ETC5521: Exploratory Data Analysis

Working with a single variable, making transformations, detecting outliers, using robust statistics

Lecturer: *Di Cook*

ETC5521.Clayton-x@monash.edu

Week 5 - Session 2

Categorical variables

This lecture is based on Chapter 4 of

Unwin (2015) Graphical Data Analysis with R

There are two types of categorical variables

Nominal where there is no intrinsic ordering to the categories
E.g. blue, grey, black, white.

Ordinal where there is a clear order to the categories.
E.g. Strongly disagree, disagree, neutral, agree, strongly agree.

Categorical variables in R

- In R, categorical variables may be encoded as **factors**.

```
data <- c(2, 2, 1, 1, 3, 3, 3, 1)
factor(data)
```

```
## [1] 2 2 1 1 3 3 3 1
## Levels: 1 2 3
```

- You can easily change the labels of the variables:

```
factor(data, labels = c("I", "II", "III"))
```

```
## [1] II  II  I   I   III III III I
## Levels: I II III
```

- Order of the factors are determined by the input:

```
# numerical input are ordered in increasing order
factor(c(1, 3, 10))
```

```
## [1] 1  3  10
## Levels: 1 3 10
```

```
# character input are ordered alphabetically
factor(c("1", "3", "10"))
```

```
## [1] 1  3  10
## Levels: 1 10 3
```

```
# you can specify order of levels explicitly
factor(c("1", "3", "10"),
  levels = c("1", "3", "10")
)
```

```
## [1] 1  3  10
## Levels: 1 3 10
```

Numerical factors in R

```
x <- factor(c(10, 20, 30, 10, 20))
mean(x)

## Warning in mean.default(x): argument is not numeric or logical: returning NA

## [1] NA
```

⚠ `as.numeric` function returns the internal integer values of the factor

```
mean(as.numeric(x))

## [1] 1.8
```

You probably want to use:

```
mean(as.numeric(levels(x)[x]))

## [1] 18
```

```
mean(as.numeric(as.character(x)))

## [1] 18
```

Numerical summaries: counts, proportions, percentages and odds

```
## # A tibble: 22 × 7
##    country   iso3   year count      p   pct  odds
##    <chr>     <chr> <dbl> <dbl>  <dbl> <dbl> <dbl>
##  1 Australia AUS    2000   982 0.0522  5.22 1    
##  2 Australia AUS    2001   953 0.0507  5.07 0.970
##  3 Australia AUS    2002  1008 0.0536  5.36 1.03 
##  4 Australia AUS    2003   926 0.0493  4.93 0.943
##  5 Australia AUS    2004  1036 0.0551  5.51 1.05 
##  6 Australia AUS    2005  1030 0.0548  5.48 1.05 
##  7 Australia AUS    2006  1127 0.0600  6.00 1.15 
##  8 Australia AUS    2007  1081 0.0575  5.75 1.10 
##  9 Australia AUS    2008  1182 0.0629  6.29 1.20 
## 10 Australia AUS    2009  1176 0.0626  6.26 1.20 
## 11 Australia AUS    2010  1146 0.0610  6.10 1.17 
## 12 Australia AUS    2011  1202 0.0640  6.40 1.22 
## 13 Australia AUS    2012  1259 0.0670  6.70 1.28 
## 14 Australia AUS    2013   512 0.0272  2.72 0.521
## 15 Australia AUS    2014   474 0.0252  2.52 0.483
```

