Statistics for Computing MA4413

Lecture 13

Statistical Estimation: Confidence Intervals

Kevin Burke

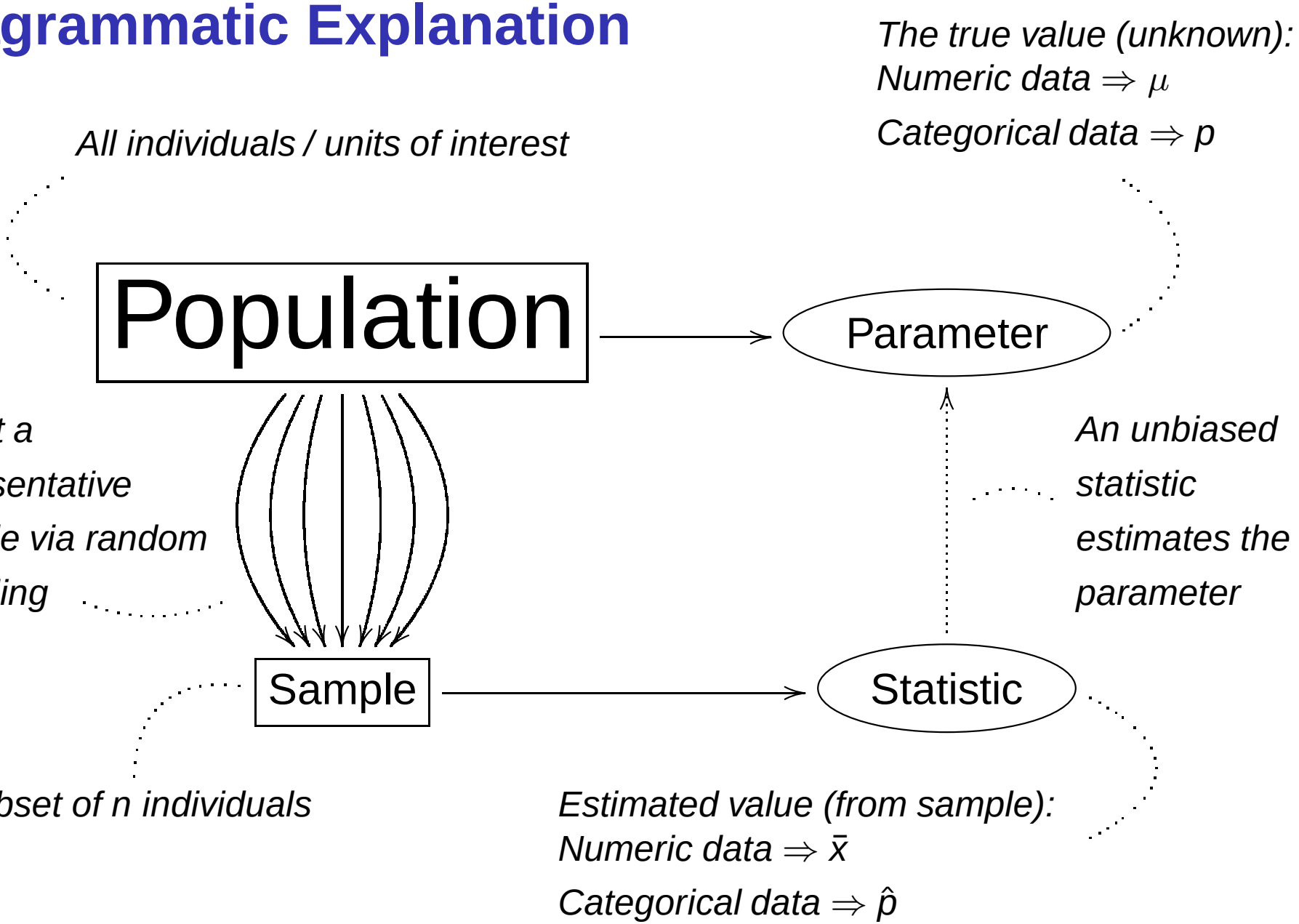
kevin.burke@ul.ie

Statistics Recap

- **Population**: *all* individuals / units of interest.
- **Parameter**: the particular *feature* of the population that we are interested in.
  - **Mean**: $\mu =$ unknown (*numeric* data).
  - **Proportion**: $p =$ unknown (*categorical* data).
- **Sample**: a set of n individuals, selected via *random sampling*, which informs us about the population.
- **Statistic**: the feature of interest calculated for the sample of data.
  - **Sample Mean**: $\bar{x}$ provides an estimate of μ.
  - **Sample Proportion**: $\hat{p}$ provides an estimate of p.

