Confidence intervals of the population mean

MATH 205–Fall 2022

The story so far...

In the first half of the course, we set up the probabilistic foundations for modern statistical analyses. In this chapter, we focus on one of the simplest (but most important) problems of this type: confidence intervals of the sample mean.

There are a few materials that should be reviewed.

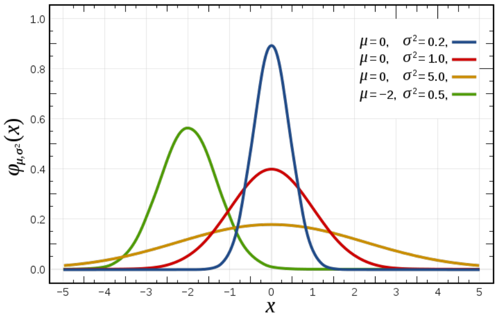


Figure 1: $\mathcal{N}(\mu,\sigma^2)$

Definition 1 (Random sample)**.** *The random variables $X_1, X_2, ..., X_n$ are said to form a (simple) random sample of size n if*

1. *the X_i's are independent random variables*
2. *every X_i has the same probability distribution*

Remark 2. *The normal distribution plays a central role in statistics, and computations with normal distributions are a must-have skill for the rest of the materials of the course.*

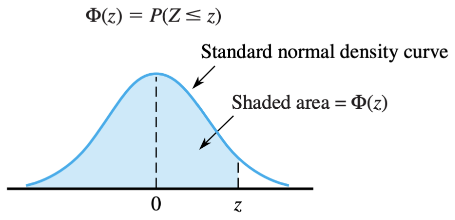


Figure 2: Computations with standard normal distribution is done through the z-table of the CDF of $\mathcal{N}(0,1)$

Theorem 3 (Linear combinations of random variables)**.** *Let $X_1, X_2, \ldots, X_n$ be independent random variables (with possibly different means and/or variances). Define*

$$T = a_1X_1 + a_2X_2 + \ldots + a_nX_n$$

then the mean and the standard deviation of T can be computed by

- $E(T) = a_1E(X_1) + a_2E(X_2) + \ldots + a_nE(X_n)$
- $Var(T) = a_1^2Var(X_1) + a_2^2Var(X_2) + \ldots + a_n^2Var(X_n)$

Moreover, if X_i's are all normal random variables, then so is T.

Theorem 4 (Mean and variance of the sample mean)**.** *Given independent random samples $X_1, X_2, ..., X_n$ from a distribution with mean μ and standard deviation σ, the mean is modeled by a random variable $\bar{X}$,*

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

Then

$$E[\bar{X}] = \mu, \quad \textit{and} \quad Var(\bar{X}) = \frac{\sigma^2}{n}$$

Theorem 5 (Law of large numbers)**.** *Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. Then*

$$\bar{X} \to \mu$$

as n approaches infinity.

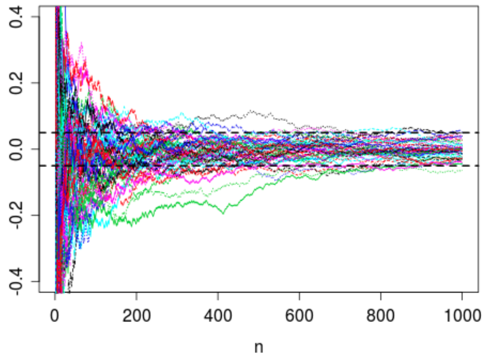


Figure 3: Law of large numbers

Theorem 6 (The Central Limit Theorem)**.** *Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. Then, in the limit when $n \to \infty$, the $\bar{X}$ follows the normal distribution.*

Recall that

$$E[\bar{X}] = \mu, \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}},$$

this means we have

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

follows the standard normal distribution.

Rule of Thumb: If $n > 30$, the Central Limit Theorem can be used for computation.

