# Estimation

Statistical Thinking (ETC2420 / ETC5242)

Week 3, Semester 2, 2025

# Outline

# Outline

# Learning goals for Week 3

- Introduce the main elements of statistical inference and estimation, especially the idea of a sampling distribution.
- Show the simplest type of estimation: that of a single parameter.
- Introduce the idea of quantifying uncertainty and describe some methods for doing so.
- Explain interval estimation, particularly confidence intervals, which are the most common type of interval estimate.
- Briefly review the t-distribution and Central Limit Theorem.
- Calculate confidence intervals for a mean or the difference of two means.
- Distinguish between independent and paired samples.
- Distinguish between probability distributions and data sets, including how concepts such as the mean and variance apply to each.

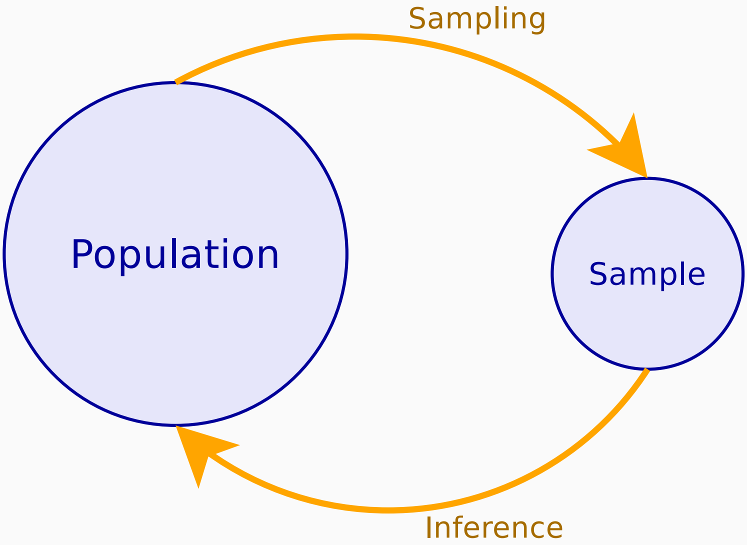
# Starting example: Travel time to Monash

- How long does it take to get to Clayton Campus?
- I asked 10 students and they gave me the following answers (in minutes):

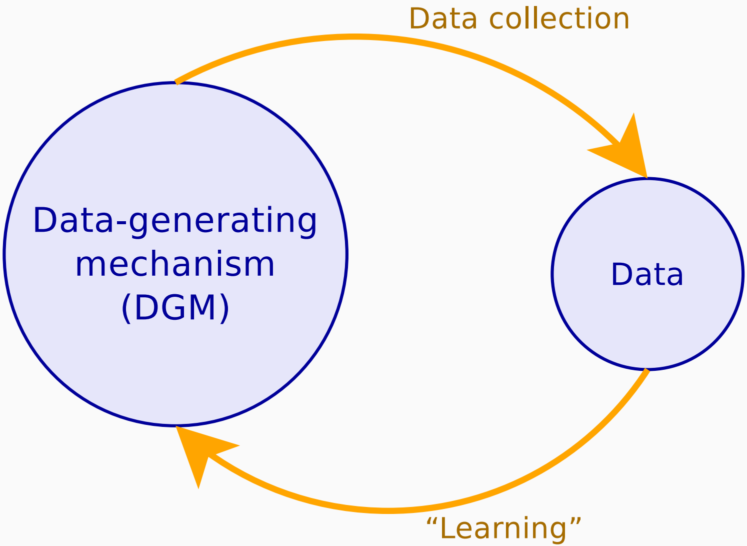
```r
travel <- c(34, 15, 18, 28, 16, 22, 24, 29, 35, 17)
```

- What is the **average** travel time…across all students?

# Statistics: the big picture



# Statistics: the big picture



# Terminology

- Population
- Sample
- Observations
- Data, data set

$$34, 15, \ldots, 17$$

# Notation for data

Actual data:

$$34, 15, \ldots, 17$$

As observed:

$$x_1, x_2, \ldots, x_n$$

As modelled via random variables:

$$X_1, X_2, \ldots, X_n$$

# Outline

# Example: Travel time to Monash (continued)

Just take the average (mean) of the data?

```
mean(travel)
```

```
[1] 23.8
```

Our estimate of the average travel time **across all students** is 23.8 minutes.