Note that for numeric data we also have the *population standard deviation*, σ, and the *sample standard deviation*, s.

Diagrammatic Explanation



Example: Android Application

Developers of an Android application wish to know the average age of their users for the purposes of advertising. They contact 100 users (selected via random sampling) and enquire about their age. Of these 100 users, 68 respond; the average age is found to be 22.54 in this sample.

- **Population**: *all* users.
- **Parameter**: The true mean age of all users; $\mu =$ unknown (must be estimated).
- **Sample**: The $n = 68$ users whose information we have.
- **Statistic**: The statistic is $\bar{x} = 22.54$ calculated based on our sample of 68 users. This is used as an estimate of μ.

Point Estimate Vs Interval Estimate

In the example just covered, the sample mean $\bar{x} = 22.54$ provides a **point estimate** for the true population mean μ.

However, an **interval estimate** is more useful, i.e., a *range* of plausible values for μ based on the sample.

For example, a statement such as "we are 95% confident that μ is contained in the interval [21.83, 23.25]" is much more useful than a single point estimate.

Interval estimates take statistical variation into account and allow us to *test hypotheses* about the true value μ.

The Central Limit Theorem

The *central limit theorem* tells us that

$$\bar{X} \sim \text{Normal}\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

Based on this fact, we know that 95% of $\bar{x}$ values (calculated using different samples) would lie in the range $\mu \pm 1.96\,\frac{\sigma}{\sqrt{n}}$.

Mathematical notation: $\Pr\left(\mu - 1.96\,\frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + 1.96\,\frac{\sigma}{\sqrt{n}}\right) = 0.95.$

(note: we usually only have *one* sample in practice, but the above result allows us to develop further theory)

Manipulating the 95% Limits

In 95% of samples $\bar{X} \in [\mu - 1.96\,\frac{\sigma}{\sqrt{n}},\ \mu + 1.96\,\frac{\sigma}{\sqrt{n}}]$,

i.e., $\bar{X}$ is within $\pm 1.96\,\frac{\sigma}{\sqrt{n}}$ units of μ.

If $\bar{X}$ is within $\pm 1.96\,\frac{\sigma}{\sqrt{n}}$ units of μ, then it is also true to say

that μ is within $\pm 1.96\,\frac{\sigma}{\sqrt{n}}$ units of $\bar{X}$.

$\Rightarrow$ 95% of the time $\mu \in [\bar{X} - 1.96\,\frac{\sigma}{\sqrt{n}},\ \bar{X} + 1.96\,\frac{\sigma}{\sqrt{n}}]$.

95% Confidence Interval

The interval $\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$ is *random*, i.e., it varies from sample to sample (since $\bar{X}$ varies).