scroll ↓

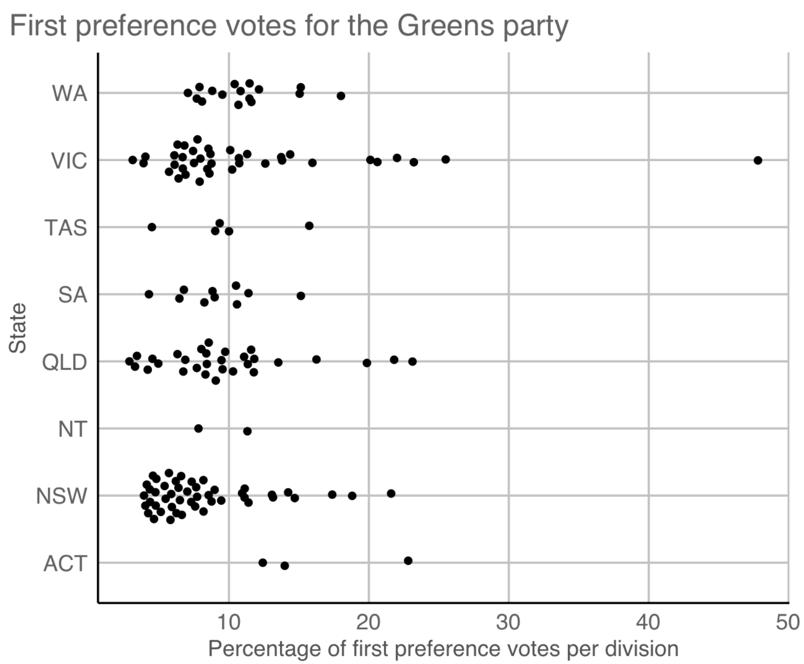
For qualitative data, compute

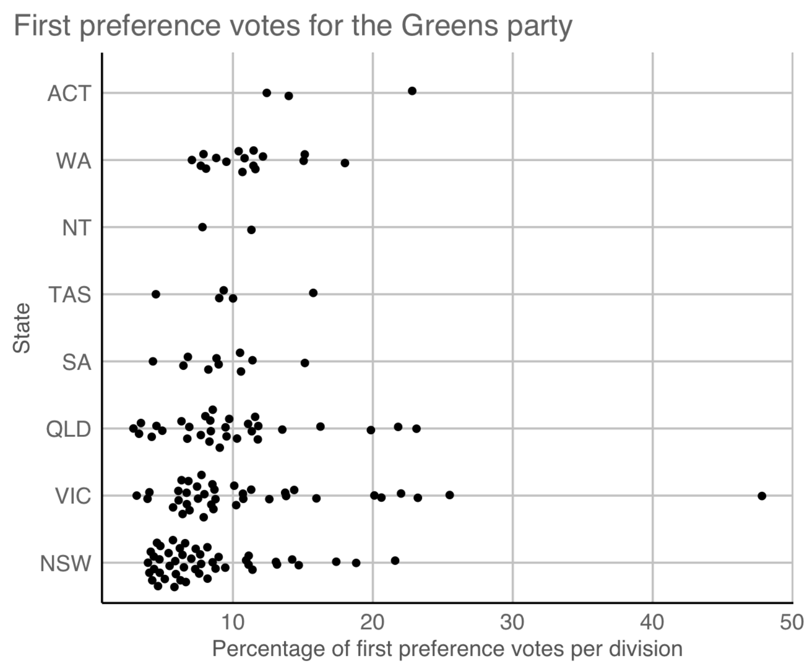
- count/frequency,
- proportion/percentage
- and sometimes, an **odds ratio**. Here we have used ratio relative to the count in year 2000.

Note: For exploration, no rounding of digits was done, but to report you would need to make the numbers pretty.

Revisiting Case study 1 2019 Australian Federal Election

📊 data R





Sorting levels is (almost) always better when plotting

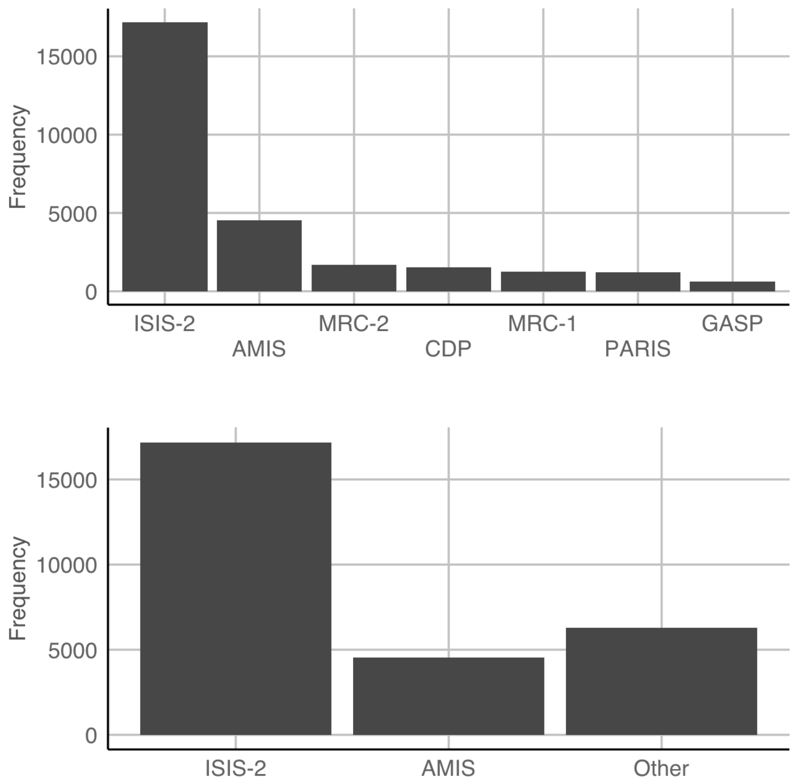
Order nominal variables meaningfully

Coding tip: use below functions to easily change the order of factor levels

```
stats::reorder(factor, value, mean)
forcats::fct_reorder(factor, value, median)
forcats::fct_reorder2(factor, value1, value2, func)
```

Case study 6 Aspirin use after heart attack

data R



- Meta-analysis is a statistical analysis that combines the results of multiple scientific studies.
- This data studies the use of aspirin for death prevention after myocardial infarction, or in plain terms, a heart attack.
- The ISIS-2 study has more patients than all other studies combined.
- You could consider lumping the categories with low frequencies together.

