ETC5242 Week 2

授课老师：Joe

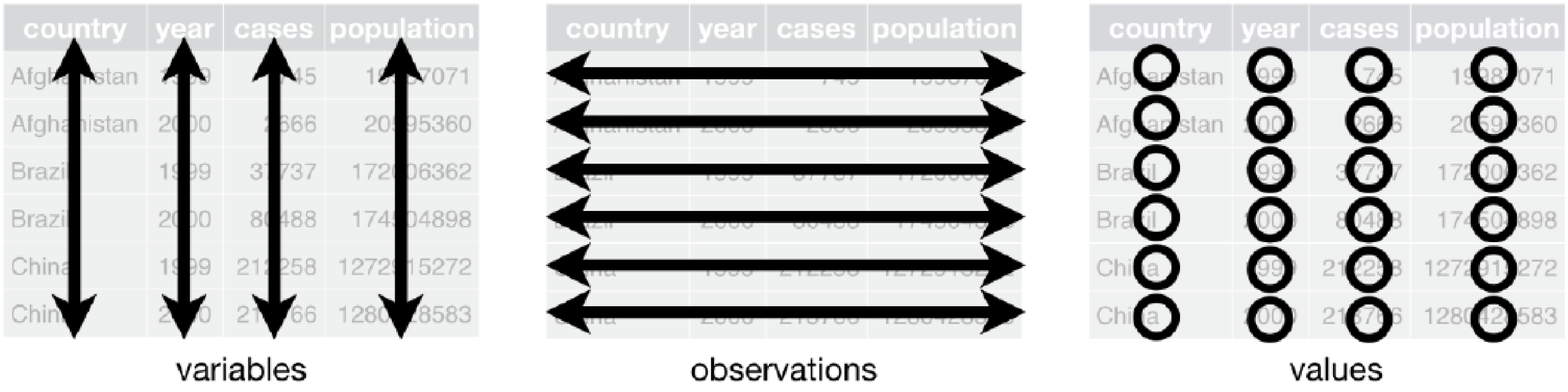
ETC5242平时班 – Week 2

Recall the **tidyverse**: An **R** package, which is itself comprised of many other **R** packages

- ggplot2: data visualisation
- dplyr: data manipulation
- tidyr: data organisation
- readr: data import
- purrr: function iteration
- tibble: data storage
- stringr: string management
- forcats: categorial data functions