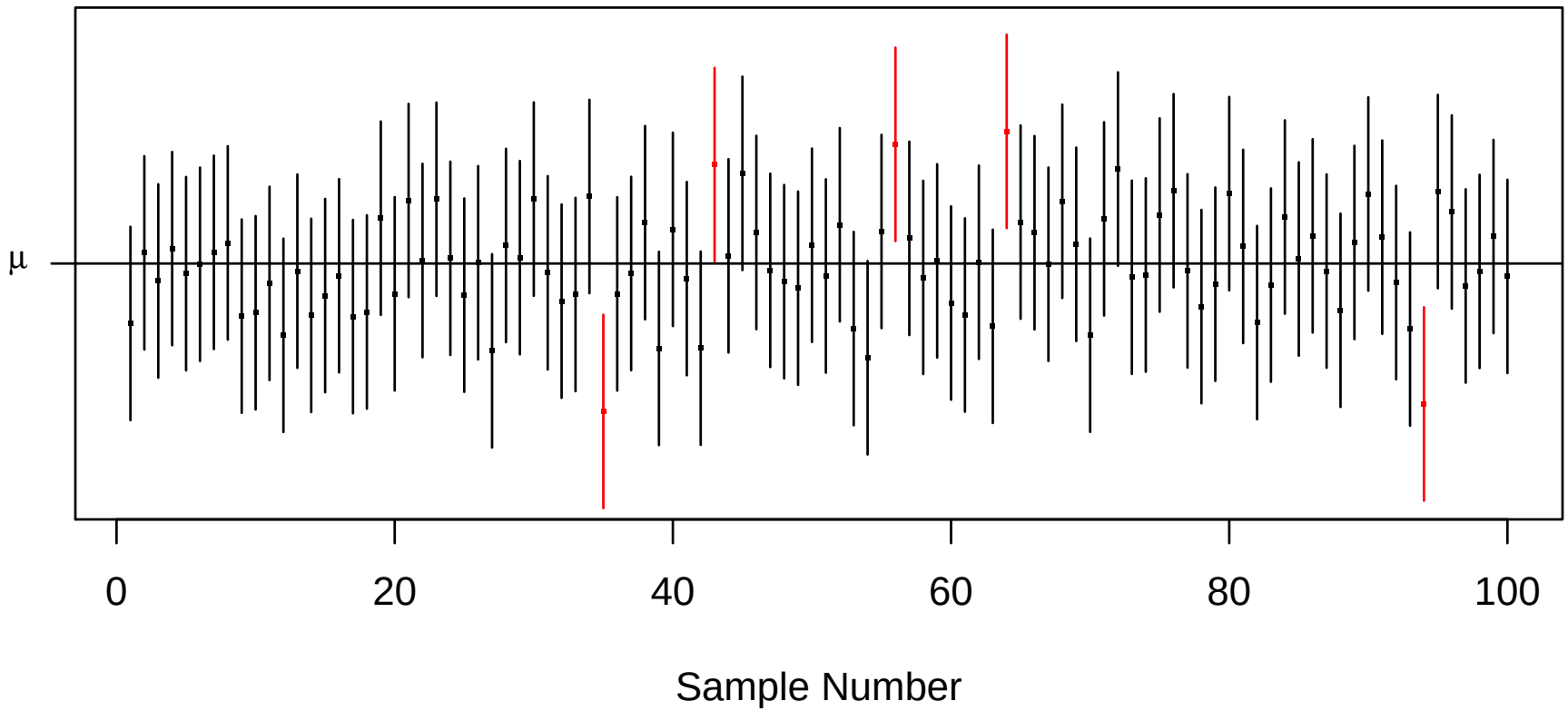
Typically we have *one* sample of data and, hence, a specific sample mean value, $\bar{x}$. The calculated interval

$$\boxed{\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}}$$

is called a **95% confidence interval**.

