

MAT 2377

Probability and Statistics for Engineers

Chapter 5

Point and Interval Estimation

P. Boily (uOttawa)

Winter 2021

Contents

5.1 – Statistical Inference (p.2)

- Statistics (p.4)
- Estimator Variance and Standard Error (p.5)

5.2 – Confidence Interval for μ when σ is Known (p.8)

- The 68 – 96 – 99.7 Rule and Confidence Intervals (p.9)
- Confidence Interval for μ when σ is Known (Reprise) (p.14)

5.3 – Choice of Sample Size (p.26)

5.4 – Confidence Interval for μ when σ is Unknown (p.30)

5.5 – Confidence Interval for a Proportion (p.35)

Appendix – Summary (p.39)

5.1 – Statistical Inference

One of the goals of **statistical inference** is to draw conclusions about a **population** based on a random sample from the population.

Examples:

- Can we assess the reliability of a product's manufacturing process by randomly selecting a sample of the final product and determining how many of them are compliant according to some quality assessment scheme?
- Can we determine who will win an election by polling a small sample of respondents?

Specifically, we seek to estimate an unknown **parameter** θ , say, using a single quantity called the **point estimate** $\hat{\theta}$.

This point estimate is obtained using a **statistic**, which is simply a function of a random sample. The probability distribution of the statistic is its **sampling distribution**. Describing these is a main research avenue.

Example: consider a process that manufactures gear wheels (in some standard gauge). Let X be the random variable that records the weight of a randomly selected gear wheel. What is the population mean $\mu_X = E[X]$?

Solution: in the absence of $f(x)$, we can estimate $\mu = X$ with the help of a random sample X_1, \dots, X_n of gear wheel weight measurements, *via* the sample mean statistic:

$$\bar{X} = \frac{X_1 + \dots + X_n}{n}, \quad \text{which is } \approx \mathcal{N}(\mu, \sigma^2/n) \text{ according to C.L.T.}$$

Statistics

Examples of statistics include:

- sample mean and sample median
- sample variance and sample standard deviation
- sample quantiles (median, quartiles, percentiles)
- test statistics (t –statistics, χ^2 –statistics, f –statistics, etc.)
- order statistics (sample maximum and minimum, sample range, etc.)
- sample moments and functions thereof (skewness, kurtosis, etc.)

Estimator Variance and Standard Error

The **standard error** of a statistic is the **standard deviation of its sampling distribution**.

For instance, if observations X_1, \dots, X_n come from a population with **unknown mean** μ and **known variance** σ^2 , then $\text{Var}(\bar{X}) = \sigma^2/n$ and the **standard error of \bar{X}** is

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

If the variance of the original population is **unknown**, then it is estimated by the sample variance S^2 and the **estimated standard error of \bar{X}** is

$$\hat{\sigma}_{\bar{X}} = \frac{S}{\sqrt{n}}, \quad \text{where} \quad S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

Examples:

1. A sample of 20 baseball player heights (in inches) is shown below.

74, 74, 72, 72, 73, 69, 69, 71, 76, 71, 73, 73, 74, 74, 69, 70, 72, 73, 75, 78.

Let \bar{X} be the sampling mean of the heights. Then,

$$\bar{X} = \frac{X_1 + \cdots + X_{20}}{20} = 72.6$$

and the sample variance S^2 is

$$S^2 = \frac{1}{20 - 1} \sum_{i=1}^{20} (X_i - 72.6)^2 \approx 5.6211.$$

The standard error of \bar{X} is thus

$$\hat{\sigma}_{\bar{X}} = \frac{S}{\sqrt{20}} \approx \sqrt{\frac{5.6211}{20}} \approx 0.5301.$$

2. Consider a sample $\{X_1, \dots, X_{100}\}$ of independent observations selected from a normal population $\mathcal{N}(\mu, \sigma^2)$ where $\sigma = 50$ is known, but μ is not. What is the best estimate of μ ? What is the sampling distribution of that estimate?

Solution: the sample mean $\bar{X} = \frac{X_1 + \dots + X_{100}}{100}$ provides the best estimate of $\mu_X = \mu_{\bar{X}}$. The standard error of \bar{X} is $\sigma_{\bar{X}} = \frac{50}{\sqrt{100}} = 5$. Since the observations are sampled independently from a normal population with mean μ and standard deviation 50, $\bar{X} \sim \mathcal{N}(\mu, 5^2) = \mathcal{N}(\mu, 25)$, according to the CLT.