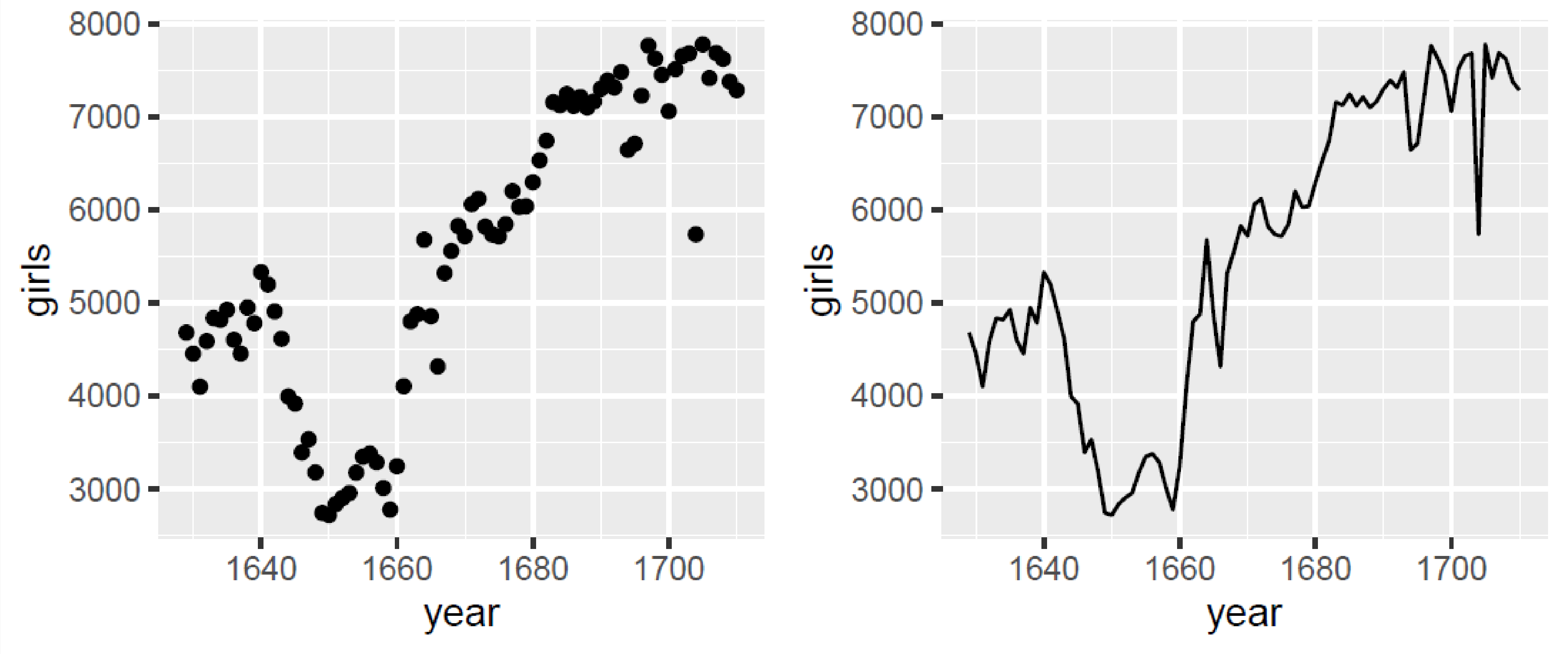
ETC5242平时班 – Week 2

- Observations in rows
- Variables in columns
- Values in cells



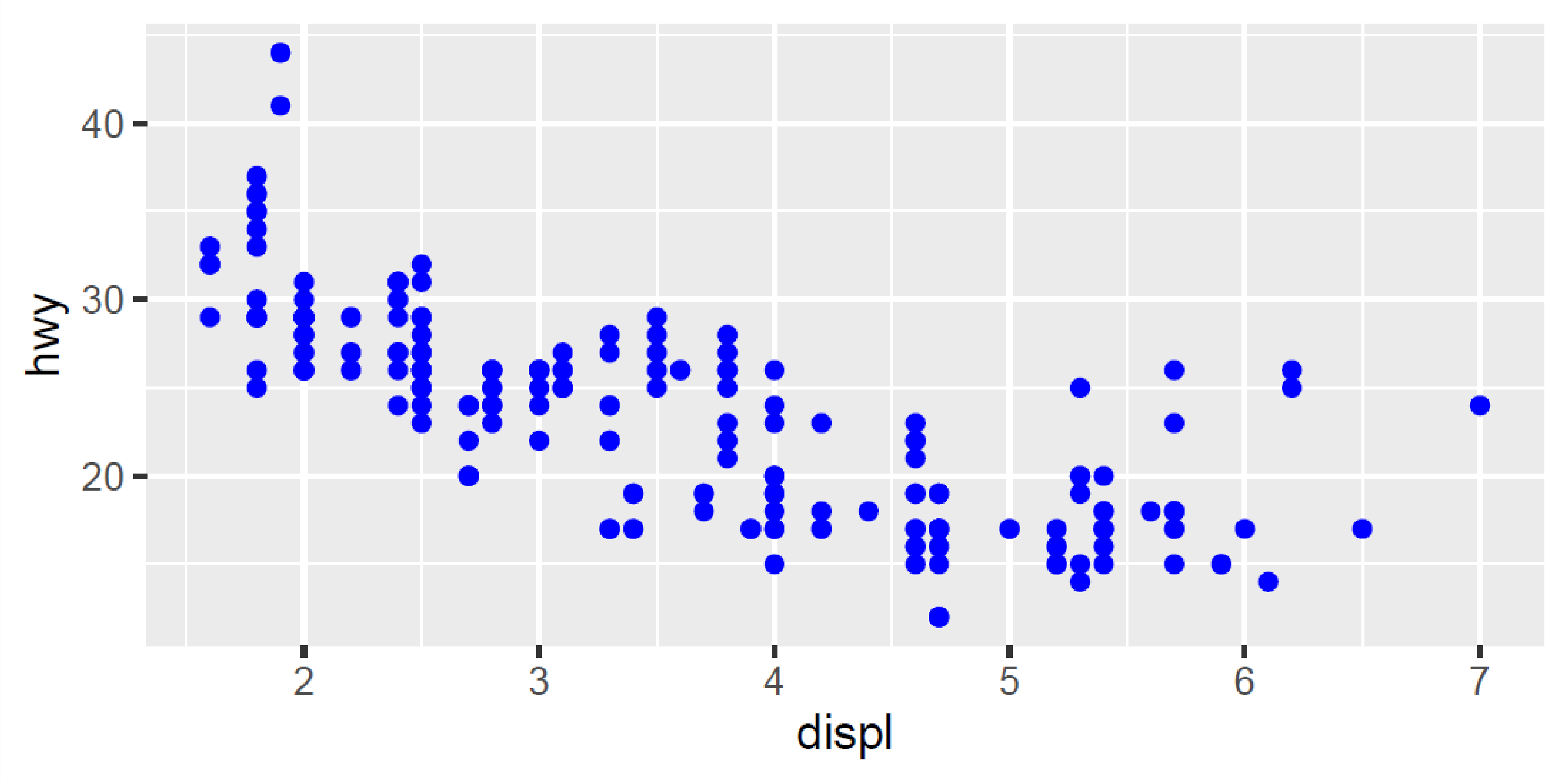
ETC5242平时班 – Week 2

```r
library(openintro)
library(gridExtra)
data(arbuthnot)
p1 <- ggplot(data = arbuthnot, aes(x = year, y = girls)) + geom_point()
p2 <- ggplot(data = arbuthnot, aes(x = year, y = girls)) + geom_line()
grid.arrange(p1, p2, ncol = 2)
```



