James Green-Armytage
Econ 229, fall 2015

Introduction to Confidence Intervals

1. Expected value and variance: properties

Let $\mathrm{E}(x)$ be the expected value of a random variable x. Let $\mathrm{var}(x) = \mathrm{E}\{[x - \mathrm{E}(x)]^2\}$ be the variance of x. Let $\mathrm{sd}(x) = \sqrt{\mathrm{var}(x)}$ be the standard deviation of x.

If x and y are two independent random variables, and a and b are two constants, the following identities hold:

$$\mathrm{E}(x + y) = \mathrm{E}(x) + \mathrm{E}(y) \qquad \mathrm{var}(x + y) = \mathrm{var}(x) + \mathrm{var}(y)$$

$$\mathrm{E}(ax) = a\mathrm{E}(x) \qquad \mathrm{var}(ax) = a^2\mathrm{var}(x)$$

$$\mathrm{E}(ax + by) = a\mathrm{E}(x) + b\mathrm{E}(y) \qquad \mathrm{var}(ax + by) = a^2\mathrm{var}(x) + b^2\mathrm{var}(y)$$

The identify $\mathrm{var}(ax) = a^2\mathrm{var}(x)$ deserves a brief argument: $\mathrm{var}(ax) = \mathrm{E}\{[ax - \mathrm{E}(ax)]^2\} = \mathrm{E}\{[ax - a\mathrm{E}(x)]^2\} = \mathrm{E}\{a^2[x - \mathrm{E}(x)]^2\} = a^2\mathrm{E}\{[x - \mathrm{E}(x)]^2\} = a^2\mathrm{var}(x)$.

2. Expected value and variance of sample means

Let $\boxed{\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i}$ be a sample mean, i.e. a simple average of n random and independent draws from a population. Let μ, σ^2, and σ be the mean, variance, and standard deviation of x.

Combining these definitions with properties above, we derive the identity $\boxed{\mathrm{E}(\bar{x}) = \mu}$

$$\mathrm{E}(\bar{x}) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n} x_i\right) = \frac{1}{n}E(x_1 + \cdots + x_n)$$

$$= \frac{1}{n}[E(x_1) + \cdots + E(x_n)] = \frac{1}{n}(n \cdot \mu) = \mu$$

Next, we derive the identity $\boxed{\mathrm{var}(\bar{x}) = \frac{\sigma^2}{n}}$

$$\mathrm{var}(\bar{x}) = \mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \left(\frac{1}{n}\right)^2 \mathrm{var}\left(\sum_{i=1}^{n} x_i\right) = \left(\frac{1}{n}\right)^2 \mathrm{var}(x_1 + \cdots + x_n)$$

$$= \left(\frac{1}{n}\right)^2 [\mathrm{var}(x_1) + \cdots + \mathrm{var}(x_n)] = \left(\frac{1}{n}\right)^2 (n \cdot \sigma^2) = \frac{\sigma^2}{n}$$

From this, it follows immediately that $\boxed{\mathrm{sd}(\bar{x}) = \frac{\sigma}{\sqrt{n}}}$. We can also call this the *standard error* of the sample mean, and write it as $SE_{\bar{x}}$.

By the central limit theorem, the distribution of our sample mean tends toward a normal distribution as n increases. So with a sufficiently large sample, the distribution of $\bar{x}$ is approximately a normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

3. Normal confidence intervals

In the next section we will proceed from the premise that $\bar{x}$ is distributed approximately according to a normal distribution with mean μ and standard deviation $\mathrm{sd}(\bar{x}) = SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, and use observable statistics to estimate a *confidence interval*, i.e. range in which we believe the true population mean is located with probability k. In this section, we will lay some groundwork, by first considering how values of normally distributed random variables map to probabilities.

