % of claim to like country music. Do men's and women's musical tastes differ?
13. Suppose I think that the country is, on average, politically moderate. In 2006, the GSS shows a mean party affiliation score of 3.32 (s = 1.71, n = 1,420). The range of the score is 1–6 (strong democrat…strong republican), so that the "middle" of the scale is 3.5. Do the data support this view that the country is moderate on average?
14. Some argue that the US population is aging. Is there any evidence of this in the GSS? In the GSS, the mean age of the sample was 43.89 in 1972 (s = 16.88, n = 1,086) and 46.10 in 2006 (s = 16.85, n = 1,420).
15. My theory says that men marry women with comparable levels of intelligence: that, on average, neither husbands nor wives are smarter than their partners. Assume I measure intelligence with IQ for 10 couples. What can I conclude about my hypothesis?

Couple	Husband's IQ	Wife's IQ
1	100	110
2	140	135
3	90	120
4	150	120
5	140	130
6	95	110
7	100	95
8	50	50
9	200	200
10	100	95

16. Some say that women are discriminated against in the labor market, that is, they do not receive comparable incomes for the same level of education as men. In a GSS subsample for men and women with comparable levels of education, mean male income was \$47,828 with a standard deviation of \$30,031. Mean female income was \$41,354 with a standard deviation of \$26,333. There were 277 men in the sample and 351 women in the sample. Is there evidence that men's and women's incomes at this education level differ?
17. Some political scientists argue that women and men differ in their political views. In a GSS subsample, among women, 487 claimed to be liberal, 448 claimed to be moderate, and 366 claimed to be conservative. Among men, 332 claimed to be liberal, 416 claimed to be moderate, and 337 claimed to be conservative. Is there any evidence that political views differ between sexes?
18. Exit polling of 400 voters in a district found that 55 % had voted for candidate A over candidate B. Given this information, would you call the election for candidate A?
19. My neighbor estimates that mean family income in the US population is \$50,000. In our GSS sample of 2,386 persons, mean family income was \$51,120.18 with a standard deviation of \$32,045.84. If we assume that the GSS sample is random, do you think my neighbor has a reasonable estimate?
20. I want to be sure to obtain a margin of error of ±2 % in a poll asking whether people support the government restricting the purchase of extra large carbonated beverages. How large should my sample be?
21. Based on anecdotal evidence (i.e., nonrandom, small samples), some have hypothesized that IQ has actually improved among recent birth cohorts. When a particular IQ test was validated in the 1950s, the mean was 100, with a standard deviation of 15. Suppose I hypothesize that IQ now has a mean of 110 (but with a standard deviation still of 15). I take a random sample of 20 persons and find $\bar{x} = 105$. What can I say about my hypothesis?

22. Now suppose I am unwilling to assume a standard deviation of 15, so I use my sample standard deviation of 12 instead. Reconduct the hypothesis test appropriately.
23. Explain why we may be more concerned about committing type I errors rather than type II errors.
24. It is a longstanding view that men tend to inflate the number of sexual partners they've had, while women tend to deflate their number. Part of the rationale for this view is that men need to appear "macho." I hypothesize that, by age 50, however, married men and women shouldn't differ in the number of sexual partners claimed. According to the GSS, the mean number of sexual partners claimed (in the last year) by 1886 married men over age 50 was 1.14 (sd = 2.58), while the mean number of sexual partners claimed by 1,750 married women over age 50 was .95 (sd = .44). Is my hypothesis reasonable?
25. Arguably, for a married person one would expect that the mean number of sexual partners reported would be 1. In the GSS, of 3,636 persons over age 50, the mean number of partners reported was 1.049 (sd = 1.89). Is it reasonable that the mean number of partners is 1 in this population?
26. Some argue that republican presidencies are worse for the economic condition of families in the U.S. than are democratic presidencies; others argue the opposite. In the GSS, 22.56 % of 30,979 respondents reported being in worse financial condition than the year before during GSS survey years in republican presidencies, while 20.42 % of 17,681 respondents reported being in worse financial condition than the year before during democratic presidencies. Is either party better, at least based on these data?
27. The term "bleeding heart liberal" is due to the (supposed) excess sympathy that liberals feel for criminals, the poor, the sick, and other marginalized groups. At the same time, however, members of marginalized groups are also more likely to be liberal, arguably because they themselves have experienced circumstances that make them more empathetic to the plight of others. If this hypothesis is true, we might expect that self-proclaimed liberals would have higher mean scores on an index capturing the number of "bad things" that have happened to them in the last year. In the GSS, of 56 persons claiming to be "strong conservatives," the mean score on such an index was 3.75 (sd = 4.15), while, of 46 persons claiming to be "strong liberals," the mean score on the index was 4.4 (sd = 2.30). Is the hypothesis reasonable?
28. An exit poll of 300 persons showed that candidate A received 53 % of the vote in a given precinct. If you were the pollster, would you call the election for candidate A based on these results?
29. In some states, if there is a difference of less than one percentage point between two candidates' final vote tallies, a recount of the cast ballots is automatically performed. In such a state, suppose that an exit poll of 200 people found 103 votes for one candidate and 97 for the other. Do you suspect there will be a recount?