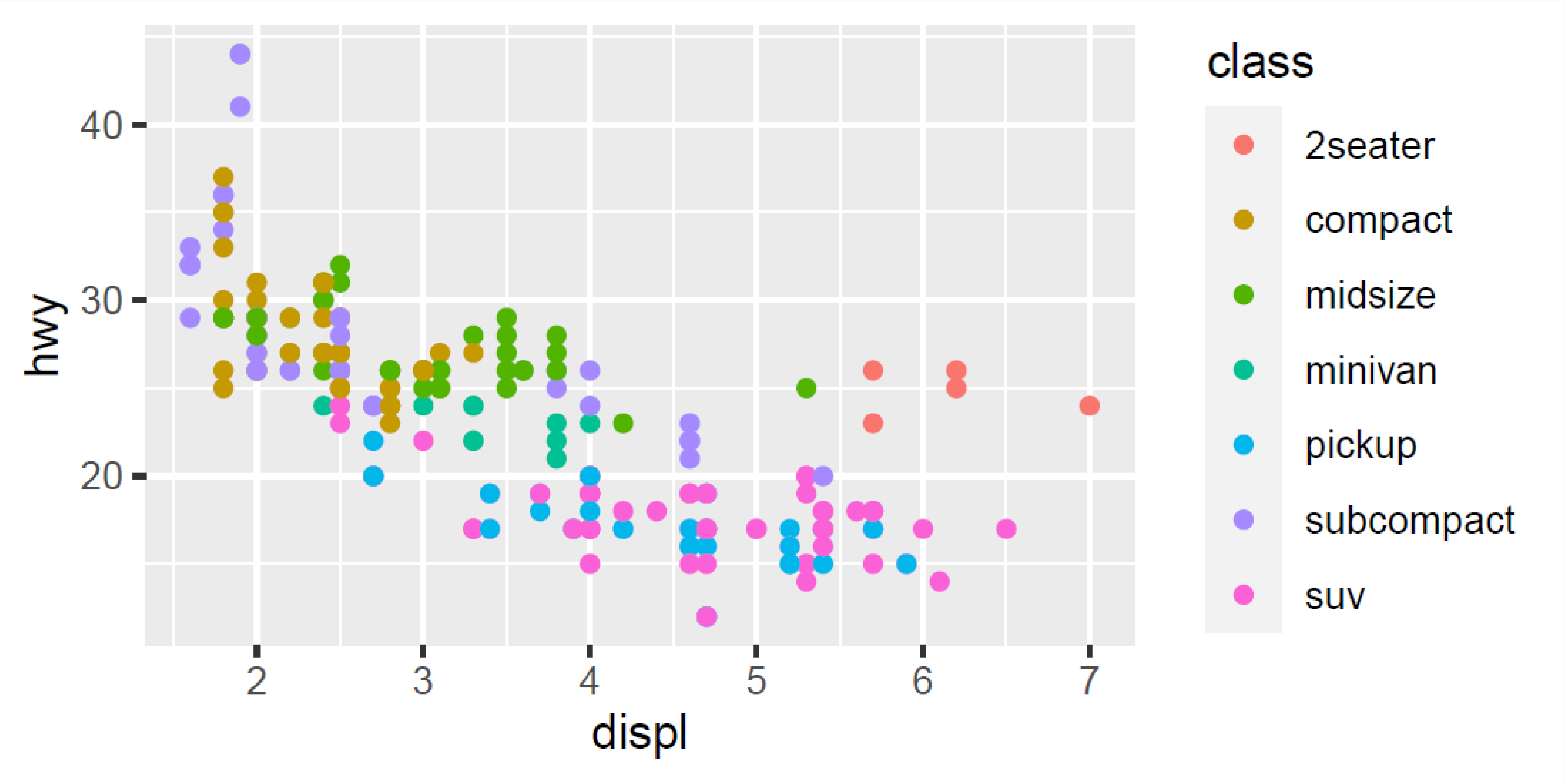
ETC5242平时班 – Week 2

```r
ggplot(mpg, aes(displ, hwy)) + geom_point(colour = "blue")
```