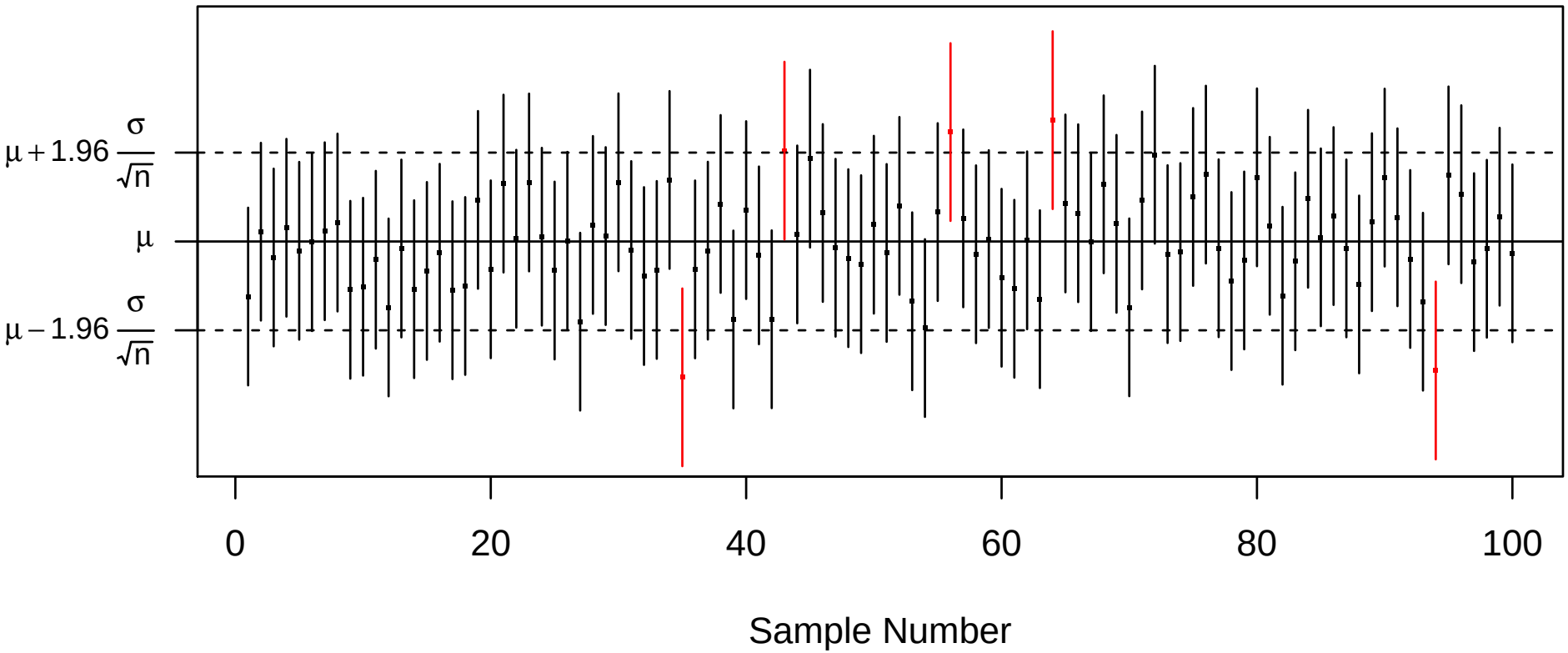
(note: if we had multiple samples of data, the true mean, μ, will be contained in 95% of the calculated intervals)

95% Confidence Intervals: Multiple Samples



- Sample means, $\bar{x}$, (dots) and 95% confidence intervals, $\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$, (vertical lines) calculated in multiple samples. Those which do not include μ are highlighted in red.

95% Confidence Intervals: Multiple Samples



- 95% limits around μ also shown. Note that if $\bar{x} \in \mu \pm 1.96\frac{\sigma}{\sqrt{n}}$, then $\mu \in \bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$. In other words, μ and $\bar{x}$ are within $\pm 1.96\frac{\sigma}{\sqrt{n}}$ units of each other in these cases (as developed in slide 7).

Terminology

Once a confidence interval is calculated, the phrase used is:

"We are 95% confident that μ is contained in this interval".

Example:

Let's assume [21.83, 23.25] is a 95% confidence interval for μ calculated using a sample of data.

Note that μ is a constant - *it does not vary*. Therefore, it is either in the calculated interval or it is not.

However, from the previous slides, we know that there is a 95% chance that it *is* in the calculated interval, i.e., we are 95% confident that it is.

$(1-\alpha)100\%$ Confidence Interval

For 95% confidence intervals we have $\alpha = 0.05 \Rightarrow \alpha/2 = 0.025$ and, hence, $z_{0.025} = 1.96$.

Naturally, the previous developments extend to other levels of confidence:

$$\boxed{\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}}$$

gives us a **$(1-\alpha)$100% confidence interval**.

Confidence Intervals: General Form

Note that the confidence interval for μ is $\bar{x} \pm z_{\alpha/2}\,\sigma(\overline{X})$ where $\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}}$ is the *standard error* of $\overline{X}$.

The general form of *all* confidence intervals that we deal with is

$$\boxed{\text{statistic} \;\pm\; z_{\alpha/2}\;\sigma(\text{statistic})}$$

where $\sigma(\text{statistic})$ is the standard error of the statistic in question.

We can be $(1-\alpha)100\%$ confident that the parameter lies in the calculated interval.