5.2 – C.I. for μ when σ is Known

Consider a sample $\{x_1, \dots, x_n\}$ from a normal population with **known** variance σ^2 and **unknown** mean μ . The sample mean

$$\bar{x} = \frac{x_1 + \dots + x_n}{n}$$

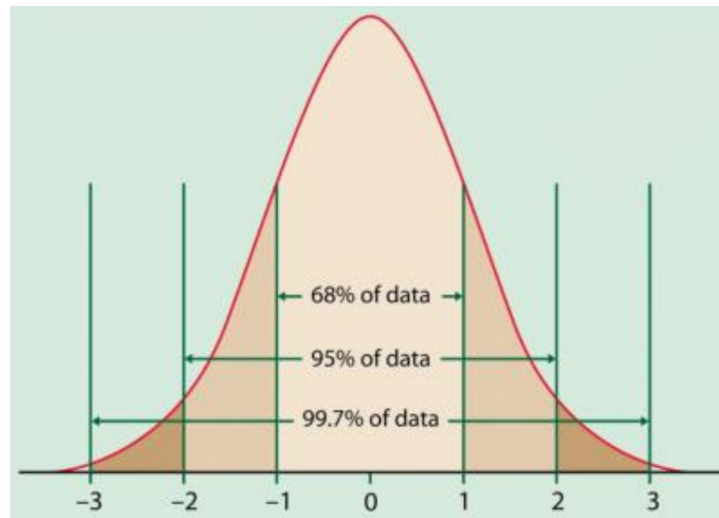
is a **point estimate** of μ .

Of course, this estimate is not exact, because \bar{x} is an observed value of \bar{X} ; it is unlikely that the observed value \bar{x} should coincide with μ .

We know that $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$, so that

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

The 68 – 96 – 99.7 Rule and Confidence Intervals



$$P(-1 < Z < 1) \approx 0.683$$

$$P(-2 < Z < 2) \approx 0.955$$

$$P(-3 < Z < 3) \approx 0.997.$$

Whenever we observe a sample mean \bar{X} from a normal population with mean μ , we would expect the inequality

$$-k < Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < k$$

to hold approximately

$$g(k) = \begin{cases} 68.3\% \text{ of the time} & \text{if } k = 1 \\ 95.5\% \text{ of the time} & \text{if } k = 2 \\ 99.7\% \text{ of the time} & \text{if } k = 3 \end{cases}$$

Equivalently, the **symmetric** $g(k)$ **confidence interval for** μ is

$$\bar{X} - k\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + k\frac{\sigma}{\sqrt{n}} \implies \bar{X} \pm k\frac{\sigma}{\sqrt{n}}.$$

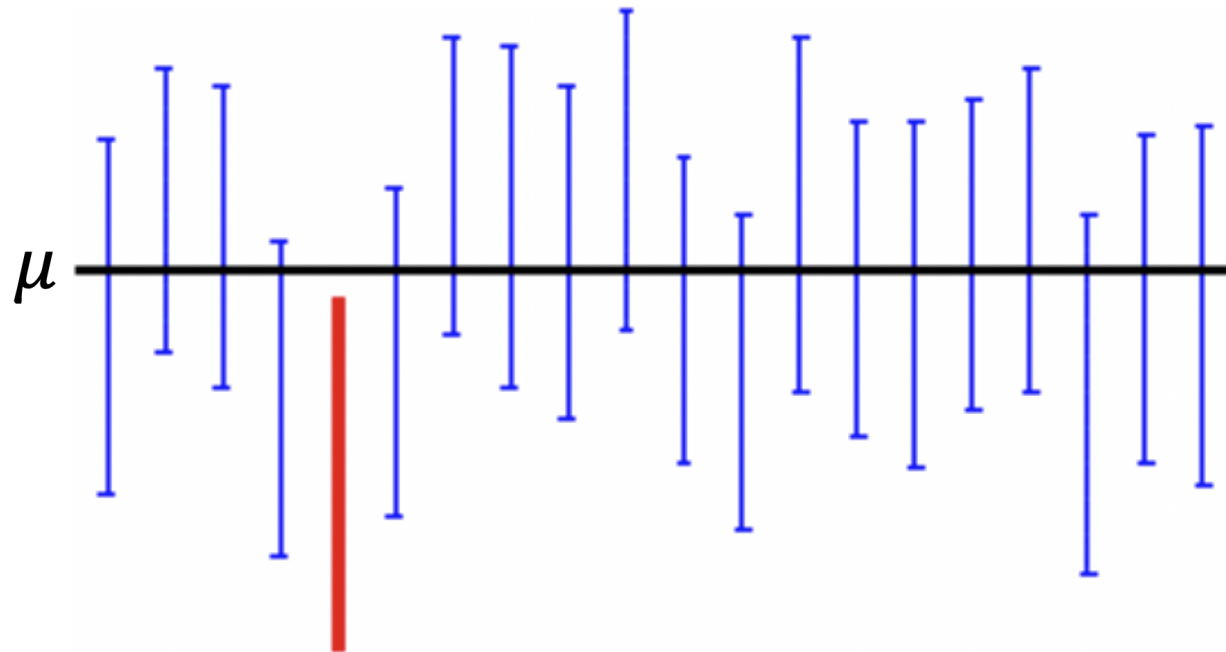
Examples:

1. Consider a sample $\{X_1, \dots, X_{64}\}$ from a normal population with standard deviation $\sigma = 72$ and unknown mean μ . The sample mean is $\bar{X} = 375.2$. Build a symmetric 68.3% confidence interval for μ .