Consider first a standard normal distribution, with mean $\mu = 0$, and standard deviation $\sigma = 1$. Using a table of the standard normal CDF we can verify that a variable x with this distribution will fall within the range $-1.96 < x < 1.96$ with probability $k = .95$. Similarly, x will fall within the larger range $-2.58 < x < 2.58$ with probability $k = .99$, and it will fall within the smaller range $-1.64 < x < 1.64$ with probability $k = .9$.

As an alternative to using a table, we can calculate these intervals using Excel. That is, if we want a value z^* such that a standard normal random variable x will fall with probability k in the interval $-z^* < x < z^*$, we can find it with the command

```
=normsinv( ( k + 1 ) / 2 )
```

We can write this equivalently in mathematical notation as $\boxed{z^* = \Phi^{-1}\left(\frac{\mathrm{k}+1}{2}\right)}$, where $\Phi^{-1}(F)$ is the inverse cumulative distribution function of a standard normal random variable.

For example, given $k = .95$, we start with $\frac{k+1}{2} = .975$. We want to know what value z^* has standard normal CDF values of $F(z^*) = .975$ and thus $F(-z^*) = .025$, so that $F(z^*) - F(-z^*) = .95$. We can calculate this with the command `=normsinv(.975)`.

Next, consider a generic normal distribution with mean μ and standard deviation σ. As before, a variable x with this distribution will fall within z^* standard deviations from the mean with probability k. In other words, with confidence k, we can say that x will take on a value in the range $\mu \pm z^* \cdot \sigma$, where z^* can still be calculated via `=normsinv( (k+1)/2 )`.

Alternatively, we can calculate the upper and lower bounds of the range, respectively, using the commands `=norminv( (k+1)/2, μ, σ )` and `=norminv( 1-(k+1)/2, μ, σ )`.

4. Confidence intervals for the sample mean

Now we are ready to consider our sample mean $\bar{x}$. By the preceding logic, we know that with probability k, $\bar{x}$ will fall in the range $\mathrm{E}(\bar{x}) \pm z^* \cdot \mathrm{sd}(\bar{x})$, which is equivalent to $\mu \pm z^* \frac{\sigma}{\sqrt{n}}$. But we must recall that, from the perspective of the statistician, μ and σ are not directly observed; indeed, μ is precisely what we are trying to estimate.

However, since we know that the sample mean $\bar{x}$ will fall less than $z^* \frac{\sigma}{\sqrt{n}}$ away from the true mean μ with probability k, we can say with the same confidence that the unobserved true mean μ will be somewhere within $z^* \frac{\sigma}{\sqrt{n}}$ of our observed sample mean $\bar{x}$. So, with probability k, the true mean μ will be in the range $\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$.

We are almost finished with the formula for our confidence interval. Only one wrinkle remains: Like μ, the true variance σ^2 and standard deviation σ may be unknown to the researcher. However, we may estimate them using the data. The standard formulae are $\boxed{s^2 = \frac{1}{n-1}\sum_{i=1}^{n}[(x_i - \bar{x})^2]}$ and $\boxed{s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}[(x_i - \bar{x})^2]}}$; these give us the *sample variance* and *sample standard deviation*, respectively. Note that we usually divide the sum of squared deviations from the mean not by the full sample size n, but instead by $n - 1$; this adjustment, called “Bessel’s correction”, removes what would otherwise be a downward bias in our estimate.

Note also that Excel can calculate the sample mean, the (corrected) sample variance, and the (corrected) sample standard deviation of some selection of numbers with the formulas `=average( )`,`=var( )`, and `=stdev( )`, respectively.

Putting all of this together, we can now construct the formula for our confidence interval: With confidence k, the population mean μ should be located in the interval

$$\boxed{\bar{x} \pm z^* \frac{s}{\sqrt{n}}}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is our sample mean, $z^* = \Phi^{-1}\left(\frac{\mathrm{k}+1}{2}\right)$ is the critical z value that corresponds to our confidence level k, $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}[(x_i - \bar{x})^2]}$ is our sample standard deviation, and n is our sample size.