Confidence Intervals: General Form

All confidence intervals will be based on the table below.

parameter	statistic	σ(statistic)
μ	$\bar{x}$	$\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}}$
p	$\hat{p}$	$\sigma(\widehat{P}) = \sqrt{\frac{p\,(1-p)}{n}}$
$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$\sigma(\overline{X}_1 - \overline{X}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	$\sigma(\widehat{P}_1 - \widehat{P}_2) = \sqrt{\frac{p_1\,(1-p_1)}{n_1} + \frac{p_2\,(1-p_2)}{n_2}}$

For example, $(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ is a confidence interval for the true difference in means of two groups, $\mu_1 - \mu_2$.

Estimating Standard Error

An important issue that we have avoided until now is the fact that

$$\sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

requires the value $\sigma =$ **unknown** (i.e., population standard deviation).

In practice, we substitute s, the sample standard deviation, in place of σ in the above formula. Thus, the **estimated standard error** is

$$\boxed{s(\bar{X}) = \frac{s}{\sqrt{n}}}$$

where the notation $s(\bar{X})$ makes it clear that we have estimated the true standard error $\sigma(\bar{X})$.

Estimating Standard Error

From the central limit theorem we know that

$$\bar{X} \sim \text{Normal}\left(\mu,\ \sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}}\right).$$

In practice we approximate the above normal distribution by substituting s in place of σ

$$\Rightarrow \text{Normal}\left(\mu,\ s(\bar{X}) = \frac{s}{\sqrt{n}}\right).$$

This only works well for large samples. Generally $\boxed{\boldsymbol{n > 30}}$ is considered to be large enough.

Confidence Intervals in Practice

In practice we cannot use the standard errors in the table on slide 14. We use the *estimated* standard errors below.

Note: we need $n > 30$ and, for two groups, both $n_1 > 30$ and $n_2 > 30$.

parameter	statistic	$s(\text{statistic})$
μ	$\bar{x}$	$s(\overline{X}) = \frac{s}{\sqrt{n}}$
p	$\hat{p}$	$s(\widehat{P}) = \sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}$
$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$s(\overline{X}_1 - \overline{X}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	$s(\widehat{P}_1 - \widehat{P}_2) = \sqrt{\frac{\hat{p}_1\,(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2\,(1-\hat{p}_2)}{n_2}}$

Confidence Intervals in Practice

Table of $(1-\alpha)100\%$ confidence intervals that we will calculate:

parameter	confidence interval
	statistic $\pm$ $z_{\alpha/2}$ **s(statistic)**
μ	$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$
p	$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
$\mu_1 - \mu_2$	$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
$p_1 - p_2$	$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

Example: Gameplay

A games developer believes that their newest game has 16 hours of gameplay. Based on a sample of 33 individuals, it was found that the average time to complete the game was 15.4 hours and the standard deviation was 2.3 hours.

A 95% confidence interval for μ is given by

$$\bar{x} \pm z_{0.025} \frac{s}{\sqrt{n}} \quad \Rightarrow \quad 15.4 \pm 1.96 \frac{2.3}{\sqrt{33}}$$

$$15.4 \pm 1.96\,(0.4004)$$

$$15.4 \pm 0.7847$$

$$[14.62, 16.18]$$

We are 95% confident that the true mean lies in the above interval which clearly supports a value of $\mu = 16$ hours.

Example: Laptop Brand

Let “Brand-A” be a particular laptop brand. We wish to find out if there is a difference in the proportions of computer science students and business students who use this brand.

A group of 130 computer science students and 150 business students were asked if they use Brand-A laptops; 95 computer science students answered yes and 103 business students answered yes.

Firstly, we must calculate $\hat{p}_1 = \frac{95}{130} = 0.731$ and $\hat{p}_2 = \frac{103}{150} = 0.687$.

Example: Laptop Brand

Let's assume that we wish to calculate a 99% confidence interval for $p_1 - p_2$:

$$(\hat{p}_1 - \hat{p}_2) \pm z_{0.005} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$(0.731 - 0.687) \pm 2.58 \sqrt{\frac{0.731(0.269)}{130} + \frac{0.687(0.313)}{150}}$$

$$0.044 \pm 2.58\,(0.0543)$$

$$0.044 \pm 0.1401$$

$$[-0.096, 0.184]$$

As this interval includes the value zero, it looks like there is no difference in the proportions from a statistical point of view.

