7 - Confidence Intervals

The aim in this lecture is to estimate a population parameter based on sample data.

	Sample Statistic	Population Parameter
proportion	$\hat{p}$	p
mean	$\bar{X}$	μ
standard deviation	s	σ^2
difference of means	$\bar{X}_1 - \bar{X}_2$	$\mu_1 - \mu_2$
mean paired difference	$\bar{d}$	μ_d

Definition A *point estimator* $\hat{\theta}$ is a formula that applies to a set of sample data to get the point estimate of a parameter θ. The resulting value is a *point estimate*.

In this course, the important point estimators are:

Point Estimator	Population Parameter
$\bar{X} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$	$\mu =$ population mean
$s^2 = \dfrac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n-1}$	$\sigma^2 =$ population variance
$\hat{p} = \frac{k}{n}$ (where $k =$ number of "successes")	$p =$ population proportion
$\bar{X}_1 - \bar{X}_2$ (based on samples from two populations)	$\mu_1 - \mu_2 =$ difference of two population means
$\hat{p}_1 - \hat{p}_2$ (based on samples from two populations)	$p_1 - p_2 =$ difference of two population proportions

Example Suppose you randomly select $n = 10$ BCIT students. For each student, you record the variable $X =$ Age, with the resulting data set:

21	20	20	28	42
31	19	20	18	25

The sample mean is: $\bar{X} = \frac{\Sigma X}{n} = \frac{24}{10} = 24.4$

Our *point estimate* of μ (population mean) is 24.4.

Note: There is no guarantee that a point estimate is correct. In fact, we expect there to be a random error associated with any point estimate:

$$\theta = \hat{\theta} + \text{error}$$

Fortunately, the Central Limit Theorem gives us a way to estimate the *size* of the error. We can then state our estimate of θ as an *interval* rather than only a single point.

What is a Confidence Interval?

A confidence interval is an *interval estimate* $(\hat{\theta} - E, \hat{\theta} + E)$ of a population parameter, θ.

The point estimate $\hat{\theta}$ is the center of the interval; the margin of error, E, extends to the ends of the confidence interval. We believe that θ lies *within* the interval with a certain level of *confidence*.

Confidence Intervals for μ

Before getting into complex details, let's see how this works in practice.

Example Suppose you randomly select $n = 100$ BCIT students. Sample data for the variable $X =$ age is shown.

26,	31,	26,	17,	26,	22,	22,	26,	31,	18,
30,	20,	30,	23,	28,	25,	26,	23,	24,	26,
33,	35,	27,	29,	25,	31,	25,	27,	22,	34,
17,	31,	22,	27,	25,	22,	17,	24,	27,	31,
32,	28,	27,	22,	26,	27,	27,	23,	21,	23,
31,	20,	25,	28,	28,	21,	23,	26,	25,	27,
18,	26,	27,	32,	26,	22,	23,	23,	26,	29,
28,	18,	17,	20,	29,	24,	28,	27,	26,	17,
25,	21,	26,	32,	32,	24,	28,	24,	25,	33,
31,	22,	25,	23,	26,	24,	25,	22,	30,	27

From the sample data, you calculate sample statistics

$$n = 100$$

$$\bar{X} = 25.52$$

$$s^2 = 16.9996$$

From this we calculate a 95% *confidence interval* as follows:

lower limit $= \bar{X} - 1.984 \times \frac{s}{\sqrt{n}} =$

upper limit $= \bar{X} - 1.984 \times \frac{s}{\sqrt{n}} =$

(The number 1.9840 is a *critical t-score*.)

Conclusion: we are 95% confident that

Using Z table (σ is known)

Confidence intervals for the population mean μ of a variable X come from:

- $\mu_{\bar{X}} = \mu$
- $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

- $\bar{X}$ is normally distributed if either:
  - X is itself a normal variable, or
  - n is large enough (typically $n \geq 30$)

These points imply that the Z-score for $\bar{X}$ follows a *standard normal distribution*

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

There is therefore a 95% probability that

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

Solving for μ tells us that

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

Note: This form of the 95% confidence interval requires knowing σ (for the underlying variable X). It is unrealistic to imagine that we would know σ when we don't know μ.