30. I want to be sure to obtain a margin of error of ±5 % in a poll asking whether people support universal background checks for purchasing guns. How large should my sample be?
31. A recent survey of 1,000 people found that 90 % of Americans support background checks for purchasing guns. Construct a 95 % confidence interval for the population proportion.
32. A group of 10 persons participated in a weight loss study evaluating the efficacy of a new drug. Based on the data below, would you conclude that the drug worked?

Person	Before	After
1	300	280
2	280	285
3	425	360
4	315	330
5	255	249
6	600	590
7	290	300
8	265	235
9	310	290
10	230	250

33. Construct a 95 % confidence interval for the weight change in the previous problem.
34. In the 2010 GSS survey, 492 persons out of a total of 1,030 said that we spend too little on law enforcement in the US. I believe the country is evenly split on the issue. Am I right?
35. In the 2008 GSS, those who said they were raised as religious fundamentalists had a mean educational attainment of 13.09 years (sd = 2.93; n = 516), while those who were not raised as fundamentalists had a mean educational attainment of 13.64 years (sd = 3.28; n = 1,083). Is there a difference in educational attainment between fundamentalists and nonfundamentalists in the population?
36. A strip of the south has been called the "Bible Belt." Is there evidence that the region is distinct from other regions, in terms of the religiosity of its people? In the 2008 GSS, 275 out of 591 southerners claimed to have been raised as fundamentalists, while 242 out of 1,012 nonsoutherners were raised as fundamentalists.
37. Suppose I'm a firm believer that no statement is so outrageous that you can't get at least 10 % of the population to believe it. So, I conduct a poll of 200 people (randomly chosen) and ask whether they believe the following statement: "The earth was constructed by invisible, purple, one-legged unicorns all named Bill exactly 5,000 years ago this Thursday at noon." In my sample, 32 people said they had no doubt this was true. What would you conclude about my hypothesis? (assume, for the present purposes, that my statement is the most outrageous possible).

38. If 50 % of the population supports a national health care plan, what is the probability of obtaining a sample of 500 persons in which less than 45 % do?
39. Some say that support for the death penalty is declining. Using GSS data, I find that, in 2010, 1,297 respondents favored the death penalty, while 627 opposed it. In 1980, 982 respondents favored it, and 390 opposed it. Construct a 95 % confidence interval for the difference in support across these two time periods. Based on the evidence, would you agree that support has declined?
40. Capture-recapture methods are statistical methods that can be used to estimate the size of an unknown population. For example, suppose I am interested in estimating the total number, N, of fish in a pond. First, I catch a number of fish (say n_1 fish), tag them, and then release them (this is the "capture" step). Next, I go fishing again and catch n_2 fish (this is the "recapture" step). I observe that x of them are tagged. Using these data I can estimate the total number of fish in the pond as follows:

$$\frac{n_1}{N} = \frac{x}{n_2} \tag{6.20}$$

$$N = \frac{(n_1)(n_2)}{x} \tag{6.21}$$

In words, when I am finished with the capture step, the proportion of the total number of fish in the pond that are tagged is $n_1/N = p_0$. I can estimate this proportion via the recapture process: $x/n_2 = \hat{p}$. Since, on average, we would expect $\hat{p}$ to equal the population proportion p_0 (under the CLT), we can set these two proportions equal and solve for the only unknown, N.

Now, if I were to engage in a second recapture process, I would most likely obtain a different $\hat{p}$, and so it is easy to see that $\hat{p}$ can be viewed as the key random quantity here. N is constant, and both n_1 and n_2 can also be held constant. So, if we were interested in comparing the number of fish in two ponds, then, we really can just compare $\hat{p}$ across the two ponds.

So, suppose I run a catfish farm, and a fish food company claims that their food is not only cheaper than that of their competitors, but it also improves the life expectancy of fish, and keeps them healthier so that they can escape from predators (e.g., predatory birds). I decide to test their claim in a year-long experiment. I have two identical holding ponds into which I randomly distributed 1,000 hatchlings each. For pond A, I used my usual feed; for pond B, I used the new feed. A year later, I decide to estimate the total number of fish in each pond, with the hypothesis being that, if the fish food company's claim is true, pond B should have more fish remaining than pond A. I spend a day fishing each pond, capturing 100 fish in each pond, tagging them, and releasing them. A few days later, I spend a half day fishing. In pond A, I catch 48 fish, 6 of which are tagged. In pond B, I catch 62 fish, 7 of which are tagged. Using the methods that have been covered in this chapter, and the logic of capture-recapture methods as described above, what would you conclude about the fish food company's claim?

In reality, capture-recapture methods are a little more complicated than this in practice, and we need to make several assumptions to justify their use. What do you think some of those assumptions are, and what are some problems that you see with the method as described here? In other words, think critically about the method like a statistician might.