Solution: according to the formula, the symmetric 68.3% confidence interval ($k = 1$) for μ in this situation is

$$375.2 \pm 1 \cdot \frac{72}{\sqrt{64}} \implies (375.2 - 9, 375.2 + 9) = (366.2, 384.2).$$

IMPORTANT: this does not say that we're 68.3% sure that the true μ is between 366.2 and 384.2. What it says is that when a sample of size 64 is taken from a normal population $\mathcal{N}(\mu, 72^2)$ and a symmetric 68.3% confidence interval for μ is built, μ will fall between the endpoints of the interval about 68.3% of the time.



A 95% C.I. indicates that we would expect 19 out of 20 samples from the same population to produce confidence intervals that contain the population parameter of interest, on average.

2. Build a symmetric 95.5% confidence interval for μ .

Solution: the same formula applies, with $k = 2$.

$$375.2 \pm 2 \cdot \frac{72}{\sqrt{64}} \implies (375.2 - 18, 375.2 + 18) = (357.2, 393.2).$$

3. Build a symmetric 99.7% confidence interval for μ .

Solution: the same formula applies, with $k = 3$.

$$375.2 \pm 3 \cdot \frac{72}{\sqrt{64}} \implies (375.2 - 27, 375.2 + 27) = (348.2, 402.2).$$

C.I. for μ when σ is Known (reprise)

Another approach to C.I. building is to specify the proportion of the area under $\phi(z)$ of interest, and then to determine the critical values (the endpoints) of the interval.

Let $\{X_1, \dots, X_n\}$ be drawn from $N(\mu, \sigma^2)$. Recall that $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$. For a **symmetric 95% confidence interval**, we need to find $z^* > 0$ such that $P(-z^* < Z < z^*) = 0.95$.

But the LHS can be re-written as

$$\begin{aligned} P(-z^* < Z < z^*) &= \Phi(z^*) - \Phi(-z^*) \\ &= \Phi(z^*) - (1 - \Phi(z^*)) = 2\Phi(z^*) - 1 \end{aligned}$$

So we are looking for z^* such that

$$0.95 = 2\Phi(z^*) - 1 \implies \Phi(z^*) = \frac{0.95 + 1}{2} = 0.975.$$

From the normal table, we see that $\Phi(1.96) \approx 0.9750$, so that

$$P(-1.96 < Z < 1.96) = P\left(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) \approx 0.95.$$

In other words, the inequality

$$-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96$$

holds with probability 0.95 (with the interpretation provided in Example 1).

Equivalently,

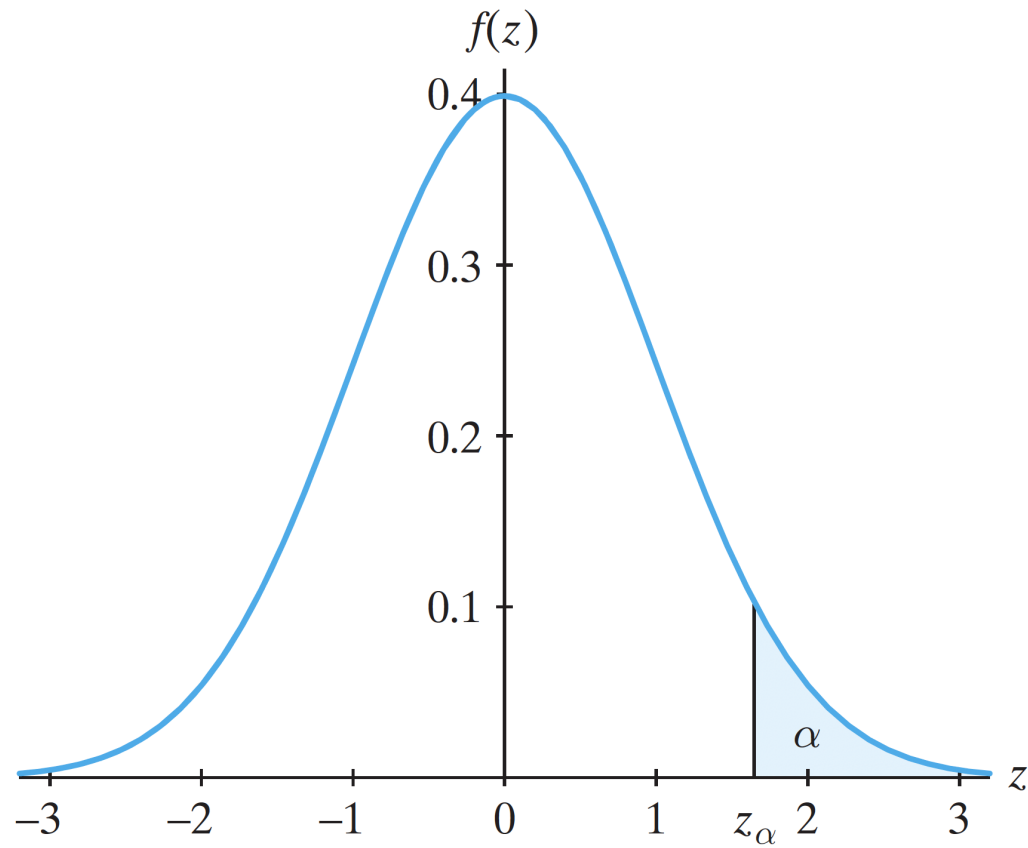
$$\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \implies \bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

