Confidence intervals for the sample mean

Confidence Interval: Suppose that θ is some unknown parameter and that we are using the statistic $\widehat{\Theta} = h(X_1, \ldots, X_n)$ as an estimator for θ. Suppose furthermore, that we can find two statistics

$$L = L(X_1, \ldots, X_n) \quad \text{and} \quad U = U(X_1, \ldots, X_n)$$

(usually based upon $\widehat{\Theta}$) such that

$$1 - \alpha = P(L \leq \theta \leq U),$$

where α is a constant between 0 and 1. Usually $1 - \alpha$ will be a large value, e.g. 90%, 95% or 99%.

Let l and u be the observed values of L and U, respectively.

We say that $[l, u]$ is a $100\,(1 - \alpha)\%$ **confidence interval** for θ.

Terminology:

- $1 - \alpha$ is called the confidence level or confidence coefficient.
- α is called the error rate.
- $u - l$=length of the interval is a measure of the precision of the estimate. A smaller length interval is interpreted as a more precise estimation.

Interpretation: Suppose that $1 - \alpha = 0.95$ and that from the n observations we get the following confidence interval $[10, 15]$ for μ. We say that we are 95% **confident** that $10 \leq \mu \leq 15$. What does that mean?

We do not know whether μ is between 10 and 15, but the technique that we used to produce the interval yields an interval that contains μ 95% of the time.

Definition: Let Z be a standard normal random variable. Its upper quantile of order A is a value z_A such that

$$P(Z > z_A) = A,$$

i.e., the area under the probability density function of Z to the right of z_A is A.

We can find the quantiles of the standard normal distribution in the table of the quantiles for the T distribution on blackboard: In the last row of the table for the T-distribution, we have $z_\alpha = t_{\alpha;\infty}$.

Estimating μ (when σ is known)

Assumptions:

- Population is normal or the sample size is large ($n \geq 30$).
- Population variance σ^2 is known.

Under these assumptions,

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

follows a standard normal distribution. It is approximate in the case of a non-normal population with a large sample size. It follows that

$$\begin{aligned} 1 - \alpha &= P\left(-z_{\alpha/2} \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) \\ &= P\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \end{aligned}$$

Therefore, if the population is normal or $n \geq 30$, a $100\,(1-\alpha)\%$ confidence interval for μ is

$$\overline{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Remarks:

- The standard deviation of the estimator can be used as a measure for the typical error, and this is called the **typical error** of the estimator. Here the typical error for the sample mean is

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}.$$

- We say that

$$E = z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}$$

  is the **maximal error** of the sample mean with confidence level of $(1-\alpha)100\%$.

- In practice, we do not know σ (σ is the true standard deviation of the population). So, we must replace the unkown constant σ with an estimator S (the sample standard deviation). We can show that

$$\lim_{n\to\infty} F_Z(z) = \Phi(z),$$

  if

$$Z = \frac{\overline{X}-\mu}{S/\sqrt{n}}.$$

  So, $\overline{X} \sim N(\mu;\sigma^2/n)$ approximately when n is large. Here we will assume that $n \geq 40$ is sufficiently large for a good approximation.

  So, if $n \geq 40$, a confidence interval of $100\,(1-\alpha)\%$ for μ is (approximately)

$$\overline{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}.$$

Example 45: A machine produces metal springs used in automobile shock absorbers. A random sampling of 10 springs is taken and the diameter of each spring is measured. Suppose that the diameter (in mm) is normally distributed with $\sigma = 0.2$.

The diameter of the sampled springs (in mm) are:

$$8.24, 8.25, 8.20, 8.23, 8.24, 8.21, 8.26, 8.26, 8.19, 8.23$$

We have also computed the following statistics (using the `R` programming language):