Question 1

A manufacturer of CPUs believes that their new model is 2 seconds faster than the previous model in completing some benchmark task. They tested 50 of each CPU type (i.e., $n_1 = n_2 = 50$). The results (in seconds) are summarised in the table below.

New	Old
$\bar{x}_1 = 8.4$	$\bar{x}_2 = 11.7$
$s_1 = 1.5$	$s_2 = 0.8$

a) Calculate a 95% confidence interval for the difference in means.

b) Based on this interval, is the manufacturer's belief justified?

Confidence Level

You may wonder how the confidence level should be chosen.

In particular, you may enquire as to why a 95% level would be chosen over a 99% level.

Intuitively it might seem that more confidence is better; *this is not the whole story*.

Example: Gameplay

Note that, in the gameplay example, a 95% confidence interval was calculated.

$$15.4 \pm 1.96\,(0.4004) \quad \Rightarrow \quad [14.62,\ 16.18]$$

We also could calculate a 90% confidence interval

$$15.4 \pm 1.64\,(0.4004) \quad \Rightarrow \quad [14.74,\ 16.06]$$

or a 99% confidence interval

$$15.4 \pm 2.58\,(0.4004) \quad \Rightarrow \quad [14.36,\ 16.43]$$

Confidence Level

We can see that

- Reduced confidence $\Rightarrow$ interval narrows.
- Increased confidence $\Rightarrow$ interval widens.

This makes sense:

In order to be more confident that the true parameter is contained in the interval, we must increase the width of the interval.

Conversely, we may wish to provide a narrower interval but we will then be less confident.

Error Probability

The confidence level, $1 - \alpha$, is the probability that the interval contains the parameter.

The remainder, α, is the **error probability**: the probability that the interval *does not* contain the parameter.

Recall that the general form of a confidence interval is

$$\text{statistic} \pm z_{\alpha/2}\, s(\text{statistic})$$

where $s(\text{statistic})$ is the standard error.

Decreasing α (and hence $\alpha/2$), increases the $z_{\alpha/2}$ value which leads to a wider interval $\Rightarrow$ less error $\Rightarrow$ greater confidence.

Extremes of Confidence: 0%

0% confidence interval $\Rightarrow \alpha = 1$ remaining $\Rightarrow \alpha/2 = 0.5$ in each tail.

$$\text{statistic} \pm z_{0.5} \times s(\text{statistic})$$
$$\text{statistic} \pm 0 \times s(\text{statistic})$$
$$\text{statistic}$$

In this case we claim that the true parameter has *exactly* the same value as the calculated statistic.

Of course, this is destined to be incorrect ($\alpha = 1$) as it is highly unlikely that any randomly selected sample has a statistic which reproduces the parameter value exactly.

Extremes of Confidence: 100%

100% confidence interval $\Rightarrow \alpha = 0$ remaining $\Rightarrow \alpha/2 = 0$ in each tail.

$$\text{statistic} \pm z_0 \times s(\text{statistic})$$
$$\text{statistic} \pm \infty \times s(\text{statistic})$$
$$[-\infty, \infty]$$

In this case we are stating that the parameter value is somewhere between $-\infty$ and ∞.

Obviously this is guaranteed to be correct ($\alpha = 0$) but it should be clear that this interval contains no useful information. We know that all parameters will be somewhere in $[-\infty, \infty]$ already.

Choosing Confidence Level

We must choose a level of confidence between 0% and 100%.

- Greater confidence comes at the expense of a wider interval.
- A narrower interval comes at the expense of being less confident.

The 95% level is the most commonly used, i.e., 5% chance of error.

(as mentioned previously, 90% and 99% are also common)

Naturally we must be willing to accept *some* probability of error and, in practice, this will be chosen based on the costs associated with making an error.

Sample Size

Recall (from slide 17) that *all* standard errors can be reduced by increasing the sample size, n. Example: $s(\overline{X}) = \frac{s}{\sqrt{n}}$.

Thus, by using a larger sample, the confidence interval

$$\text{statistic} \pm z_{\alpha/2}\, s(\text{statistic})$$

can be narrowed *without* sacrificing the desired level of confidence.