is the **symmetric 95% confidence interval for μ when σ is known.**

A similar argument shows that

$$\bar{X} - 2.575 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 2.575 \frac{\sigma}{\sqrt{n}} \implies \bar{X} \pm 2.575 \frac{\sigma}{\sqrt{n}}$$

is the **symmetric 99% confidence interval for μ when σ is known.**



$$P(Z > z_{\alpha}) = \alpha$$

$$P(Z > z) = 1 - \Phi(z) = \Phi(-z)$$

The **confidence level** $1 - \alpha$ is usually expressed in terms of a **small** α , e.g. $\alpha = 0.05 \Rightarrow 1 - \alpha = 0.95$ confidence level.

For $\alpha = 0.01, 0.02, \dots, 0.98, 0.99$, the corresponding z_α are called the **percentiles** of the standard normal distribution. In general,

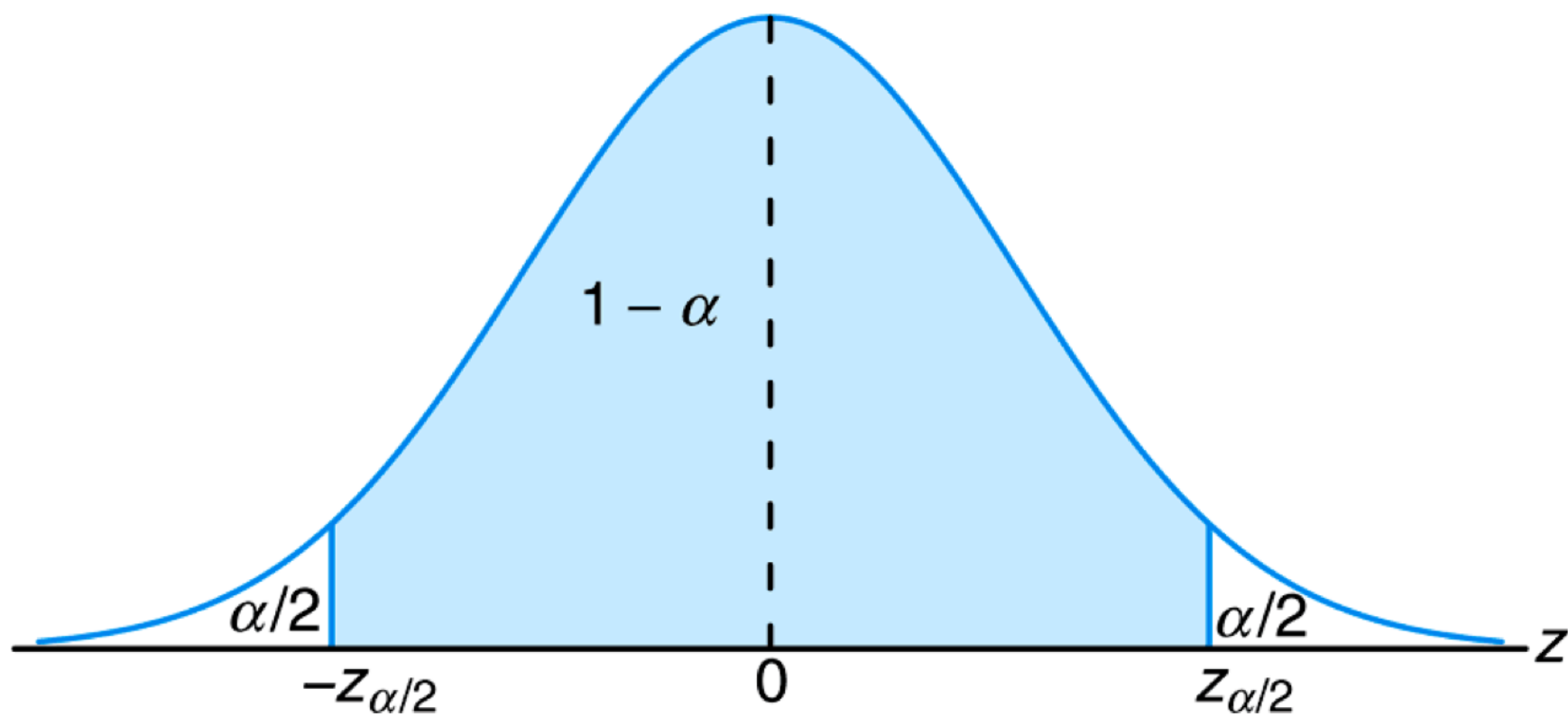
$$P(Z > z_\alpha) = \alpha \implies z_\alpha \text{ is the } 100(1 - \alpha) \text{ percentile.}$$

