Confidence Intervals

I. Interval estimation.

The particular value chosen as most likely for a population parameter is called the point estimate. Because of sampling error, we know the point estimate probably is not identical to the population parameter.

The accuracy of a point estimator depends on the characteristics of the sampling distribution of that estimator. If, for example, the sampling distribution is approximately normal, then with high probability (about .95) the point estimate falls within 2 standard errors of the parameter.

Because the point estimate is unlikely to be exactly correct, we usually specify a range of values in which the population parameter is likely to be. For example, when $\overline{X}$ is normally distributed, the range of values between $\overline{X} \pm 1.96\sigma_{\overline{X}}$ is called the 95% confidence interval for μ. The two boundaries of the interval, $\overline{X} - 1.96\sigma_{\overline{X}}$ and $\overline{X} + 1.96\sigma_{\overline{X}}$ are called the 95% confidence limits. That is, there is a 95% chance that the following statement will we true:

$$\overline{X} - 1.96\sigma_{\overline{X}} \leq \mu \leq \overline{X} + 1.96\sigma_{\overline{X}}$$

Similarly, when $\overline{X}$ is normally distributed, the 99% confidence interval for the mean is

$$\overline{X} - 2.58\sigma_{\overline{X}} \leq \mu \leq \overline{X} + 2.58\sigma_{\overline{X}}$$

The 99% confidence interval is larger than the 95% confidence interval, and thus is more likely to include the true mean.

α = the probability a confidence interval will not include the population parameter, 1 - α = the probability the population parameter will be in the interval. The 100(1 - α)% confidence interval will include the true value of the population parameter with probability 1 - α, i.e., if α = .05, the probability is about .95 that the 95% confidence interval will include the true population parameter. On the other hand, 2.5% of the time the highest value in the confidence interval will be smaller than the true value, while 2.5% of the time the smallest value in the confidence interval will be greater than the true value.

Be sure to note that the population parameter is not a random variable. Rather, the probability statement is made about samples. If we drew 100 samples of the same size, we would get 100 different sample means and 100 different confidence intervals. We expect that in 95 of those samples the population parameter will lie within the estimated 95% confidence interval, in the other 5 the 95% confidence interval will not include the true value of the population parameter.

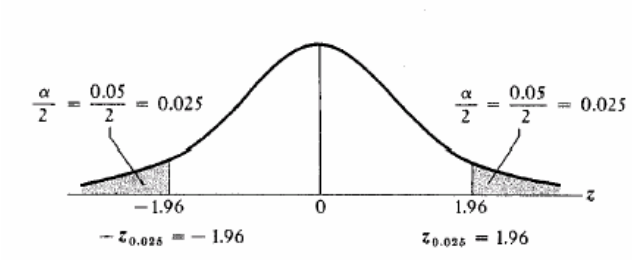
Of course, $\overline{X}$ is not always normally distributed, but this is usually not a concern so long as N ≥ 30. More importantly, σ is not always known. Therefore, how we construct confidence intervals will depend on the type of information and data available.

II. Computing confidence intervals. To compute confidence intervals, proceed as follows:

A. Calculate critical values for $z_{\alpha/2}$ or, if appropriate, $t_{\alpha/2,v}$, such that

$P(Z \leq z_{\alpha/2}) = 1 - \alpha/2 = F(z_{\alpha/2})$, or
$P(T_v \leq t_{\alpha/2,v}) = 1 - \alpha/2 = F(t_{\alpha/2,v})$

For example, as we have seen many times, if $\alpha = .05$, then $z_{\alpha/2} = 1.96$ since $F(1.96) = 1 - \alpha/2 = .975$. We divide α by 2 to reflect the fact that the true value of the parameter can be either greater than or less than the range covered by the confidence interval.



B. Confidence intervals for E(X) (or p) are then estimated by the following. Again, recall that the following statements will be true 100(1 - α)% of the time, that is, for all possible sample of size N, the true value of the population parameter will lie within the specified interval 100(1 - α)% of the time.