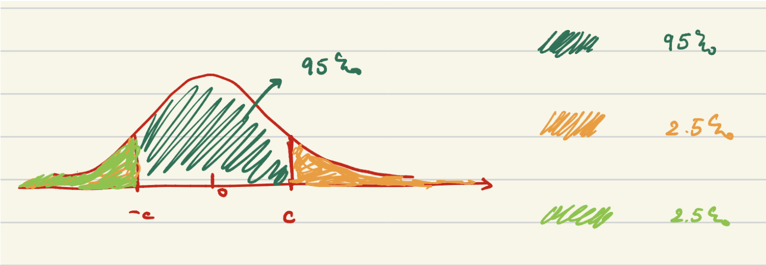
Normal distribution with known σ

Example 7. *Suppose the sediment density (g/cm) of a randomly selected specimen from a certain region is normally distributed with mean μ (unknown) and standard deviation $\sigma = 0.85$. A random sample of $n = 25$ specimens is selected with sample average $\bar{X}$.*

Find a number c such that

$$P\left[-c < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < c\right] = 0.95$$

Solution. We first note that $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ follows normal distribution.



- Since the normal density is symmetric: light green = orange = 2.5%.
- $\Phi(c)$ = light green + dark green = 0.975

We then can look up the z-table and obtain $c = 1.96$ □

A detailed construction of the 95% confidence interval

- We have

$$P\left[-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right] = 0.95$$

- Rearranging the inequalities gives

$$P\left[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right] = 0.95$$

- This means that if you use

$$\left[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

  as a range to estimate μ, then you are correct 95% of the time!

Definition 8. *Choose some fraction f. An f confidence interval for a population mean is an interval constructed using the sample mean. It has the property that for that fraction f of all samples, the population mean will lie inside the interval constructed from each sample's mean.*

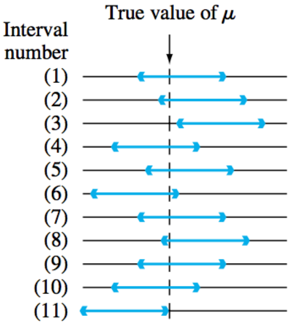


Figure 4: 95% confidence interval: If we repeat the experiment many times, the interval contains μ about 95% of the time

Theorem 9 (95% CIs). *Suppose that the parameter of interest is a population mean μ and that*

- *The population distribution is normal.*
- *The value of the population standard deviation s is known.*

If after observing $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, we compute the observed sample mean $\bar{x}$ and then substitute $\bar{x}$ in place of $\bar{X}$, the resulting fixed interval is called a 95% confidence interval for μ. This CI can be expressed as

$$\left[\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

Theorem 10 (General levels of confidence). *A $100(1-\alpha)\%$ confidence interval for the mean μ of a normal population when the value of σ is known is given by*

$$\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

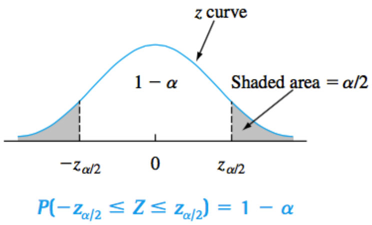


Figure 5: $100(1-\alpha)\%$ confidence interval

Example 11. *Assume that the helium porosity (in percentage) of coal samples taken from any particular seam is normally distributed with true standard deviation $\sigma = .75$.*

- *Compute a 90% CI for the true average porosity of a certain seam if the average porosity for 20 specimens from the seam was 4.85.*
- *How large the sample size would be if we want the width of the 90% interval to be .40?*

Example 12. *Suppose the sediment density (g/cm) of a randomly selected specimen from a certain region is normally distributed with mean μ (unknown) and standard deviation $\sigma = 0.85$. A random sample of $n = 25$ specimens is selected with sample average $\bar{X}$.*

Find a number c such that

$$P\left[\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < b\right] = 0.95$$

Theorem 13 (One-sided confidence intervals). *For a normal population when the value of σ is known:*

- *A $100(1-\alpha)\%$ upper confidence bound for μ is given by*

$$\mu < \bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}}$$

- *A* $100(1-\alpha)\%$ *lower confidence bound for* μ *is given by*

$$\mu > \bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}}$$

Example 14. *A sample of 66 obese adults was put on a low-carbohydrate diet for a year. The (sample) average weight loss was 7.7 lb. It is known that the standard deviation of weight loss is 19.olb.*

Calculate a 99% lower confidence bound for the true average weight loss. From the result, do you think that the diet is effective?