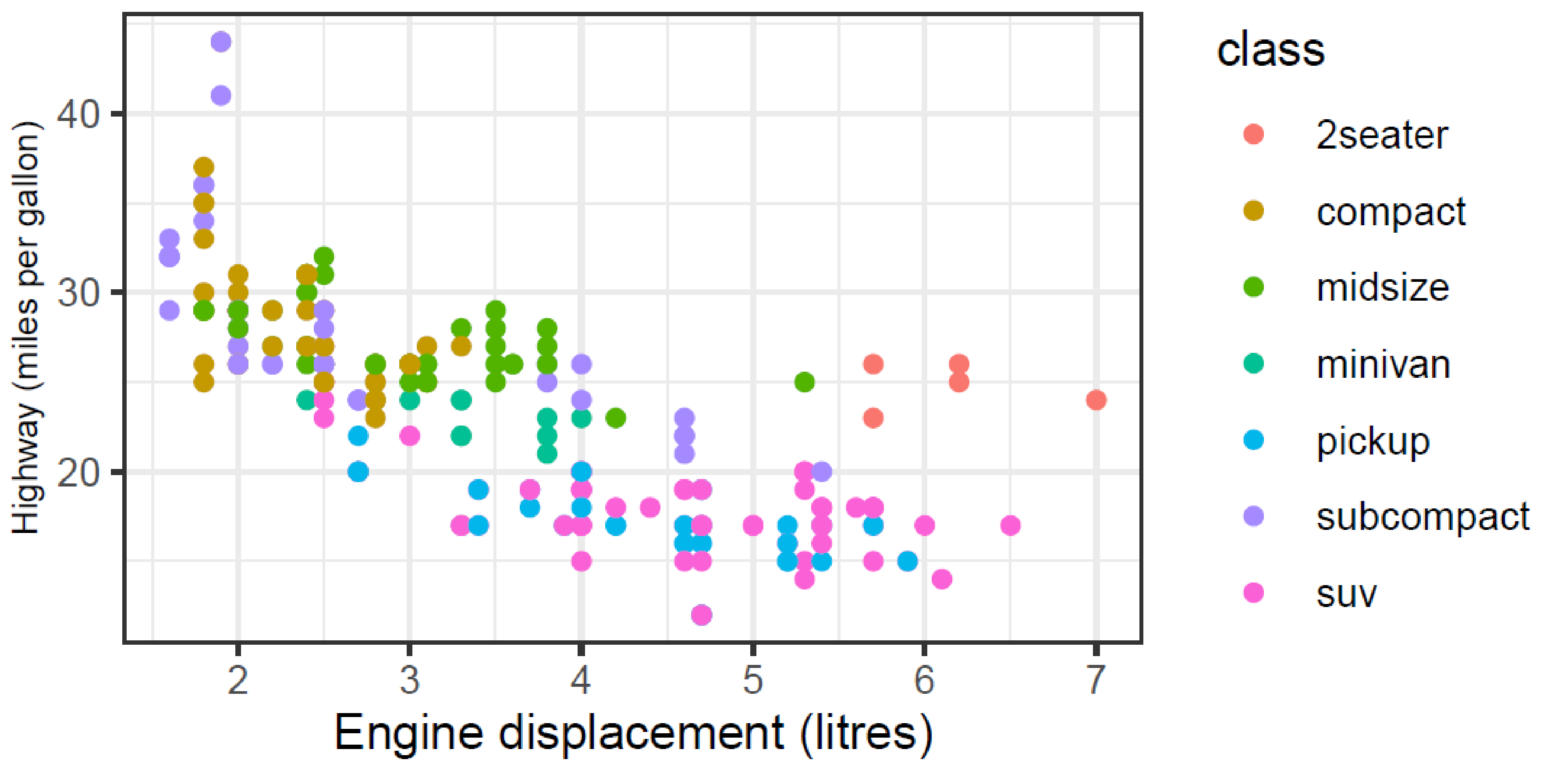
ETC5242平时班 – Week 2

```
mpg %>% ggplot(aes(displ, hwy, colour = class)) + geom_point()
```