1. Case I. Population normal, σ known:

$$\overline{x} \pm (z_{\alpha/2} * \sigma/\sqrt{N}), i.e.,$$

$$\overline{x} - (z_{\alpha/2} * \sigma/\sqrt{N}) \leq \mu \leq \overline{x} + (z_{\alpha/2} * \sigma/\sqrt{N})$$

Recall that $\sigma/\sqrt{N}$ is the true standard error of the mean. Note also that, as N gets bigger and bigger, the standard error gets smaller and smaller, and the confidence interval gets smaller and smaller too. This means that, the larger your sample size, the more precise your point estimate will be. For example, different samples of size 20 could produce widely different estimates of the sample mean. Different samples of size 1000 will tend to produce fairly similar estimates of the sample mean. This is true, not just of case 1, but of all cases in general.

2. Case II. Binomial parameter p: An approximate confidence interval, which often works fairly well for large samples, is given by

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{N}}, i.e.$$

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{N}} \le p \le \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{N}}$$

The reason this is an approximation is because $\hat{p}\hat{q}$ is only an estimate of the variance. It turns out that there is actually a fair amount of controversy over how binomial confidence intervals should be estimated. Various formulas include Clopper-Pearson (which Stata questionably labels as "exact"), Wilson, Agresti-Coull, and Jeffreys. None of these are particularly easy to compute by hand, but the Wilson confidence interval (which may be the best, along with Jeffreys) is computed using this formula:

$$\frac{N}{N + z_{\alpha/2}^2}\left[\ \hat{p} + \frac{z_{\alpha/2}^2}{2N} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{N} + \frac{z_{\alpha/2}^2}{4N^2}}\ \right], i.e.$$

$$\frac{N}{N + z_{\alpha/2}^2}\left[\ \hat{p} + \frac{z_{\alpha/2}^2}{2N} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{N} + \frac{z_{\alpha/2}^2}{4N^2}}\ \right] \le p \le \frac{N}{N + z_{\alpha/2}^2}\left[\ \hat{p} + \frac{z_{\alpha/2}^2}{2N} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{N} + \frac{z_{\alpha/2}^2}{4N^2}}\ \right]$$

Caution: Prior to September 2004 I always used the Wilson CI and called it the "exact" CI. However, in 2001, Brown et al (see citation below) wrote a detailed discussion of the various methods that were available and the pros and cons of each. In Statalist on Sept. 7, 2004, Nick Cox wrote this comment about the various options:

> -exact- is something of a propaganda term. It just means the method due to Clopper and E.S. Pearson from 1934 or thereabouts. Even then a method due to E.B. Wilson in 1927 was available which (we know now) has generally better coverage properties. And the Jeffreys method, which although it has a Bayesian frisson to it, is interpretable as a continuity-corrected variant of the exact method. The Jeffreys method requires you to know -invibeta()- and is thus not congenial for hand calculation, but a doddle with current Stata.
>
> If you download -cij- and -ciw- from SSC you won't get any extra functionality (well, you will, but it's not documented), but you will get sets of references on the topic embedded in the help files.
>
> Above all, go straight to the paper by Brown, Cai and DasGupta in Statistical Science in 2001. This was the avalanche that started the stone rolling that led eventually to these changes to -ci-, added since the initial release of Stata 8.
>
> My reading of the literature, and some practical experience, is that especially for proportions near 0 or 1 the -exact- method can perform distinctly poorly while -jeffreys- and -wilson- can be relied on to give much more plausible answers.
>
> Turning this around, if the methods disagree the problem is thereby flagged as more difficult.

> It's arguable that we have here a bizarre situation, namely: many statistics texts have recommended the Clopper-Pearson method for decades and all along at least one better method was already available.

Roberto G. Gutierrez of StataCorp then followed up with this comment:

> The exact interval used by -ci, binomial- is the Clopper-Pearson interval, but you must realize that "exact" is a bit of a misnomer. It is exact in the sense that it uses the binomial distribution as the basis of the calculation. However, the binomial distribution is a discrete distribution and as such its cumulative probabilities will have discrete jumps, and thus you'll be hard pressed to get (say) exactly 95% coverage.
>
> What Clopper-Pearson does do is guarantee that the coverage is AT LEAST 95% (or whatever level you specify) and so it is desirable in that sense. It is able to accomplish this goal by using the exact binomial distribution in its calculations.
>
> However, by guaranteeing 95% coverage, Clopper-Pearson can be a bit conservative (wide) for some tastes, since for some n and p the true coverage can even get quite close to 100%. The other intervals (Jeffrey's, Agresti, Wilson) offered by -ci- are an attempt to not be so conservative, but yet still get the right coverage without the constraint of having to be at least the stated coverage level. These new intervals were added (by popular demand) after the release of Stata 8, and so you won't find them in the manual.
>
> The definitive article covering all this, including definitions for Jeffrey's, Agresti, and Wilson, is
>
> Brown, Cai, & DasGupta. Interval Estimation for a Binomial Proportion.
> Statistical Science, 2001, 16, pp. 101-133.
>
> Great article if you are into this sort of thing.