For 2–**sided confidence intervals**, the appropriate numbers are found by solving $P(|Z| > z^*) = \alpha$ for z^* . By the properties of $\mathcal{N}(0, 1)$,

$$\alpha = P(|Z| > z^*) = 1 - P(-z^* < Z < z^*) = 1 - (2\Phi(z^*) - 1) = 2(1 - \Phi(z^*)),$$

so that

$$\Phi(z^*) = 1 - \alpha/2 \implies z^* = z_{\alpha/2}.$$



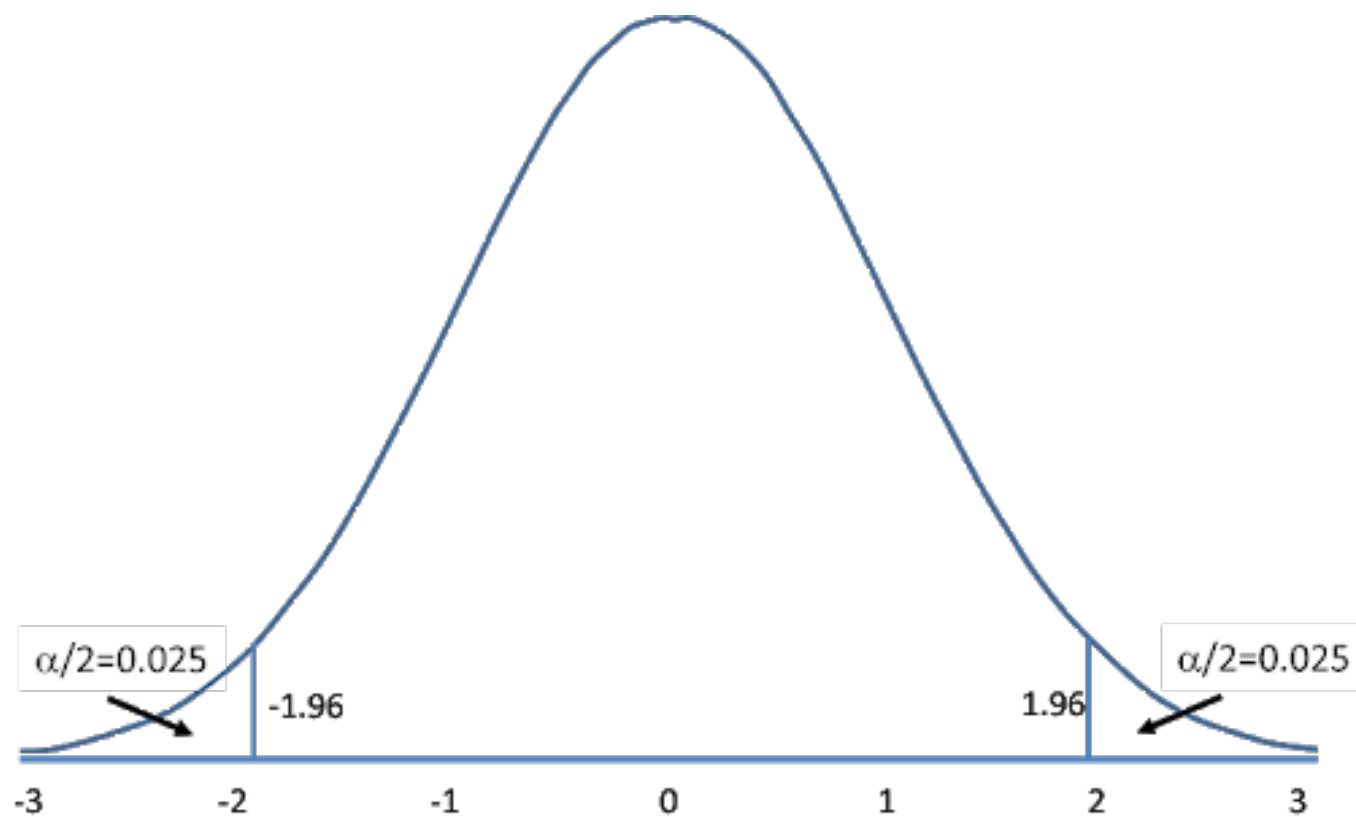
For instance,

$$P(|Z| > z_{0.025}) = 0.05 \implies z_{0.025} = 1.96$$

$$P(|Z| > z_{0.005}) = 0.01 \implies z_{0.005} = 2.575.$$

The symmetric $100(1 - \alpha)\%$ confidence interval can generally be written as

$$\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \implies \bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$