# Estimation

- A common type of inference
- Goal: determine (approximately) some numerical property of the population
- Use the data to do this

# More terminology & notation

- **Parameter**: a property of the *population*
- **Statistic**: a quantity calculated from the *sample* (*data*)

## Examples

- Population mean:

$$\mu$$

- Sample mean:

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

- Sample median:

$$m = \mathrm{median}(x_1, x_2, \ldots, x_n)$$

# Example (continued)

```r
mean(travel)
```

```
[1] 23.8
```

```r
median(travel)
```

```
[1] 23
```

Our estimate of $\mu$ (the population mean) could be 23.8 minutes or 23 minutes, depending on which statistic we decide to use.

# More terminology & notation

- **Estimator** (or **point estimator**):
  - A statistic used for estimating a parameter.
  - Refers specifically to the random variable version.
- **Estimate** (or **point estimate**):
  - The observed/realised value of an estimator.

**Examples**

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

$$\bar{x} = \frac{1}{10}(34 + 15 + \cdots + 17) = 23.8$$

# 'Hat' notation

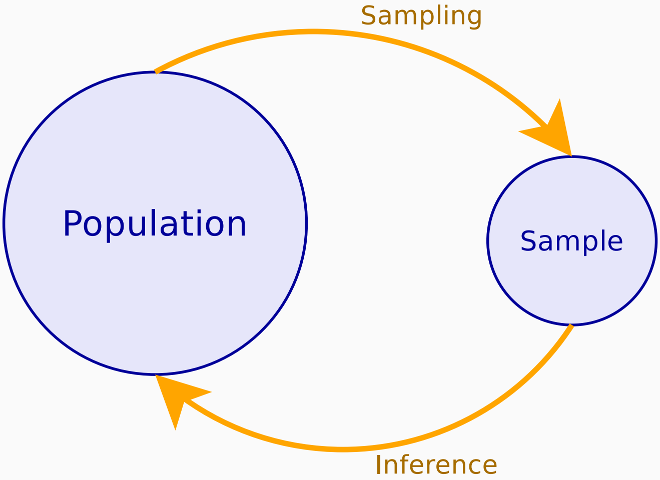
- Let a statistic $T$ be an estimator for a parameter $\theta$
- We often refer to $T$ as $\hat{\theta}$ for convenience, i.e. $\hat{\theta} = T$
- Example: $\hat{\mu} = \bar{X}$

Keep in mind:

- $\theta$ is a parameter
- $\hat{\theta}$ is a statistic
- $\hat{\theta}$ could be an estimator or an estimate (often ambiguous)
- These are common sources of confusion for students!

# Which statistic should we use?

Need to understand the sampling part in order to ‘reverse’ it



# Variation (in data)

- Repeat the data collection $\Rightarrow$ different observations
- This is **sampling variation**
- Can be described using probability distributions
- It will depend on the population and the method of data collection
- Typically not known exactly, but we can make some assumptions
- Such assumptions are known as **modelling assumptions**

# Sampling distribution

- Each observation and statistic will have a **sampling distribution**
- Use random variables for describing these
- Example:

$$X_1, X_2, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$$

- If the observations are independent, then we also know:

$$\bar{X} \sim \mathrm{N}(\mu, \sigma^2/n)$$

# Random sample

- Suppose the observations are *independent* and *identically distributed* (iid)
- This is known as a **random sample**
- A widely used model for data

## Examples

- Selecting people from a population, where each person has the same chance of being selected.
- Running replicate experiments, under the same experimental conditions.

# Properties of estimators

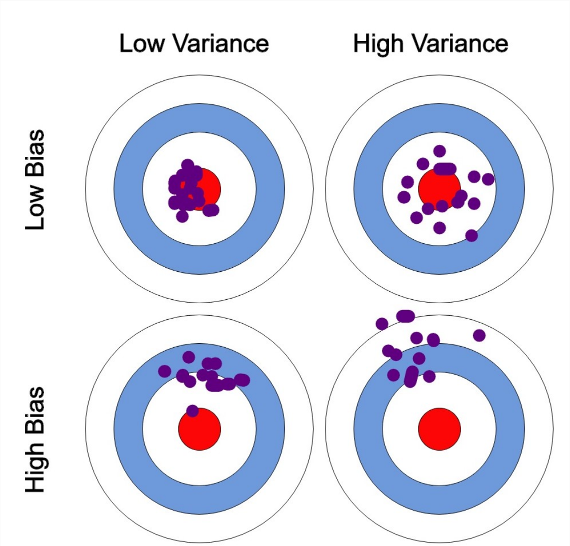
- What makes a "good" estimator?
- Want it to be *likely* to be *close* to the value of the parameter.
- Some concepts:
  - **Bias**: on average how much does the estimator differ to the truth?

$$\mathbb{E}(\hat{\theta}) - \theta$$

  - **Variance**: on average, how much does the estimator vary (i.e. how precise is it)?

$$\mathrm{var}(\hat{\theta})$$

# Properties of estimators



# Outline

# Example: Disability in Melbourne

- Surveying residents in Melbourne to find out how many are disabled.
- The results will be used to set a budget for disability support.