ETC5242平时班 – Week 2

```r
ggplot(mpg, aes(displ, hwy, colour = class)) + geom_point() +
    xlab("Engine displacement (litres)") + ylab("Highway (miles per gallon)") +
    theme_bw() + theme(axis.title.y = element_text(size = 8))
```



ETC5242 Week 3

授课老师：Joe

ETC5242平时班 – Week 3

- Hypothesis testing

Do males and females have the same chance of being promoted?

- Population proportion of **males** promoted: p_M
- Population proportion of **females** promoted: p_F

- Test **Null hypothesis** H_0: $p_M = p_F$
  - (Gender has no effect on promotion decision)
- Against **Alternative hypothesis** H_1: $p_M > p_F$
  - Gender has an effect on promotion decision, with a man more likely to be promoted than a woman

		decision		
		promoted	not promoted	Total
gender	male	21	3	24
	female	14	10	24
	Total	35	13	48

ETC5242平时班 – Week 3

- Use the sample proportions as **point estimates** of the “true” p_M and p_F
  - **observed** $\hat{p}_M = \dfrac{\#\text{ males promoted}}{\#\text{ males considered for promotion}}$
  - **observed** $\hat{p}_F = \dfrac{\#\text{ females promoted}}{\#\text{ females considered for promotion}}$
- Do **observed** values satisfy $\hat{p}_M > \hat{p}_F$?
  - Equivalently, is $x_{obs} =$ **observed** $\hat{p}_M - \hat{p}_F > 0$?
- We take x_{obs} is our **point estimate** for $p_M - p_F$
- **Could $x_{obs} > 0$ be due to “chance”?**
- YES. Even when $p_M = p_F$ we can get $x_{obs} > 0$

ETC5242平时班 – Week 3

- Restating the hypotheses of interest:
  - H_0: $p_M - p_F = 0$ (Null hypothesis)
  - H_1: $p_M - p_F > 0$ (Alternative hypothesis)

- Need the **decision rule** to decide whether to reject H_0
- We want to reject H_0 when x_{obs} is far from zero
  - zero is the value of the parameter (here $p_M - p_F$) under H_0

- $\Rightarrow$ Choose the **decision rule**:
  - Reject H_0 when $x_{obs} \geq x^{crit}$
  - Otherwise: Do not Reject H_0

- Here x^{crit} is the **critical value**
  - how is it set?

ETC5242平时班 – Week 3

- Critical value determined by desired to control **Type I error**

		Decision	
		Do not reject H_0	Reject H_0
Truth	H_0 true	no error	**Type I Error**
	H_1 true	**Type II Error**	no error

Table 1: Decision errors from an hypothesis test

- Fix $\Pr(\text{Type I error}) = \alpha$, the **significance level**
  - choose α to be 'small' (e.g. $\alpha = 0.05$)

ETC5242平时班 – Week 3

- *"If H_0 is true and we repeated the experiment, what's the chance we would observed a value of $\hat{p}_M - \hat{p}_F$ that is '**as or more extreme**' than we already have observed with our data?"*

- The "chance" is a probability known as a **p-value**

$$p\text{-value} = \Pr\left(\hat{p}_M - \hat{p}_F \geq x_{obs} \mid H_0 \text{ is true}\right)$$

  - A one-sided test: 'as or more extreme' implies $\geq x_{obs}$ values
  - 'Probability' for a (hypothetical) repeated experiment (under H_0)

- $\Rightarrow$ p-value approach yields same conclusion if use the same α
- $\Rightarrow$ decision rule for p-value approach
  - If p-value $< \alpha$: Reject H_0
  - Otherwise: Do not reject H_0

ETC5242平时班 – Week 3

1 A **randomisation test**: Use variability in observed data

- **NEW**: A **modern computational** approach

2 A test based on the **Central Limit Theorem (CLT)**

Let $X = \hat{p}_M - \hat{p}_F$ denote the unobserved (random variable)

- Either before the data is collected
- Or from a hypothetical repeated experiment

Under the CLT:

