Approximate Confidence Intervals for the Difference Between Two Proportions

We have learned that, thanks to the Central Limit Theorem and the Law of Large Numbers,

$$\overline{X}_n \pm Z_{\alpha/2}\sqrt{Var\left[\overline{X}_n\right]}$$

is an approximate confidence interval for the expected value, E[X], when the sample size (n) is large, even if the observations are coming from a distribution other than the normal.

We call it a "*large sample confidence interval*" or an "*approximate confidence interval*" to emphasize that the coverage probability is only approximate.

Thus,

$\overline{X}_n \pm Z_{\alpha/2}\sqrt{\frac{Var[X]}{n}}$ is an "approximate (1-α)100% CI".

When we wished to compare the difference between two means E[X] and E[Y], with

$X_1,\ldots,X_n$ are i.i.d. with E[X] and Var[X]
$Y_1,\ldots,Y_m$ are i.i.d. with E[Y] and Var[Y]

the same reasoning applied, and we derived the approximate (1-α)100% confidence interval for $E[X]-E[Y]=\mu_X-\mu_Y$ given by

$$\overline{X}_n-\overline{Y}_m \pm Z_{\alpha/2}\sqrt{Var\left[\overline{X}_n-\overline{Y}_m\right]}$$

$$\overline{X}_n-\overline{Y}_m \pm Z_{\alpha/2}\sqrt{\frac{S_X^2}{n}+\frac{S_Y^2}{m}}$$

Which works because

$$Z=\frac{\overline{X}_n-\overline{Y}_m-\left(E[X]-E[Y]\right)}{\sqrt{\frac{Var[X]}{n}+\frac{Var[Y]}{m}}} \overset{approx}{\sim} N(0,1) \quad \text{as n getslarge}$$

Notice that because the X's and Y's are independent we get $Var\left[\overline{X}_n-\overline{Y}_m\right]=Var\left[\overline{X}_n\right]+Var\left[\overline{Y}_m\right]$. Later we will run into situations when this independence is violated and we will need to `correct' the variance term. However the CLT approximation still holds.

When the X's and Y's are i.i.d. Ber(θ_x) and Ber(θ_y) the same large sample confidence interval applies. We see that an approximate (1-α)100% confidence interval for $E[X]-E[Y]=\theta_X-\theta_Y$ given by

$$\overline{X}_n-\overline{Y}_m \pm Z_{\alpha/2}\sqrt{Var\left[\overline{X}_n-\overline{Y}_m\right]}$$

$$\overline{X}_n-\overline{Y}_m \pm Z_{\alpha/2}\sqrt{\frac{Var[X]}{n}+\frac{Var[Y]}{m}}$$

$$\hat{\theta}_X-\hat{\theta}_Y \pm Z_{\alpha/2}\sqrt{\frac{\hat{\theta}_X\left(1-\hat{\theta}_X\right)}{n}+\frac{\hat{\theta}_Y\left(1-\hat{\theta}_Y\right)}{m}}$$

Because E[X]= θ_x and Var[X]= $\theta_x(1-\theta_x)$.

The alternative CI's that we just saw (the ones that replace the factor z by some t-value) are generally <u>not</u> used in this case, where the distribution of the *X*'s is not "nearly normal".

Example

Construct a 95% confidence intervals for the difference between two event rates when:

$X_1,\ldots,X_n$ are i.i.d. Poisson with $E[X]=Var[X]=\lambda_X$
$Y_1,\ldots,Y_m$ are i.i.d. Poisson with $E[Y]=Var[Y]=\lambda_Y$

Interpretation of a Confidence Interval

The confidence coefficient, (1-α)100%, tells us the coverage probability of our confidence interval *procedure* $\overline{X}_n \pm Z_{\alpha/2}\frac{S_n}{\sqrt{n}}$ (here a random interval).

It says that over many different samples, approximately (1-α)100% of the fixed (because we have data now) confidence intervals will contain the true mean.