Example Let $X =$ the amount of time (in months) a person uses their phone before replacing it. Suppose we do not know μ but, somehow, we know that $\sigma = 13.0$ months.

Find a 95% confidence interval for μ.

1. Collect X data for a random sample of size $n = 50$.

2. Calculate sample statistics:

$$\bar{X} = 32.5$$
$$s = 13.2$$

   For this question, we assume that $\boldsymbol{\sigma = 13.0}$.

3. Calculate the margin of error:

$$E =$$

4. Calculate the limits of the confidence interval:

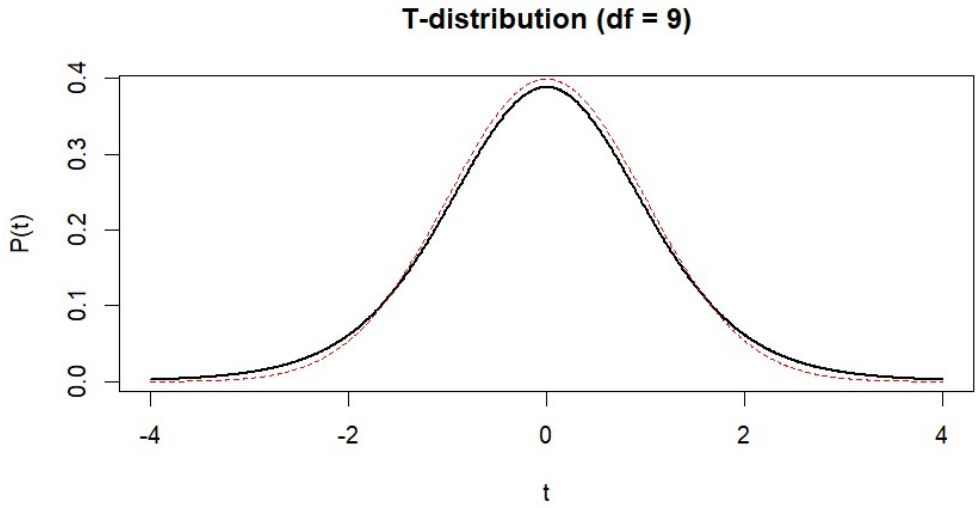
5. Conclusion:

Using T table (σ is unknown)

In realistic scenarios, we do not know σ. We then work with the t-statistic:

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

The variable T follows *Student's t-distribution* T_{n-1} with $n-1$ "degrees of freedom". The distribution T_{n-1} is close to a standard normal but contains more probability in the "tails".



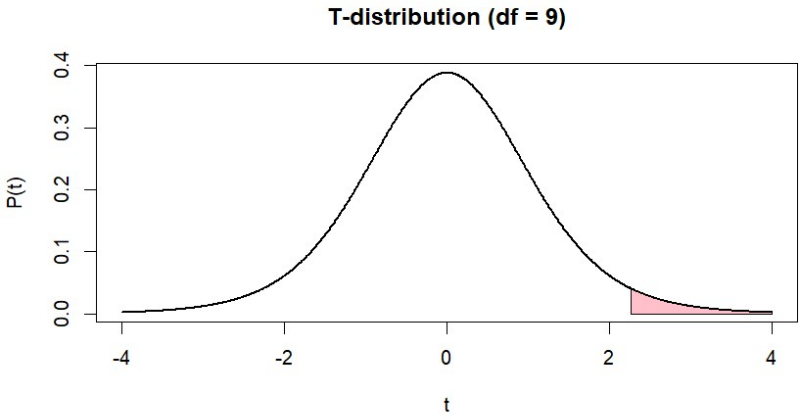
Critical t-Values

For confidence intervals, we frequently need the *critical t-value*

$t_{\alpha/2}$ = the value such that $P(T > t_{\alpha/2}) = \frac{\alpha}{2}$

We calculate critical t-values using:

- Students' t-table, *or*
- the R function:
  `qt(1-alpha/2, df = n-1)`



Example Namzor Inc. wishes to test the ability of its hard drives to withstand high temperatures. To keep costs down, 10 hard drives are randomly selected and tested. The failure temperatures appear to follow a normal distribution. From the sample, we get:

$$\bar{X} = 50.0\ ^\circ\mathrm{C}$$

$$s = 3.0\ ^\circ\mathrm{C}$$

a. Construct a 95% confidence interval for μ, the population mean of failure temperatures.