For a given confidence level α , **shorter confidence intervals are better** in relation to estimating the mean:

- estimates become better when the sample size n increases;
- estimates become better when σ decreases.

If $\alpha_1 > \alpha_2$, the $100(1 - \alpha_1)\%$ C.I. is smaller than the $100(1 - \alpha_2)\%$ C.I. (i.e. a 95% C.I. is always shorter than a 99% C.I.)

If the sample comes from a normal population, then the C.I. is **exact**. Otherwise, if n is large, we may use the CLT and get an **approximate** C.I.

Examples:

1. A sample of 9 observations from a normal population with known standard deviation $\sigma = 5$ yields a sample mean $\bar{X} = 19.93$. Provide a 95% and a 99% C.I. for the unknown population mean μ based on this sample.

Solution: the estimate of μ is $\bar{X} = 19.93$. The $100(1 - \alpha)\%$ confidence intervals are

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

$$95\% : \bar{X} \pm z_{0.025} \frac{\sigma}{\sqrt{n}} \Rightarrow 19.93 \pm 1.96 \frac{5}{\sqrt{9}} \Rightarrow 19.93 \pm 3.27 \text{ or } (16.66, 23.20)$$

$$99\% : \bar{X} \pm z_{0.005} \frac{\sigma}{\sqrt{n}} \Rightarrow 19.93 \pm 2.575 \frac{5}{\sqrt{9}} \Rightarrow 19.93 \pm 4.29 \text{ or } (15.64, 24.22)$$

2. A sample of 25 observations from a normal population with known standard deviation $\sigma = 5$ yields a sample mean $\bar{X} = 19.93$. Provide a 95% and a 99% C.I. for the unknown population mean μ based on this sample.

Solution: the estimate of μ is $\bar{X} = 19.93$. The $100(1 - \alpha)\%$ confidence intervals are

$$95\% : \bar{X} \pm z_{0.025} \frac{\sigma}{\sqrt{n}} \Rightarrow 19.93 \pm 1.96 \frac{5}{\sqrt{25}} \Rightarrow 19.93 \pm 1.96 \text{ or } (17.97, 21.89)$$

$$99\% : \bar{X} \pm z_{0.005} \frac{\sigma}{\sqrt{n}} \Rightarrow 19.93 \pm 2.575 \frac{5}{\sqrt{25}} \Rightarrow 19.93 \pm 2.58 \text{ or } (17.35, 22.51)$$

3. A sample of 25 observations from a normal population with known standard deviation $\sigma = 10$ yields a sample mean $\bar{X} = 19.93$. Provide a 95% and a 99% C.I. for the unknown population mean μ based on this sample.

Solution: the estimate of μ is $\bar{X} = 19.93$. The $100(1 - \alpha)\%$ confidence intervals are

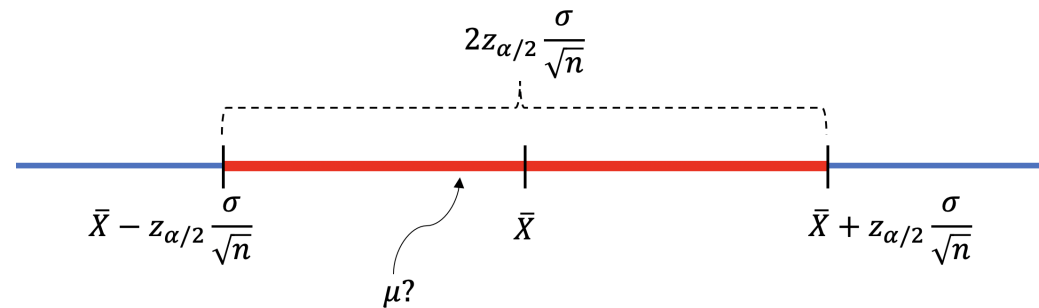
$$95\% : \bar{X} \pm z_{0.025} \frac{\sigma}{\sqrt{n}} \Rightarrow 19.93 \pm 1.96 \frac{10}{\sqrt{25}} \Rightarrow 19.93 \pm 3.92 \text{ or } (16.01, 23.85)$$

$$99\% : \bar{X} \pm z_{0.005} \frac{\sigma}{\sqrt{n}} \Rightarrow 19.93 \pm 2.575 \frac{10}{\sqrt{25}} \Rightarrow 19.93 \pm 5.15 \text{ or } (14.78, 25.08)$$

Note how the confidence intervals are affected by α , n , and σ .