Estimate from survey: 5% of residents are disabled

*What can we conclude?*

Estimate from a second survey: 2% of residents are disabled

*What can we now conclude?*

# Going beyond point estimates

- Point estimates are usually only a starting point
- Insufficient to conclusively answer real questions of interest
- Perpetual lurking questions:
  - How confident are you in the estimate?
  - How accurate is it?
- We need ways to **quantify** and **communicate** the **uncertainty** in our estimates.

# Simple scenario: sampling from a normal distribution

Assume a random sample from a normal distribution:

$$X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$$

Standard estimators:

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

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

In R you calculate these using `mean()` and `var()`.

Also, `sd()` gives $S$.

# Estimating $\mu$

Sampling distribution of $\hat{\mu}$:

$$\hat{\mu} \sim \mathrm{N}(\mu, \sigma^2/n)$$

Properties:

$$\mathbb{E}(\hat{\mu}) = \mu$$

$$\mathrm{sd}(\hat{\mu}) = \sigma/\sqrt{n}$$

How close is our estimate?

# Standard error

- We can *estimate* the precision our estimator.
- The **standard error** is an estimate of the standard deviation of the estimator.

# Simple scenario (continued)

We already know the following:

$$\mathrm{sd}(\hat{\mu}) = \frac{\sigma}{\sqrt{n}}$$

$$\hat{\sigma}^2 = S^2$$

Put them together:

$$\mathrm{se}(\hat{\mu}) = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{S}{\sqrt{n}}$$

This gives a quantitative measure of uncertainty.

(Also, it is an example of “plug-in” estimator.)

# Reporting the standard error

There are many approaches.

Suppose that $\hat{\mu}$ = 17 and $\mathrm{se}(\hat{\mu})$ = 1.3.

Some examples:

- 17 (1.3)
- 17 $\pm$ 1.3
- 17 $\pm$ 2.6    [= 2 $\times$ $\mathrm{se}(\hat{\mu})$]

This now gives us some useful information about the (estimated) accuracy of our estimate.

# Example: Disability in Melbourne (continued)

More info:

- First survey: $5\% \pm 4\%$
- Second survey: $2\% \pm 0.1\%$

*What would we now conclude?*

*What information should we use for setting the disability support budget?*

# Outline

# Motivation

- Let's go one step further...
- The form $\text{estimate} \pm \text{error}$ can be expressed as an interval: $(\text{estimate} - \text{error},\ \text{estimate} + \text{error})$
- This is an example of an **interval estimate**
- More general and more useful than just reporting a standard error
- How can we calculate interval estimates?

# Confidence intervals

- We now need **two** statistics, to define an interval: $(L, U)$
- Want the interval to have a high chance of capturing the value of the parameter. For example:

$$\Pr(L < \mu < U) = 0.95$$

- Such an interval is called a **confidence interval** (CI).
- It has a **confidence level** of 95%.
- We say that it is a "95% confidence interval for $\mu$".

# Example: sampling from a normal distribution

Like before, assume a random sample from a normal distribution:

$$X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$$

Let $c = t_{n-1,0.975}$ (details on a later slide) and define:

$$L = \bar{X} - c \times \frac{S}{\sqrt{n}}$$

$$U = \bar{X} + c \times \frac{S}{\sqrt{n}}$$

$(L, U)$ is a 95% confidence interval for $\mu$.

It is of the form $\hat{\mu} \pm c\,\mathrm{se}(\hat{\mu})$.