Large-sample confidence intervals

The analysis in the previous section sets up the foundations for deriving CIs for the population mean. However, its limitations are very clear:

- The analysis only applies to normal distributions. While we can empirically verify this condition, this imposes a limit on the type of populations that we can analyze.
- The analysis assumes that σ is known. This is very unlikely in practice.

Thankfully, we have some powerful tools to address those limits:

- By the Central Limit Theorem, when $n > 30$, $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ follows (approximately) the standard normal distribution, regardless of the distribution of X.
- We also know that when $n > 40$, we can approximate σ by the sample standard deviation

  $$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}}$$

  or its variation

  $$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$$

The conclusion is that when $n > 40$,

$$\frac{\bar{X}-\mu}{s/\sqrt{n}}$$

follows (approximately) the standard normal distribution.

A detailed construction of the 95% confidence interval

- We have

  $$P\left[-1.96 < \frac{\bar{X}-\mu}{s/\sqrt{n}} < 1.96\right] = 0.95$$

- Rearranging the inequalities gives

  $$P\left[\bar{X} - 1.96\frac{s}{\sqrt{n}} \le \mu \le \bar{X} + 1.96\frac{s}{\sqrt{n}}\right] = 0.95$$

- This means that if you use

  $$\left[\bar{X} - 1.96\frac{s}{\sqrt{n}}, \bar{X} + 1.96\frac{s}{\sqrt{n}}\right]$$

  as a range to estimate μ, then you are correct 95% of the time!

Theorem 15 (Large-sample confidence intervals)**.** *When $n > 40$, a $100(1-\alpha)\%$ confidence interval for the mean μ of a population is given by*

$$\left[\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}\right]$$

Theorem 16 (Large-sample confidence bound)**.** *When $n > 40$:*

- *A $100(1-\alpha)\%$ upper confidence bound for μ is given by*

$$\mu < \bar{x} + z_{\alpha}\frac{s}{\sqrt{n}}$$

- *A $100(1-\alpha)\%$ lower confidence bound for μ is given by*

$$\mu > \bar{x} - z_{\alpha}\frac{s}{\sqrt{n}}$$

Example 17. *A random sample of 50 patients who had been seen at an outpatient clinic was selected, and the waiting time to see a physician was determined for each one, resulting in a sample mean time of 40.3 min and a sample standard deviation of 28.0 min.*

- *Construct a 95% confidence interval of the true average waiting time.*
- *Assuming that the true standard deviation of the waiting time is 27 min, construct a 95% confidence interval of the true average waiting time.*

Solution. We note that in this problem, $n > 40$ and we can use large sample intervals:

$$\left[\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}\right]$$

where

$$n = 50, \quad \bar{x} = 40.3, \quad s = 28.0, \quad \alpha = 0.05, \quad z_{\alpha/2} = 1.96$$

When $\sigma = 27$ is known, as in the second part, we don't have to approximate σ by s, and can directly use

$$\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

as the confidence interval. □

Example 18. *A sample of 66 obese adults was put on a low-carbohydrate diet for a year. The (sample) average weight loss was 7.7 lb and the sample standard deviation was 19.1 lb. Calculate a 99% lower confidence bound for the true average weight loss. From the result, do you think that the diet is effective?*

Solution. Again, $n > 40$ and we can use

$$n = 66, \quad \bar{x} = 7.7, \quad s = 19.1, \quad \alpha = 0.01, \quad z_{\alpha} = 2.33$$

to derive the large-sample the lower confidence bound:

$$\bar{x} - z_{\alpha}\frac{s}{\sqrt{n}} \approx 2.22$$

The lower bound of the true average weight loss is strictly positive, so we can tell that the diet is effective. □

Small-sample normal populations

Assumption 19. *The population of interest is normal, so that $X_1, X_2, \ldots, X_N$ constitutes a random sample from a normal distribution with both μ and σ unknown. We will also assume that $n < 40$.*

Remark 20. *When $n < 40$, we will use the adjusted formula*

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

for the sample standard deviation. It is worth noting that when $n < 40$, S is no longer close to σ. Thus

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

does not follow the standard normal distribution.

Definition 21. *The t distribution with degree of freedom ν has the following probability density function*

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

When $\nu > 40$, $t_\nu \approx \mathcal{N}(0, 1)$.