I don't want to spend hours going over the pros and cons of these different formulas! For our purposes, it probably won't matter too much which formula you use. But if your life depended on doing things as accurately as possible, it would be a good idea to check out the Brown article. The help files for the above-mentioned `cij` and `ciw` routines (which you can install by using the `findit` command in Stata) also contain brief discussions and extensive sets of references.

---

3. Case III. Population normal, σ unknown:

$$\bar{x} \pm (t_{\alpha/2,\nu} * s/\sqrt{N}), i.e.$$

$$\bar{x} - (t_{\alpha/2,\nu} * s/\sqrt{N}) \leq \mu \leq \bar{x} + (t_{\alpha/2,\nu} * s/\sqrt{N})$$

Case III is probably the most common. Even when Case II technically holds, treating it as though it were Case III often won't matter too much so long as N is large.

III. Examples.

1. $\overline{X}$ = 24.3, σ = 6, n = 16, X is distributed normally. Find the 90% confidence interval for the population mean, E(X).

Solution: Since σ is known, this falls under Case I. Note that the true standard error of the mean $= \sigma/\sqrt{N} = 6/\sqrt{16} = 6/4 = 1.5$. Also, α = .10, α/2 = .05, so the critical value of Z is 1.65 (since F(1.65) = 1 - α/2 = .95). Hence, the c.i. is

$$\bar{x} \pm (z_{\alpha/2} * \sigma/\sqrt{N}), i.e.,$$
$$24.3 - (1.65 * 6/4) \le \mu \le 24.3 + (1.65 * 6/4), i.e.,$$
$$21.83 \le \mu \le 26.78$$

Note: Again, remember that this does NOT mean that μ definitely lies between 21.83 and 26.78. Rather, we are merely saying that, 90% of the time, intervals constructed in the above fashion will include the true population value.

2. N = 100, $\hat{p}$ = .40. Construct a 95% c.i.

Solution. We want to construct a confidence interval for p (Case II). Since n is large, a normal approximation is appropriate. α = .05 and α/2 = .025, so the critical value for Z is 1.96 (since F(1.96) = 1 - α/2 = .975). Using the formula for the approximate binomial confidence interval, we get

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{N}}, i.e.$$
$$.4 - 1.96\sqrt{\frac{.4 * .6}{100}} \le p \le .4 + 1.96\sqrt{\frac{.4 * .6}{100}}, i.e.,$$
$$.304 \le p \le .496$$

Of course, there is no reason to use approximations when it is such a simple matter to get the real thing. So, using the Wilson formula for the binomial confidence interval, we get

$$\frac{N}{N + z_{\alpha/2}^2}\left[\hat{p} + \frac{z_{\alpha/2}^2}{2N} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{N} + \frac{z_{\alpha/2}^2}{4N^2}}\right], i.e.$$

$$\frac{100}{100 + 1.96^2}\left[.4 + \frac{1.96^2}{200} - 1.96\sqrt{\frac{.4 * .6}{100} + \frac{1.96^2}{40{,}000}}\right] \le p \le \frac{100}{100 + 1.96^2}\left[.4 + \frac{1.96^2}{200} + 1.96\sqrt{\frac{.4 * .6}{100} + \frac{1.96^2}{40{,}000}}\right], i.e.$$

$$.3094 \le p \le .4980$$

Note that, because N is large, the answers differ only slightly in this case. (Also note that I won't ever make you do this by hand).

3. X is distributed normally, n = 9, $\overline{X}$ = 4, s² = 9. Construct a 99% c.i. for the mean of the parent population.

Solution. σ is not known, and the sample size is small, hence we will have to use the T distribution (Case III) with v = 8. α = .01, so the critical value for T_8 is 3.355. (See Table III, Appendix E, v = 8 and 2Q = .01).

$$\bar{x} \pm (t_{\alpha/2,v} * s/\sqrt{N}), i.e.$$

$$4 - (3.355 * 3/3) \leq \mu \leq 4 + (3.355 * 3/3), i.e.$$

$$.645 \leq \mu \leq 7.355$$