5.3 – Choice of Sample Size

The **error** we commit by estimating μ via the sample mean \bar{X} is smaller than $z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$, with probability $100(1 - \alpha)\%$.



If we want to control the error, the only thing we can really do is control the sample size:

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

Examples:

1. A sample $\{X_1, \dots, X_n\}$ is selected from a normal population with standard deviation $\sigma = 100$. What sample size should be used to insure that the error on the population estimate is at most $E = 10$, at a confidence level $\alpha = 0.05$?

Solution: as long as

$$n > \left(\frac{z_{\alpha/2}\sigma}{E} \right)^2 = \left(\frac{z_{0.025} \cdot 100}{10} \right)^2 = (19.6)^2 = 384.16,$$

then the error committed by using \bar{X} to estimate μ will be at most 10, with 95% probability.

2. Repeat the first example, but with $\sigma = 10$.

Solution: we need

$$n > \left(\frac{z_{\alpha/2}\sigma}{E} \right)^2 = \left(\frac{z_{0.025} \cdot 10}{10} \right)^2 = (1.96)^2 = 3.8416.$$

3. Repeat the first example, but with $E = 1$.

Solution: we need

$$n > \left(\frac{z_{\alpha/2}\sigma}{E} \right)^2 = \left(\frac{z_{0.025} \cdot 100}{1} \right)^2 = (196)^2 = 38416.$$

4. Repeat the first example, but with $\alpha = 0.01$.

Solution: we need

$$n > \left(\frac{z_{\alpha/2}\sigma}{E} \right)^2 = \left(\frac{z_{0.005} \cdot 100}{10} \right)^2 = (25.75)^2 = 663.0625.$$

The relationship between α , σ , E , and n is not always intuitive!

5.4 – C.I. for μ when σ is Unknown

So far, we have been in the fortunate situation of sampling from a population with known variance σ^2 .

What do we do when the population variance is **unknown**?

We estimate σ using the **sample variance**

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

(remember that the true population mean μ is also unknown... that's what we're trying to find!) and the **sample standard deviation** $S = \sqrt{S^2}$.

If σ is known, we know from the CLT that $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ is approximately $\mathcal{N}(0, 1)$.

If σ is unknown, it can be shown that $\frac{\bar{X} - \mu}{S/\sqrt{n}}$ follows approximately $t(n - 1)$, the **Student T -distribution with $n - 1$ degrees of freedom**.

Consequently, for a confidence level α ,

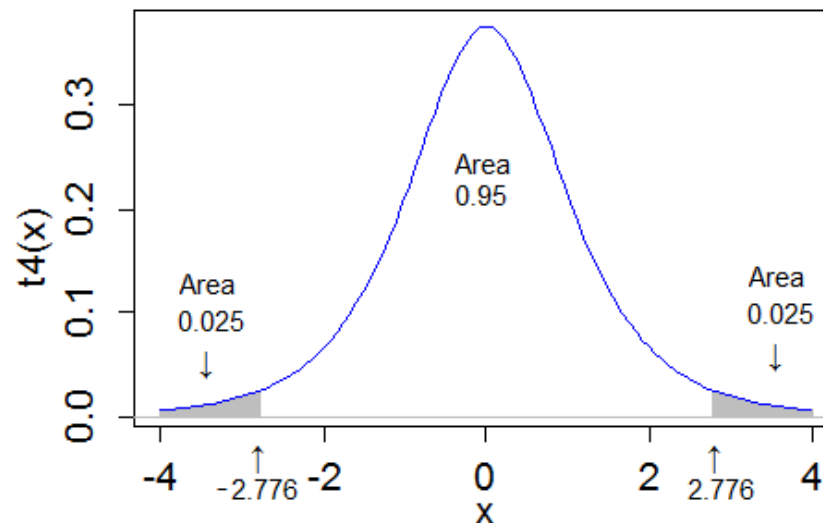
$$P \left(-t_{\alpha/2}(n - 1) < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{\alpha/2}(n - 1) \right) \approx 1 - \alpha,$$

where $t_{\alpha/2}(n - 1)$ is the $100(1 - \alpha/2)^{\text{th}}$ percentile of $t(n - 1)$ (these can be read from the table). Equality is reached if the underlying population is normal.

$$100(1 - \alpha)\% \text{ C.I. for } \mu : \bar{X} \pm t_{\alpha/2}(n - 1) \frac{S}{\sqrt{n}}.$$

For instance, if $\alpha = 0.05$ and $\{X_1, X_2, X_3, X_4, X_5\}$ are samples from a normal distribution with unknown mean μ and unknown variance σ^2 , then

$$t_{0.025}(5 - 1) = 2.776 \quad \text{and} \quad P\left(-2.776 < \frac{\bar{X} - \mu}{S/\sqrt{5}} < 2.776\right) = 0.95.$$



Examples:

1. For a given year, 9 measurements of ozone concentration are obtained:

3.5 5.1 6.6 6.0 4.2 4.4 5.3 5.6 4.4

Assume that the measured ozone concentrations follow a normal distribution with variance $\sigma^2 = 1.21$, build a 95% C.I. for the population mean μ . Note that $\bar{X} = 5.01$ and that $S = 0.97$.