Theorem 22. *When $\bar{X}$ is the mean of a random sample of size n from a normal distribution with mean μ, the random variable*

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

follows t distribution with $n - 1$ degree of freedom (df).

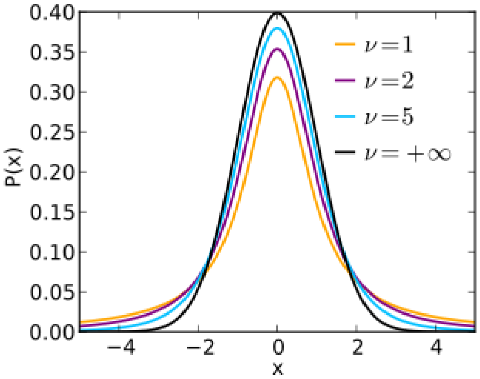


Figure 6: t distributions

A detailed construction of the 95% confidence interval

- We have

$$P\left[-t_{\alpha/2,n-1} < \frac{\bar{X} - \mu}{s/\sqrt{n}} < t_{\alpha/2,n-1}\right] = 1 - \alpha$$

- Rearranging the inequalities gives

$$P\left[\bar{X} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} \le \mu \le \bar{X} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right] = 1 - \alpha$$

- This means that if you use

$$\left[\bar{X} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}, \bar{X} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right]$$

  as a range to estimate μ, then you are correct $100(1 - \alpha)\%$ of the time!

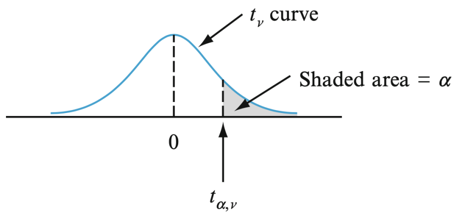


Figure 7: t critical values

Theorem 23. *Let $\bar{x}$ and s be the sample mean and sample standard deviation computed from the results of a random sample from a normal population with mean μ.*

- *A $100(1-\alpha)\%$ confidence interval for the mean μ of a population is given by*

$$\left[\bar{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right]$$

- *A $100(1-\alpha)\%$ upper confidence bound for μ is given by*

$$\mu < \bar{x} + t_{\alpha,n-1}\frac{s}{\sqrt{n}}$$

- *A $100(1-\alpha)\%$ lower confidence bound for μ is given by*

$$\mu > \bar{x} - t_{\alpha,n-1}\frac{s}{\sqrt{n}}$$

Example 24. *Let X be the amount of butterfat in pounds produced by a typical cow during a 305-day milk production period between her first and second calves. Assume that the distribution of X is $N(\mu, \sigma^2)$.*

To estimate μ, a farmer measured the butterfat production for n = 20 cows and obtained the following data:

481	537	513	583	453	510	570	500	457	555
618	327	350	643	499	421	505	637	599	392

- *Construct a 90% confidence interval for μ.*
- *Find a 90% one-sided confidence interval that provides an upper bound for μ.*

Solution. We first check the assumptions:

- $n < 40$
- Normal distribution
- σ is unknown

Thus, we need to use the t-distribution-based formulas

$$\left[\bar{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right]$$

and

$$\mu < \bar{x} + t_{\alpha,n-1}\frac{s}{\sqrt{n}}.$$

Here

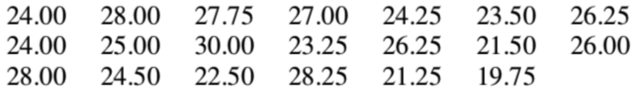
$$n = 20, \quad \bar{x} = 507.5, \quad s = 89.75$$

and

$$\alpha = 0.01, \quad \alpha/2 = 0.005, \quad t_{\alpha/2,n-1} = t_{0.005,19} = 2.86, t_{\alpha,n-1} = t_{0.01,19} = 2.54.$$

□

Example 25. *Here is a sample of ACT scores for students taking college freshman calculus:*



24.00	28.00	27.75	27.00	24.25	23.50	26.25
24.00	25.00	30.00	23.25	26.25	21.50	26.00
28.00	24.50	22.50	28.25	21.25	19.75	

Assuming that ACT scores are normally distributed, calculate a two-sided 95% confidence interval for the population mean.