This is why after observing data we can not say that the probability that the true mean is in the interval is interpretation (1-α)100%. (This probability only applies to the random interval otherwise known as the confidence interval procedure.)

Thus, the confidence coefficient is a property of the Confidence Interval procedure and <u>not</u> a property of any single fixed confidence itself.

So, after observing some data, what can we say about the individual confidence interval itself? How do we interpret the confidence interval?

- A common interpretation is this one:

"I'm (1-α)100% confident that the true mean is within the (fixed or observed) confidence interval."

But this is not really right, as my confidence about the interval will vary depending on a variety of this, such as the quality of the study, my algebraic skills, quality of the data, etc…

- A better interpretation is the following:

1. "A CI provides the best possible values for the unknown mean, at the (1-α)100% level."

2. "The CI provides the possible mean values that are consistent with the data at the (1-α)100% level."

3. "The data suggest or estimate the unknown mean to be in the range of _____ at the (1-α)100% level."

All these interpretations distinguish between properties of the CI procedure and those for the observed CI itself.

Example

Suppose we observe the weights of 18 insulin-dependent diabetics (measured as percentages of ideal body weights), and construct a 95% CI, we can be pretty sure that interval will contain the true expected weight of a person like those in our sample.

Suppose we do this (using the data in Pagano and Gauvreau) we find that $\overline{X}_{18} = 112.8$, and $S_{18} = 14.42$, giving an estimated standard error of $S_{18} / \sqrt{18} = 3.40$. For a 95% confidence level, the t table value for 17 degrees freedom is 2.11, so that our interval is

$$112.8 \pm (2.11)(3.40) \quad \text{which is} \quad 112.8 \pm 7.17$$

or

$$(105.6, 120.0).$$

The 95% CI consists of all those values of the parameter that are "consistent with the observations", or "consistent with the data."

Values outside the interval are "inconsistent with the data" (at the 95% confidence level).

We cannot say that our data on diabetics' weight proves that their weights tend to be greater than normal.

But these data do lend support to that hypothesis. They are evidence that the expected value of a diabetic's weight (as a percentage of normal) is greater than 100.

The values of μ inside the 95% CI, i.e., the values that are within ± 7.17 of the estimate, 112.8, are "consistent with the observations." Those that are so far from the estimate that they fall outside the 95% CI are "inconsistent with the observations (at the 95% confidence level)."

The CI shows which values of the parameter are compatible, or consistent, with the observations (at the specified confidence level), and which are not. It is one way of representing "what these observations tell us" about the parameter.

If diabetics were no different from non-diabetics, with respect to the distribution of body weight, then the true value of μ would be 100.

Since our interval, (105.6, 120.0), excludes this value, either μ is not 100, or we happen to have observed one of the relatively rare (one in twenty) cases where a 95% CI misses the true value.

In this sense, our sample represents evidence that μ is not 100, i.e., that diabetics' weights really differ, on the average, from non-diabetics' weights.

The evidence in our sample is actually somewhat stronger than this analysis shows — since the 99% CI, (102.9, 122.6) also excludes 100, either μ is not 100 or we have observed one of the rare one-in-a-hundred cases where a 99% CI misses the true value.

If the value 100 were outside the 99.9% CI, the evidence against it would be really strong. (99.9% CIs miss the true value only one time in 1000.) The evidence in our sample is not this strong, since the 99.9% CI is 112.8 ± (3.965)(3.40), or (99.3, 126.3), which includes the value 100.

One-Sided Confidence Intervals
(Upper and Lower Confidence Bounds)

Here we will consider a simple case, namely the case where the data are normal distributed with known variance, but the concepts apply more generally with little effort.