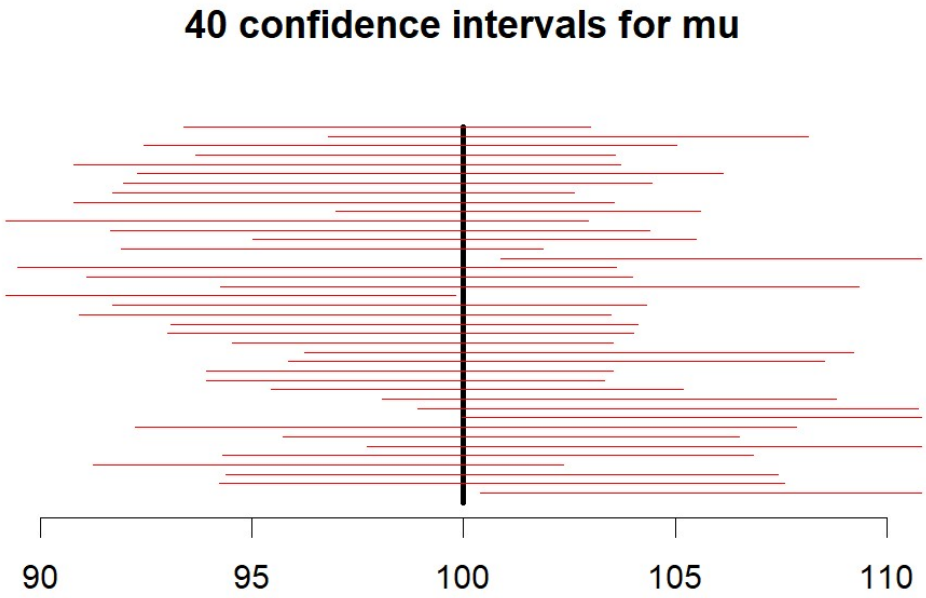
b. Construct a 90% confidence interval for the population mean.

What Do We Mean by 95% Confidence?

We know that there is a 95% chance that $\bar{X}$ lies within the margin of error E, relative to μ.

However, it does *not* make sense to say that there is a 95% chance that μ lies within the confidence interval $(\bar{X} - E, \bar{X} + E)$. Why?

It is the confidence interval that is random, not μ!



What we *can* say:

If we collect data from a random sample, there is a 95% chance that the resulting confidence interval contains the true μ.

In other words, if we could produce confidence intervals for *many* different random samples, we would find μ lies within the resulting confidence interval for 95% of samples.

To keep it simple, we say:

> "We are 95% confident that μ lies in the interval."

Example Accessing disk storage is slow compared to accessing internal memory (RAM). For a particular system, 35 measurements are taken of the time to access disk storage. A mean of 0.0293 seconds and a standard deviation of 0.0032 were determined.

a. Find a 95% confidence interval for the true average disk access time, μ.

b. If we mistakenly used a critical Z value (1.96) instead of T, what would the 95% confidence interval be?

Determining Sample Size

We have built a confidence interval based on known data. Suppose we haven't yet collected the sample. We would then be interested in determining what sample size n is necessary to yield a certain margin of error in our estimate of the population parameter.

Starting with

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

we can solve for n, obtaining:

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

Example (Determining Sample Size) Find the sample size needed to find an estimate for the mean age of computers at BCIT. Assume a 95% degree of confidence that the sample mean will be in error by no more than 0.25 years. In a previous study the standard deviation for BCIT computer age was $\sigma = 0.5$ years.