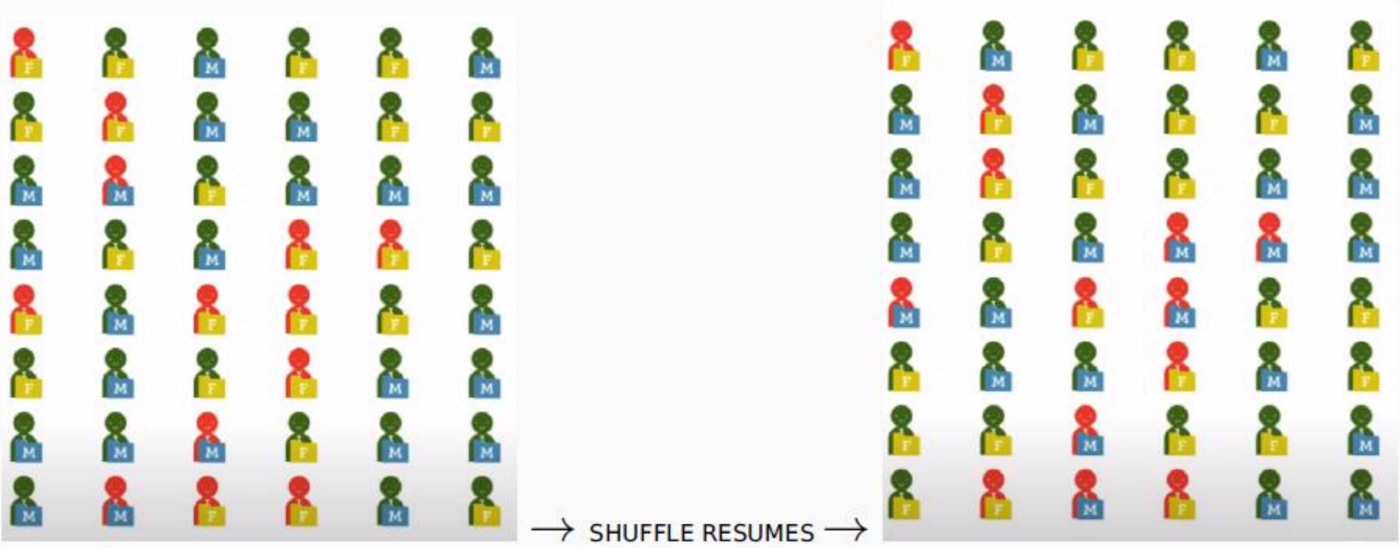
$$X \overset{approx}{\sim} N\left(\mu_X,\ \sigma_X^2\right)$$

- $\mu_X = p_M - p_F$ is the **mean** of X
- σ_X^2 is (an appropriate) **variance** of X
- $\Rightarrow \sigma_X = \sqrt{\sigma_X^2}$ is the **standard error (SE)**
  - but must be estimated since it will depend upon p_M and p_F

ETC5242平时班 – Week 3

Idea: Use the variability in the data to approximate the p-value

- **Shuffle = Randomly permute** resumes (gender) assigned to supervisor (promotion outcome)
  - to simulate X under H_0
  - by breaking association (if present) between gender and promotion outcome