# Example: Travel time to Monash (continued)

```r
mean(travel)
```

```
[1] 23.8
```

```r
sd(travel)
```

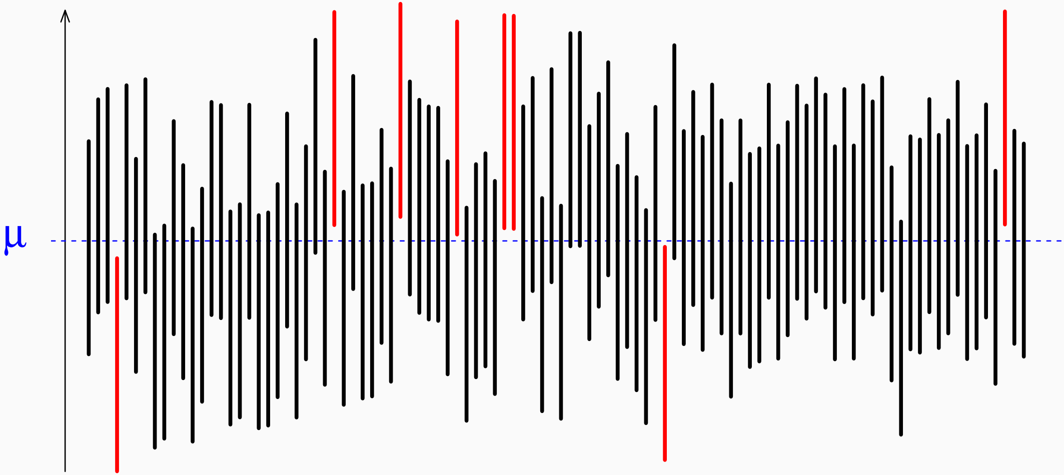
```
[1] 7.42069
```

```r
mean(travel) + c(-1, 1) * 2.26 * sd(travel)/sqrt(10)
```

```
[1] 18.4966 29.1034
```

# Sampling distribution of an interval estimator?

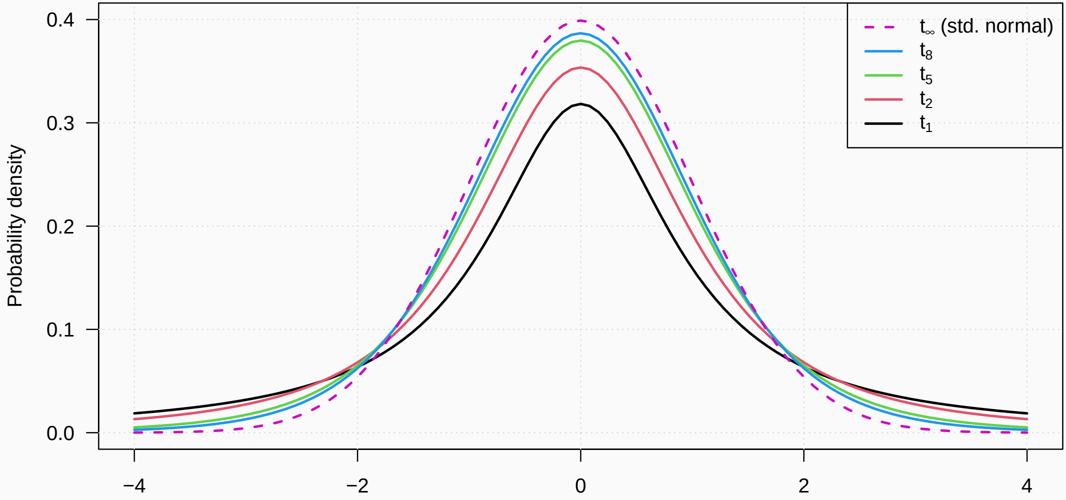
Let's visualise (hypothetical) realisations of the interval:



# The t quantile

Where does the $c = t_{n-1,0.975}$ come from?

# t-distribution



# t-distribution

- Also known as **Student's** *t*-distribution
- Similar to a normal distribution ("bell shaped") but with wider tails.
- Single parameter: $k$, the **degrees of freedom**
- Notation: $T \sim \mathrm{t}_k$ or $T \sim \mathrm{t}(k)$
- As $k \to \infty$, then $\mathrm{t}_k \to \mathrm{N}(0, 1)$
- Describes sampling distributions of statistics from a normal distribution, in particular:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim \mathrm{t}_{n-1}$$

# Derivation

- Let $c = t_{n-1,0.975}$ be the 0.975 quantile from a $\mathrm{t}_{n-1}$ distribution:

$$\Pr(T < c) = 0.975$$

- Due to symmetry:

$$\Pr(-c < T < c) = 0.95$$

- Substitute $T$:

$$\Pr\left(-c < \frac{\bar{X} - \mu}{S/\sqrt{n}} < c\right) = 0.95$$

- Rearrange the double inequality:

$$\Pr\left(-c < \frac{\bar{X}-\mu}{S/\sqrt{n}} < c\right) = 0.95$$

$$\Pr\left(\bar{X} - c \times \frac{S}{\sqrt{n}} < \mu < \bar{X} + c \times \frac{S}{\sqrt{n}}\right) = 0.95$$