Solution: since we know the variance, we need to use the standard normal percentile $z_{\alpha/2} = z_{0.025} = 1.96$:

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 5.01 \pm 1.96 \frac{\sqrt{1.21}}{\sqrt{9}} = 5.01 \pm 0.72 \text{ or } (4.29, 5.73).$$

2. Same thing, but assume that the variance of the underlying population is unknown.

Solution: since we do not know the variance, we need to use the Student percentile $t_{\alpha/2}(n - 1) = t_{0.025}(8) = 2.306$ (make sure you understand how to get this value from the table):

$$\bar{X} \pm t_{\alpha/2}(n - 1) \frac{S}{\sqrt{n}} = 5.01 \pm 2.306 \frac{0.97}{\sqrt{9}} \text{ or } (4.26, 5.76).$$

The 95% C.I. when we know the variance is **tighter** (smaller), which is natural as we are more confident about our results when we have more information.

5.5 – C.I. for a Proportion

If $X \sim \mathcal{B}(n, p)$ (number of successes in n trials), then the point estimator for p is $\hat{P} = \frac{X}{n}$.

Recall that $E[X] = np$ and $\text{Var}[X] = np(1 - p)$.

We can standardize any random variable:

$$Z = \frac{X - \mu}{\sigma} = \frac{n\hat{P} - np}{\sqrt{np(1 - p)}} = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

is approximately $\mathcal{N}(0, 1)$.

Thus, for sufficiently large n ,

$$P \left(-z_{\alpha/2} < \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} < z_{\alpha/2} \right) \approx 1 - \alpha.$$

Using the previous approach, an **approximate** $100(1 - \alpha)\%$ C.I. for p is:

$$\hat{P} - z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} < p < \hat{P} + z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}},$$

but this is not really useful because we don't actually know p ! Instead:

$$\hat{P} - z_{\alpha/2} \sqrt{\frac{\hat{P}(1 - \hat{P})}{n}} < p < \hat{P} + z_{\alpha/2} \sqrt{\frac{\hat{P}(1 - \hat{P})}{n}}.$$

Examples:

1. Two candidates (A and B) are running for office. A poll is conducted: 1000 voters are selected randomly and asked for their preference: 52% support A , while 48% support their rival, B . Provide a 95% C.I. for the support of each candidate.

Solution: we use $\alpha = 0.05$ and $\hat{P} = 0.52$. The 95% C.I. for A is

$$\hat{P} \pm z_{\alpha/2} \sqrt{\frac{\hat{P}(1 - \hat{P})}{n}} = 0.52 \pm 1.96 \sqrt{\frac{0.52 \cdot 0.48}{1000}} \approx 0.52 \pm 0.031.$$

The 95% C.I. for B is 0.48 ± 0.031 .

2. On the strength of this polling result, a newspaper prints the following headline: “Candidate A Leads Candidate B !” Is the headline warranted?

Solution: although there is a 4–point gap in the poll numbers, the true support for candidate A is in the 48.9% – 55.1% range, and, the true support for candidate B is in the 44.9% – 51.1% range, with probability 95% (that is to say, 19 times out of 20).

Since there is overlap in the confidence intervals, the race is more likely to be a dead heat.

Appendix – Summary

Sample: $\{X_1, \dots, X_n\}$. **Objective:** predict μ with confidence level α .

- If population is **normal** with **known** variance σ^2 , the **exact** $100(1 - \alpha)\%$ C.I. is

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

- If population is **non-normal** with **known** variance σ^2 and n is '**big**', the **approximate** $100(1 - \alpha)\%$ C.I. is

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

- If population is **normal** with **unknown** variance, the **exact** $100(1 - \alpha)\%$ C.I. is

$$\bar{X} \pm t_{\alpha/2}(n - 1) \frac{S}{\sqrt{n}}.$$

- If population has **unknown** variance and n is '**big**', the **approximate** $100(1 - \alpha)\%$ C.I. is

$$\bar{X} \pm z_{\alpha/2} \frac{S}{\sqrt{n}}.$$

- If population has **unknown** variance and n is '**small**', you are S.O.O.L.