墨学教育
MELBSTUDY

ETC5242 Week 4

授课老师：Joe

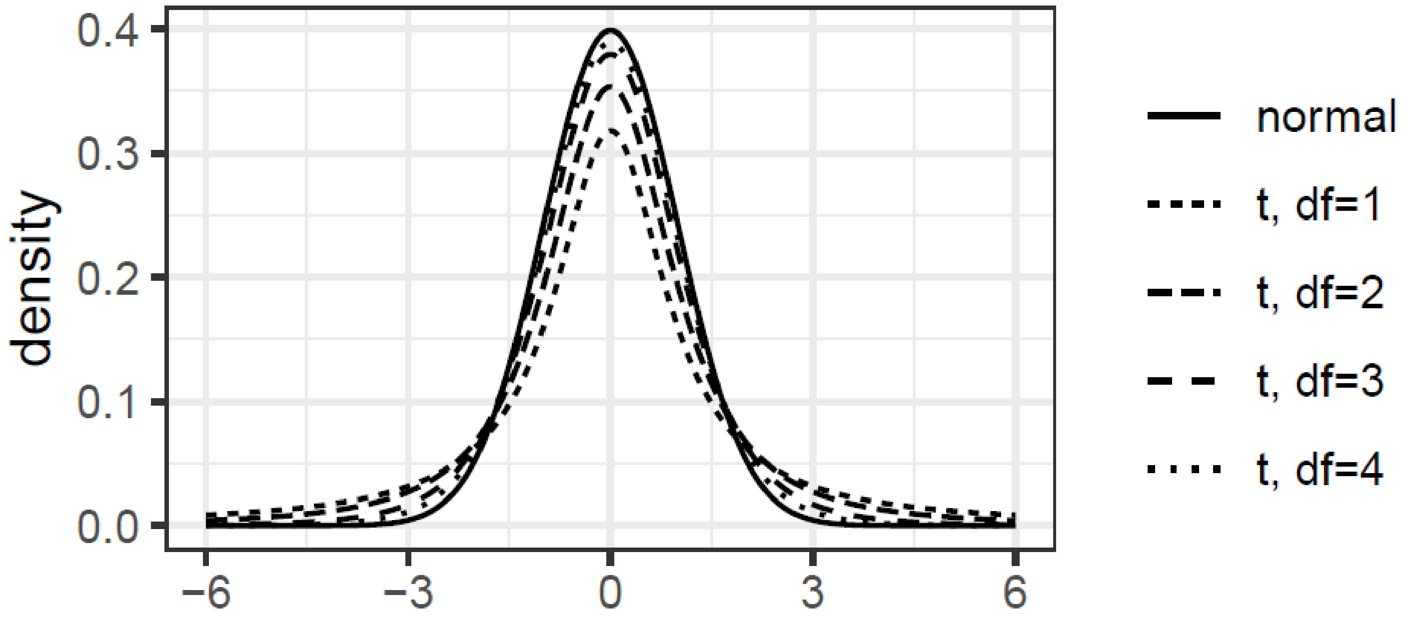
ETC5242平时班 – Week 4

Review: What purpose does a large sample serve?

- As long as observations are independent
- And the population distribution is not "extremely skewed"
- A "large" sample would ensure that...
  - the sampling distribution of the mean is nearly normal
  - the estimate of the standard error ($\frac{s}{\sqrt{n}}$) is reliable
  - (if skewed, even larger sample size is needed)
- $\Rightarrow$ s will be a good estimate of population standard deviation, σ

ETC5242平时班 – Week 4

- Use when σ is unknown (almost always the case) to address the uncertainty of standard error estimate
- Is “bell shaped” but with thicker tails than the normal
  - centered at zero
  - one parameter: degrees of freedom (df) determine thckness of tails
  - compare with $N(\mu, \sigma^2)$, two parameters (mean=μ and SD =σ)



ETC5242平时班 – Week 4

- for inference on a mean where
  - σ unknown, which is almost always
- calculated the same way

$$T = \frac{obs - null}{SE}$$

Here:

- *obs* refers to the value of an observed statistic
  - $\bar{x}$ from one sample, including $\bar{x}_{DIff}$ for paired samples, where $d_i = x_{1i} - x_{2i}$ is the "DIff" for pair i
  - $\bar{x}_1 - \bar{x}_2$ from two independent samples
- *null* refers to the corresponding value of the population quantity under H_0
- *SE* refers to the *Standard Error*, which is the standard deviation of the statistic

- p-value calculated using *R*
  - one or tail areas, based on H_1

ETC5242平时班 – Week 4

- Find the following probabilities:
  a. $\Pr(|Z| > 2) = 0.0455$ ($\rightarrow$ reject)
  b. $\Pr(|t_{df=50}| > 2) = 0.0509$ ($\rightarrow$ fail to reject?)
  c. $\Pr(|t_{df=10}| > 2) = 0.0734$ ($\rightarrow$ fail to reject)
- (And suppose you have a two sided hypothesis test, and your test statistic is 2. Under which scenario would you be able to reject H_0 and the 5% significance level?)
- Generally degrees of freedom (df) is tied to the sample size

ETC5242平时班 – Week 4

- Biscuits after lunch study

biscuit intake	$\bar{x}$	s	n
solitaire	52.1 g	45.1 g	22
no distraction	27.1 g	26.4 g	22

- estimating the mean (single sample): point estimate $\pm$ margin of error

$$\bar{x} \pm t^{\star}_{df} SE_{\bar{x}}$$

$$\bar{x} \pm t^{\star}_{df} \frac{s}{\sqrt{n}}$$

$$\bar{x} \pm t^{\star}_{n-1} \frac{s}{\sqrt{n}}$$

- Degrees of freedom for t statistic for inference on one sample mean: $df = n - 1$
- Find the critical t score using R