- 95% confidence interval is:

$$\bar{X} \pm c \times \frac{S}{\sqrt{n}}$$

# Outline

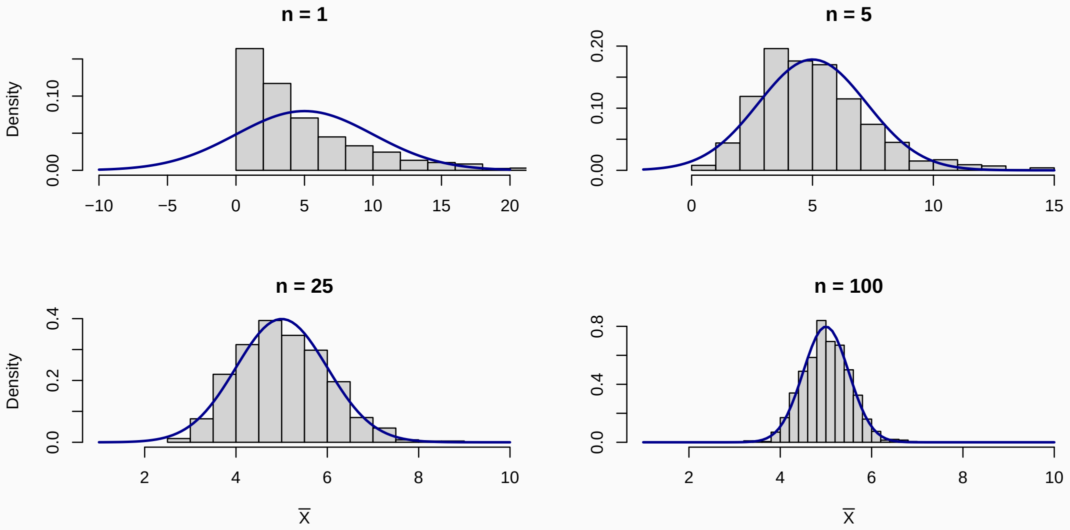
# Central Limit Theorem

- Describes the sampling distribution of $\bar{X}$, as the sample size **increases**
- Let $X_1, X_2, \ldots, X_n$ be a random sample (iid), with
  - $\mathbb{E}(X_i) = \mu$
  - $\mathrm{var}(X_i) = \sigma^2 < \infty$ (finite variance)
- Then
$$\sqrt{n}\left(\bar{X} - \mu\right) \xrightarrow{\text{dist}} \mathrm{N}\left(0, \sigma^2\right), \quad \text{as } n \to \infty.$$
- The sample mean becomes closer to normally distributed... even if the data are **not** normally distributed!
- Can use this approximate the sampling distribution:
$$\bar{X} \overset{\text{approx}}{\sim} \mathrm{N}\left(\mu, \sigma^2/n\right)$$

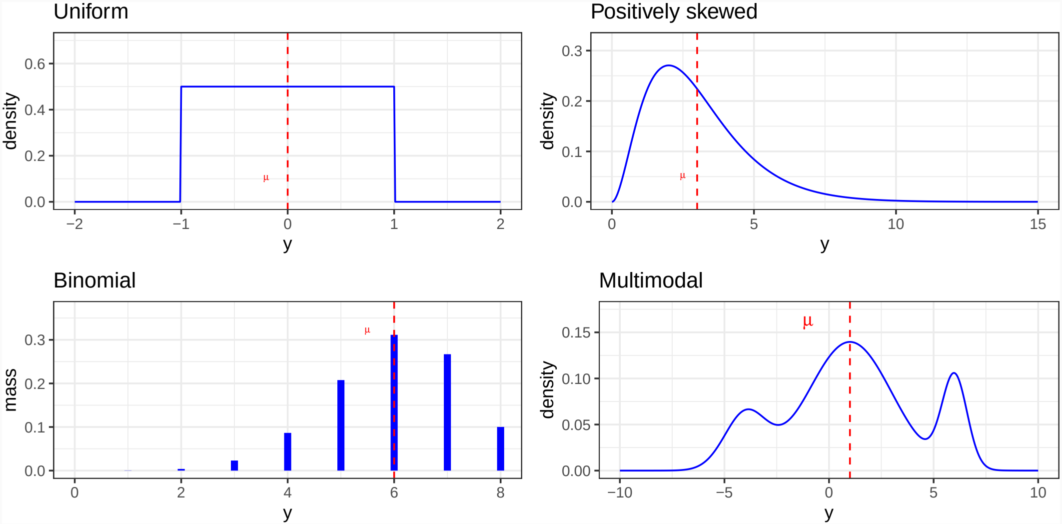
# The CLT is fundamental

- This is an extremely important theorem.
- It provides the ‘magic’ that powers many statistical analyses.
- It ‘kicks in’ even for relatively small samples sizes.