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  8.190   8.215   8.235   8.231   8.248   8.260
```

(a) Give a 95% confidence interval for μ.

solution: $\overline{x} = 8.231$, $\sigma = 0.2$
$1 - \alpha = 0.95$ so $\alpha = 0.05$. Therefore,

$$z_{\frac{\alpha}{2}} = z_{0.025} == t_{0.025,\infty} = 1.96$$

so a $100(1-\alpha)\%$ confidence interval for μ is

$$\overline{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} 8.231 \pm 1.96(\frac{0.2}{\sqrt{10}})$$

(b) Give a 97% confidence interval for μ.
solution: Same as part (a), but use $\alpha = 0.03$

(c) Give a 99% confidence interval for μ.
solution: Same as part (a), but use $\alpha = 0.01$

(d) Compare the length of these three intervals.
solution: The intervals with a smaller value of α have a larger width.

Precision:

The precision of the interval is defined as the length of the interval, that is

$$\overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} - (\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 2\,\frac{z_{\alpha/2}\sigma}{\sqrt{n}}.$$

Remarks:

- The precision is a function of the confidence level and of the sample size.
- As we increase the confidence level, the estimation is less precise.
- As we increase the sample size, we increase precision.
- In practice, we would like high precision and high confidence. To do so, we can fix the confidence level, and then choose the appropriate sample size to control the precision of the estimate.

Sample Size:

If $\overline{x}$ is used as an estimate of μ, we can be $100\,(1-\alpha)\%$ confident that the error $|\overline{x} - \mu|$ will not exceed a specified amount E when the sample size is

$$n \geq \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2.$$

Remarks:

- If the value of n that you compute is not an integer, then it must be rounded-up to the nearest integer.
- If σ is unknown, then it is common practice to collect a preliminary sample and use the sample standard deviation s instead of σ in the formula to compute n.

Example 46: Consider a population with standard deviation $\sigma = 9.6$. Suppose that we want to be 95% confident that the sample mean $\overline{x}$ has an error less than 3 units when estimating the mean μ. What sample size should we use?

solution: $\alpha = 0.05$. We want $|\overline{x} - \mu| < E$, where $E = 3$. So we need

$$n \geq \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2.$$

but

$$z_{\alpha/2} = z_{0.025} = t_{0.025,\infty} = 1.96$$

so

$$n \geq \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2 = \left(\frac{1.96 \times 9.6}{3}\right)^2 = 39.34,$$

so we need to use at least $n = 40$ samples.

One-Sided Confidence Intervals

We sometimes want a lower-bound (or upper bound) for μ. We can do this with a one-sided confidence interval.

If the population is normal or the sample size is large ($n \geq 30$), then
(i) a $100\,(1 - \alpha)\%$ upper-confidence bound for μ is

$$\mu \leq \overline{x} + z_\alpha \frac{\sigma}{\sqrt{n}};$$

(ii) a $100\,(1 - \alpha)\%$ lower-confidence bound for μ is

$$\mu \geq \overline{x} - z_\alpha \frac{\sigma}{\sqrt{n}}.$$

Example 47: We study the yield of a chemical process. It is important that the yield be at least 0.9. If we have strong evidence that the mean yield μ is less than 0.9, we should make improvements to the chemical process. Suppose that the yield is normally distributed with standard deviation $\sigma = 0.02$. From a random sample of size $n = 10$, we have a sample mean of $\overline{x} = 0.88$. Construct a 95% upper-confidence bound for μ. Are we confident that the yield $\mu < 0.9$?

solution: $\alpha = 0.05$, a $100(1-\alpha)\%$ confidence bound for μ is

$$\mu \leq \overline{x} + z_\alpha \frac{\sigma}{\sqrt{n}}$$

but $z_\alpha = z_{0.05} = 1.645$ so

$$\overline{x} + z_\alpha \frac{\sigma}{\sqrt{n}} = 0.88 + 1.645 \frac{0.02}{\sqrt{10}} = 0.8904$$

so

$$\mu \leq 0.8904$$

with 95% confidence.

Therefore, yes, we are at least 95% confident that $\mu < 0.9$.

Example 48: We sampled the water in $n = 53$ lakes in a mining region. Here are statistics for the concentration of mercury (in parts per billion)

n	mean	standard deviation
53	0.52	0.349

Calculate a 95% lower-confidence bound for μ.

solution: $\alpha = 0.05$ so $z_\alpha = z_{0.05} = 1.645$ so a 95% lower confidence bound for μ is

$$\mu \geq \overline{x} - z_\alpha \frac{\sigma}{\sqrt{n}}.$$

Now,

$$\overline{x} - z_\alpha \frac{\sigma}{\sqrt{n}} = 0.52 - 1.645 \frac{0.349}{\sqrt{53}} = 0.441.$$

Therefore, $\mu \geq 0.441$ with 95% confidence.

Estimating the mean of a normal population (σ unknown)

In the special case of a normal population, it is possible to construct a confidence interval for the mean even when σ is unknown.

Definition: Consider the random variable T with probability density function

$$f(t) = \frac{\Gamma[(\nu+1)/2]}{\sqrt{\pi\,\nu}\Gamma(\nu/2)} \left[\frac{x^2}{\nu} + 1\right]^{-(\nu+1)/2}, \quad -\infty < t < \infty.$$

We say that T follows a t distribution with ν degrees of freedom.

Properties: Consider a t distribution with ν degrees of freedom.

- Its upper quantile of order α is denoted $t_{\alpha,\nu}$. We find them in a table on blackboard.
- The density $f(t)$ is symmetric about $t = 0$, hence

$$t_{1-\alpha,\nu} = -t_{\alpha,\nu}.$$

- When $\nu \to \infty$, then the t distribution converges to a standard normal, hence

$$z_\alpha = t_{\alpha,\infty}.$$

  So we can use this same table to find quantiles for the $N(0, 1)$ distribution. For example, $z_{.025} = t_{.025,\infty} = 1.96$.

Theorem: Let $X_1, \ldots, X_n$ be a random sample from a **normal** population with mean μ and variance σ^2. The random variable

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

has a t distribution with $\nu = n - 1$ degrees of freedom.

Therefore, if the population is normal, then a $100\,(1 - \alpha)\%$ confidence interval for μ is

$$\overline{x} \pm t_{\alpha/2,n-1}\,\frac{s}{\sqrt{n}}.$$

Example 49: Consider the scenario in Example 45, but now suppose that σ^2 is unknown. We will assume that the diameter of the springs is normally distributed. Below are values for statistics for the diameters of a random sampling of $n = 10$ springs. Construct a 95% confidence interval for the mean spring diameter.

```
  Min.   1st Qu.     Median    Mean     3rd Qu.    Max.
 8.190   8.215        8.235     8.231    8.248     8.260
> standard deviation
   0.02424413
```

solution: $\overline{x} = 8.231$, $s = 0.0242$, $n = 10$
$\alpha = 0.05$, so $\frac{\alpha}{2} = 0.025$. so

$$t_{\alpha/2,n-1} = t_{0.025,9} = 2.262$$

so a 95% confidence interval is

$$\overline{x} \pm t_{\alpha/2,n-1} \frac{s}{\sqrt{n}} = 8.231 \pm 2.262 \frac{0.0242}{\sqrt{10}} = 8.231 \pm 0.0152$$