ETC5242平时班 – Week 4

- Estimate the average after-lunch snack compumption (in grams) of people who eat lunch **distracted** using a 95% confidence interval

$$
\begin{aligned}
\bar{x} \pm t^{\star} SE &= 52.1 \pm 2.08 \times \frac{45.1}{\sqrt{22}} \\
&= 52.1 \pm 2.08 \times 9.62 \\
&= 52.1 \pm 20 \\
&\Rightarrow (32.1,\ 72.1)
\end{aligned}
$$

Hypothesis test application

- Suppose the suggested service size of these biscuits is 30g. Do these data provide convincing evidence that the amount of snacks consumed by distracted eaters post-lunch is different than the suggested serving size?

$$H_0 : \mu = 30 \quad \text{vs.} \quad H_1 : \mu \neq 30$$

$$T = \frac{52.1 - 30}{9.62} = 2.3$$

- p-value: use 2 $\times$ probability greater than T under t distribution with df=21:

```
[1] 0.0318023
```

- $\Rightarrow$ Reject H_0 at level $\alpha = 0.05$ since p-value = 0.0318 < 0.05

Video 3: Analysing paired data

- When two sets of observations have this special correspondence (not independent), they are said to be **paired**
- To analyse paired data, it is often useful to look at the difference in outcomes of each pair of observations:

$$DIff_i = read_i - write_i, \text{ for each } i = 1, 2, \dots, n = 200$$

- parameter of interest: Average difference between the reading and writing scores of **all** high school students

xbar_DIff	s_DIff	n_DIff	tstat
-0.545	8.88667	200	-0.867

- point estimate? Average difference between the reading and writing scores of **sampled** high school students: $\bar{x}_{DIff}$

ETC5242平时班 – Week 4

- $H_0 : \mu_{Diff} = 0$ vs. $H_1 : \mu_{Diff} \neq 0$
- Same structure as one-sample mean test
- Test statistic:

$$T = \frac{-0.545 - 0}{8.887/\sqrt{200}} = -0.867$$

- degrees of freedom: 200 - 1 = 199
- $\Rightarrow$ p-value: 2*pt(-0.867, df=199) = 0.387
- Interpretation?: *p-value of obtaining a random sample of 200 students where the average difference between the reading and writing scores is at least 0.545 (in either direction), if in fact the true average difference between the scores is zero.*

$$\Pr(\text{observed or more extreme outcome} \mid H_0 \text{ is true})$$

ETC5242平时班 – Week 4

- Refer back to earlier **Biscuits after lunch study**
- point estimate $\pm$ margin of error

$$(\bar{x}_1 - \bar{x}_2) \pm t^{\star} SE_{(\bar{x}_1 - \bar{x}_2)}$$

- Standard error of difference between two independent means: $SE_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
- We add the two variances (inside the square root) even though we are looking for the Standard error of the difference ($SE_{(\bar{x}_1 - \bar{x}_2)}$)
- Video advocates:
  DF for t statistic for inference on difference of two independent means: $df = min(n_1 - 1, n_2 - 1)$

ETC5242平时班 – Week 4

Video 1: Independence:

- Within groups: sampled observations must be independent
  - random sample/assignment
  - if sampling without replacement, $n < 10\%$ of population
- Between groups: the two groups must be independent of each other (non-paired)

2 Sample size/skew: The more skewed the populations, the large the sample size we need from those distributions

ETC5242平时班 – Week 4

- $\Rightarrow$ Confidence interval for $\mu_{wd} - \mu_{wod}$: (1.83g, 48.17g)
  - obtained from:

$$(\bar{x}_{wd} - \bar{x}_{wod}) \pm t^{\star}_{df}\, SE = (52.1 - 27.1) \pm 2.08 \times \sqrt{\frac{45.1^2}{22} + \frac{26.4^2}{22}}$$

$$= 25 \pm 2.08 \times 11.14$$

$$= 25g \pm 23.17g$$

- $\Rightarrow$ Hypothesis test:

$$H_0 : \mu_{wd} - \mu_{wod} = 0 \quad \text{vs.} \quad H_1 : \mu_{wd} - \mu_{wod} \neq 0$$

- $T_{21} = \frac{25-0}{11.14} = 2.24 \Rightarrow$ p-value: 2*pt(2.24, df=21, lower.tail=FALSE) = 0.036
  - Reject $H_0 : \mu_{wd} - \mu_{wod} = 0$ at the 5% level