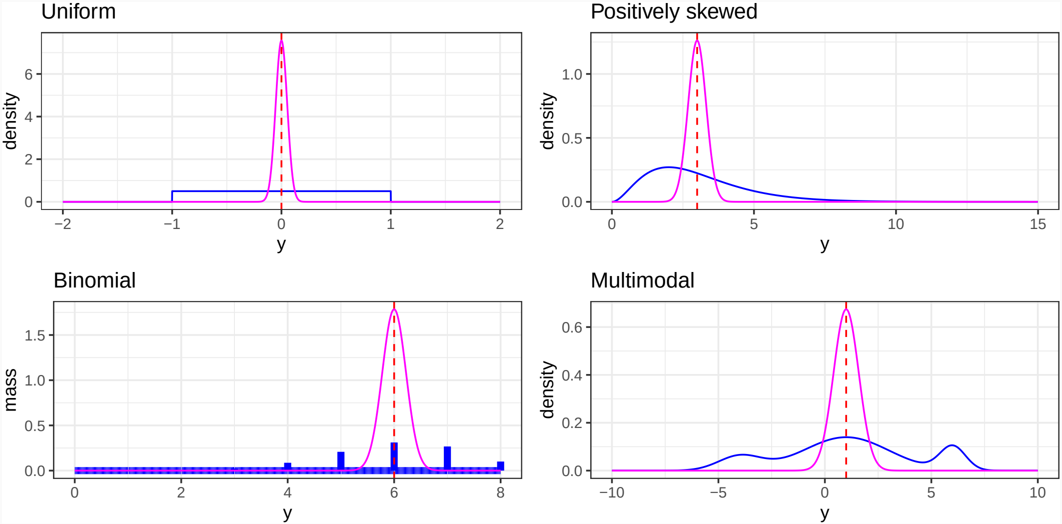
# CLT demonstration



# Some non-normal populations



# CLT approximations (of $\bar{X}$) with $n = 30$



# Outline

# Estimating the mean, in general

- The CLT approximation:

$$\hat{\mu} = \bar{X} \overset{\text{approx}}{\sim} \mathrm{N}\left(\mu, \sigma^2/n\right)$$

- Then proceed the same as before…
- An approximate 95% confidence interval is

$$\bar{X} \pm c \times \frac{S}{\sqrt{n}}$$

# Calculating quantiles

- Let $T \sim \mathrm{t}_k$ and $Z \sim \mathrm{N}(0, 1)$
- Let their $p$ quantiles be $t_{k,p}$ and $z_p$ respectively

$$\Pr\left(T < t_{k,p}\right) = p$$

$$\Pr\left(Z < z_p\right) = p$$

- Calculate these in R as follows:

```
qt(p, k)
qnorm(p)
```

- What happens when `k` is large?

# A ‘short cut’ in R

```r
t.test(travel)$conf.int
```

```
[1] 18.4916 29.1084
attr(,"conf.level")
[1] 0.95
```

# Interpreting CIs

- Our interval estimator is

$$\bar{X} \pm c \times \frac{S}{\sqrt{n}}$$

- Substituting the observed values:

$$\bar{x} \pm c \times \frac{s}{\sqrt{n}}$$

- For example, if $\bar{x} = 10$, $s = 1.2$, $n = 20$, we have $c = t_{19,0.975} = 2.09$, which gives the 95% confidence interval:

$$\bar{x} \pm c \times \frac{s}{\sqrt{n}} = [7.5, 12.5]$$

- What does the "95%" confidence level refer to?

# Interpreting CIs

- **Before** any observations
  - $\bar{X}$ is random
  - 95% probability the interval contains $\mu$
- **After** we observe the data
  - $\bar{x}$ is fixed (not random)
  - The realised interval either contains $\mu$ or it doesn't
  - There is no longer any probability!
- Note: $\mu$ is always fixed

# Interpreting CIs

- The confidence level relates to the **sampling procedure**.
- It refers to **hypothetical repeated samples**.
- It does not describe any specific realised interval.
- Once a specific sample is observed and a CI is calculated, the confidence level **cannot** be interpreted probabilistically with regard to the specific data at hand.