To derive the CI we rearranged the terms in the probability statement

$$P\left(-1.96 < \frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{\sigma} < 1.96\right) = 0.95$$

to obtain the 95% CI for the mean, μ: $\overline{X}_n \pm 1.96\ \sigma/\sqrt{n}$.
If we start with a single inequality,

$$P\left(\frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{\sigma} < 1.645\right) = 0.95$$

we can again rearrange to find the 95% confidence interval:

$$P\left(\overline{X}_n - 1.645\frac{\sigma}{\sqrt{n}} < \mu\right) = 0.95$$

$$\overline{X}_n - 1.645\frac{\sigma}{\sqrt{n}}$$

We can be 95% certain that the random point

$$\bar{X}_n - 1.645\frac{\sigma}{\sqrt{n}}$$

will fall below μ. This point , $\bar{X}_n - 1.645\frac{\sigma}{\sqrt{n}}$, is a 95% Lower Confidence Bound for μ.

The interval ($\bar{X}_n - 1.645\frac{\sigma}{\sqrt{n}}$, ∞) is a 95% one-sided confidence interval for μ. When we make observations and calculate the value of $\bar{X}_n - 1.645\frac{\sigma}{\sqrt{n}}$, we have some basis for confidence that the unknown value of μ is at least that large.

In the same way, we can find a 95% upper confidence bound, for μ: $\bar{X}_n + 1.645\frac{\sigma}{\sqrt{n}}$.

Likewise, from

$$P\left(\frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{S_n} < t_{\alpha/2}^{n-1}\right) = 0.95$$

we can obtain 95% a lower confidence bound for μ when the variance is unknown: $\overline{X}_n - t_{\alpha/2}^{n-1}\frac{S_n}{\sqrt{n}}$ (and similarly for an upper bound).

How Large Is "Large"?

If our random variables $X_1, X_2, \ldots, X_n$ are independent with a common probability distribution that is not normal, when is the sample size n large enough that we can use the standard normal distribution to approximate the distribution of

$$\frac{\sqrt{n}\left(\overline{X}_n - E[X]\right)}{\sqrt{\hat{V}ar[X]}}$$ with a normal distribution?

That is, when can we say that the probability that the interval

$$\overline{X}_n \pm 1.96\frac{S_n}{\sqrt{n}}$$

will contain *E(X)* really is approximately 0.95? There is no simple answer to this question.

This creates some awkwardness — as soon as any number is proposed, such as " n = 30 is large enough," some smart-aleck will point out that he can find a distribution for which that value of n is not large enough.

On the other hand, for any distribution that he chooses, I can find an n large enough that the approximation is very accurate whenever the sample is at least that large.

Whether the approximation is good enough, for a specific probability distribution and sample size, depends primarily on the skewness of the distribution. For distributions that are like the normal in the sense that they are symmetric (not skewed at all), the approximation works well, even in small samples.

It is the skewness of the distribution that really dictates how large a sample is required for the distribution of the sample average to be almost symmetrical.

Sample Size for Confidence intervals

Suppose $X_1,\ldots,X_n$ are i.i.d. Normal with $E[X]=\mu$ and $Var[X]=\sigma$.

A (1-α)100% CI is $\overline{X}_n \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

The length of the interval is

$$L = 2 \times Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

And the Margin of Error is defined as

$$MOE = Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Solving for the n gives us two very desirable equations!

$$n = \left(2 \times Z_{\alpha/2}\frac{\sigma}{L}\right)^2 = 4Z_{\alpha/2}^2\frac{\sigma^2}{L^2}$$

$$n = \left(Z_{\alpha/2}\frac{\sigma}{MOE}\right)^2 = Z_{\alpha/2}^2\frac{\sigma^2}{MOE^2}$$

The general formula is to see as

$$n = 4Z^2_{\alpha/2}\frac{Var[X]}{L^2}$$

For example, if you are interested in proportions then

$$\text{Var}[X]=\theta(1-\theta)$$

and

$$n = 4Z^2_{\alpha/2}\frac{\theta(1-\theta)}{L^2}$$

Note that we almost never use a t-distribution to calculate sample size. (Why?)