# Explaining CIs

**Don't say**:

- This CI has a 95% chance of including the true value
- We can be "95% confident" that this CI includes the true value

Can say:

- If we were to repeat the data collection, then 95% of the time the CI we calculate will cover the true value.

(This is a bit of a mouthful…)

# Explaining CIs in practice

- If you are reporting results to people who know what they are, you can just state that the "95% confidence interval is…"
- If people want to know what this means, use an intuitive notion like "it is the set of plausible values of the parameter that are consistent with the data".
- If you need to actually explain what a CI is precisely, you need to explain it in terms of repeated sampling. (No shortcuts!)

# Tips for communicating results

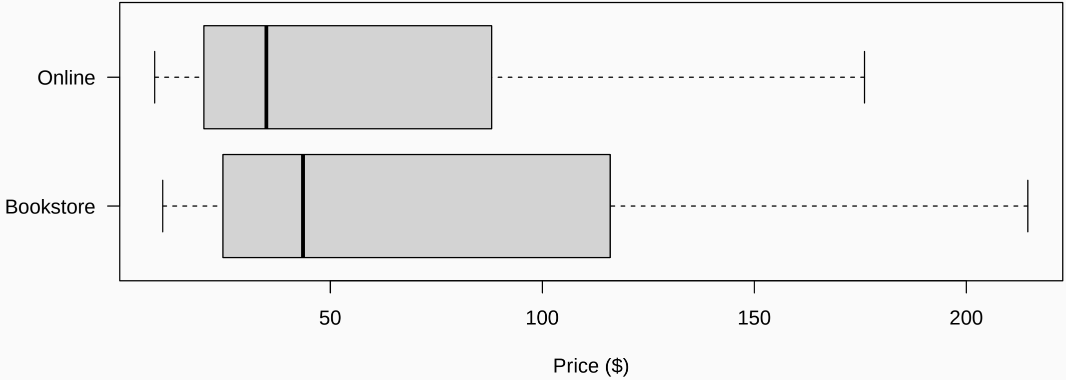
- Describe the extent of your uncertainty
- Emphasise a **range** a plausible values, not just a point estimate

# Outline

# Example: Textbook prices

Are textbooks cheaper online?

Prices collected for 73 randomly sampled textbooks at a university bookstore and from Amazon's website.



# Comparing two independent groups

- Collect separate samples from **two** populations
- Parameter of interest: $\mu_1 - \mu_2$
- Point estimate: $\bar{x}_1 - \bar{x}_2$
- A few ways to calculate CIs
- Simple approximate way:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{k,0.975} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

  with $k = \min(n_1 - 1, n_2 - 1)$.
- R will use a more accurate formula for $k$, called Welch's approximation.

# Example: Textbook prices (comparing independent groups)

```r
mean(textbooks$ucla_new)  # Bookshop
```

```
[1] 72.2219
```

```r
mean(textbooks$amaz_new)  # Online
```

```
[1] 59.4603
```

```r
t.test(textbooks$ucla_new, textbooks$amaz_new)$conf.int
```

```
[1] -5.10348 30.62677
attr(,"conf.level")
[1] 0.95
```

# Paired samples

- Sometimes we observe measurements in **pairs**

$$(X_{11}, X_{12}), (X_{21}, X_{22}), \dots, (X_{n1}, X_{n2})$$

- Each pair is independent of other pairs
- But measurements *within* the pair could be related
- We can exploit this extra information to improve the estimate
- Let $D_i = X_{i1} - X_{i2}$ be the difference of each pair
- Work just with the differences
- Our data are now simpler: $D_1, D_2, \dots, D_n$
- Use same approach as estimating a mean for a single population

# Example: Textbook prices (as paired samples)

```r
t.test(textbooks$ucla_new - textbooks$amaz_new)$conf.int
```

```
[1]  9.43564 16.08765
attr(,"conf.level")
[1] 0.95
```

```r
t.test(textbooks$ucla_new, textbooks$amaz_new, paired = TRUE)$conf.int
```

```
[1]  9.43564 16.08765
attr(,"conf.level")
[1] 0.95
```

# Independent vs paired

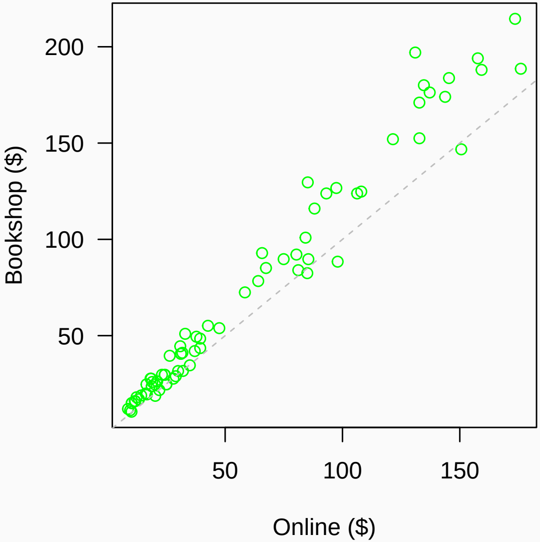
```r
t.test(textbooks$ucla_new, textbooks$amaz_new)$conf.int
```

```
[1] -5.10348 30.62677
attr(,"conf.level")
[1] 0.95
```

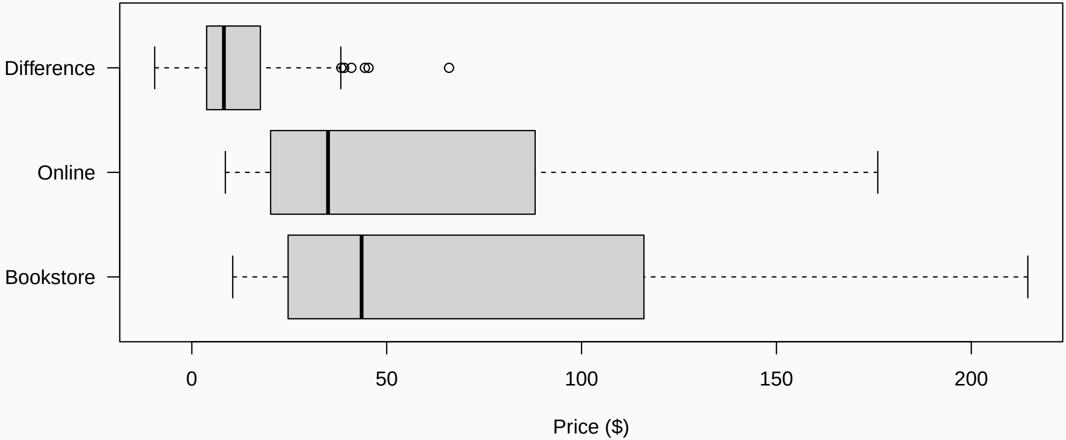
```r
t.test(textbooks$ucla_new, textbooks$amaz_new, paired = TRUE)$conf.int
```

```
[1]  9.43564 16.08765
attr(,"conf.level")
[1] 0.95
```

# Independent vs paired

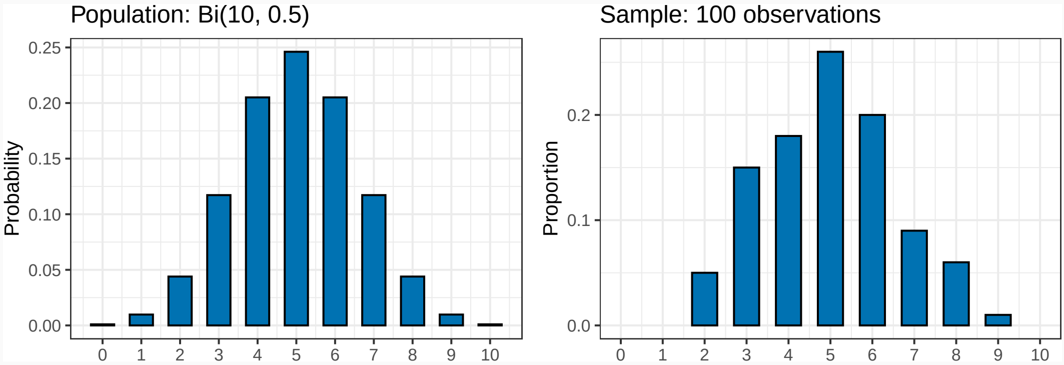


# Independent vs paired



# Outline

# Bar plots / Histograms



# Mean

Population:

$$\mu = \mathbb{E}(X) = \sum_{j=1}^{k} x_j \Pr(X = x_j)$$

Sample:

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

# Variance

Population:

$$\sigma^2 = \mathrm{var}(X) = \sum_{j=1}^{k} (x_j - \mu)^2 \Pr(X = x_j)$$

Sample:

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

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

# Outline

# Learning to ‘think statistically’

- Population vs sample; parameters vs statistics
- Variability (data, statistics; sampling distributions)
- Quantifying uncertainty
- Understanding how the data were generated
- Relevance of the data to the study question
- Exploiting structure in the data, e.g. paired samples