Fleiss JL (1993): The statistical basis of meta-analysis. *Statistical Methods in Medical Research* **2** 121–145

Balduzzi S, Rücker G, Schwarzer G (2019), How to perform a meta-analysis with R: a practical tutorial, Evidence-Based Mental Health.

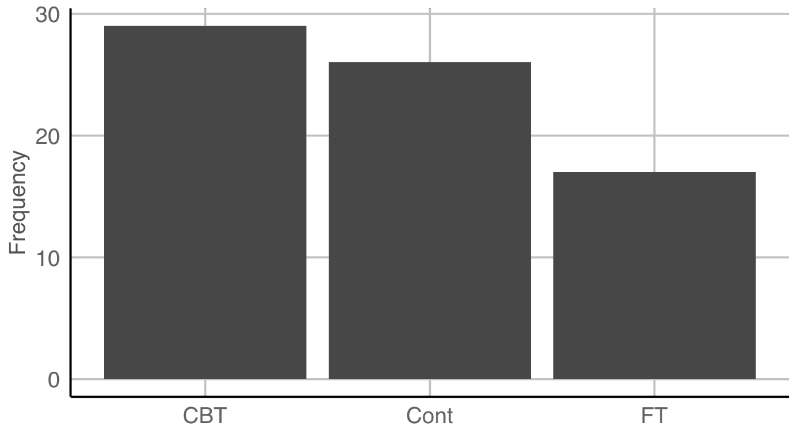
Consider combining factor levels with low frequencies

</> **Coding tip**: the following family of functions help to easily lump factor levels together:

```
forcats::fct_lump()
forcats::fct_lump_lowfreq()
forcats::fct_lump_min()
forcats::fct_lump_n()
forcats::fct_lump_prop()
# if conditioned on another variable
ifelse(cond, "Other", factor)
dplyr::case_when(
  cond1 ~ "level1",
  cond2 ~ "level2",
  TRUE ~ "Other"
)
```

Case study 7 Anorexia

data R

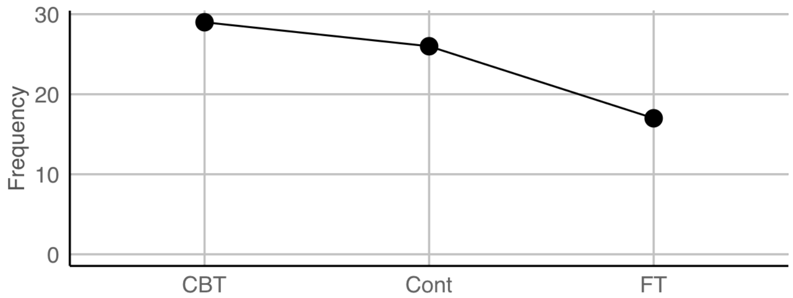


Treatment	Frequency
CBT	29
Cont	26
FT	17

Table or Plot?

- Table for accuracy, plot for visual communication

Why not a point or line?

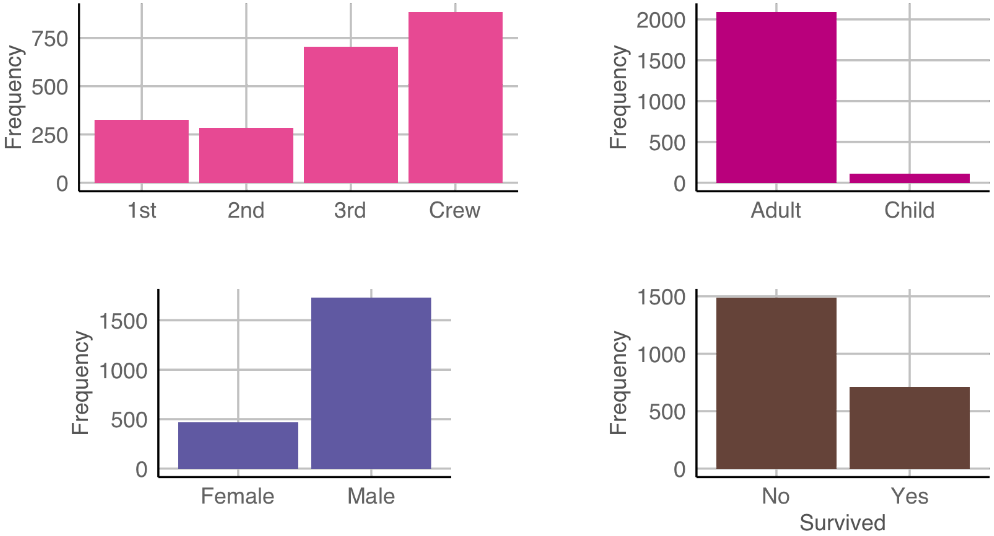


- This can be appropriate depending on what you want to communicate
- A barplot occupies more area compared to a point and the area does a better job of communicating size
- A line is suggestive of a trend

Hand, D. J., Daly, F., McConway, K., Lunn, D. and Ostrowski, E. eds (1993) A Handbook of Small Data Sets. Chapman & Hall, Data set 285 (p. 229)
Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

Case study 8 Titanic

data R

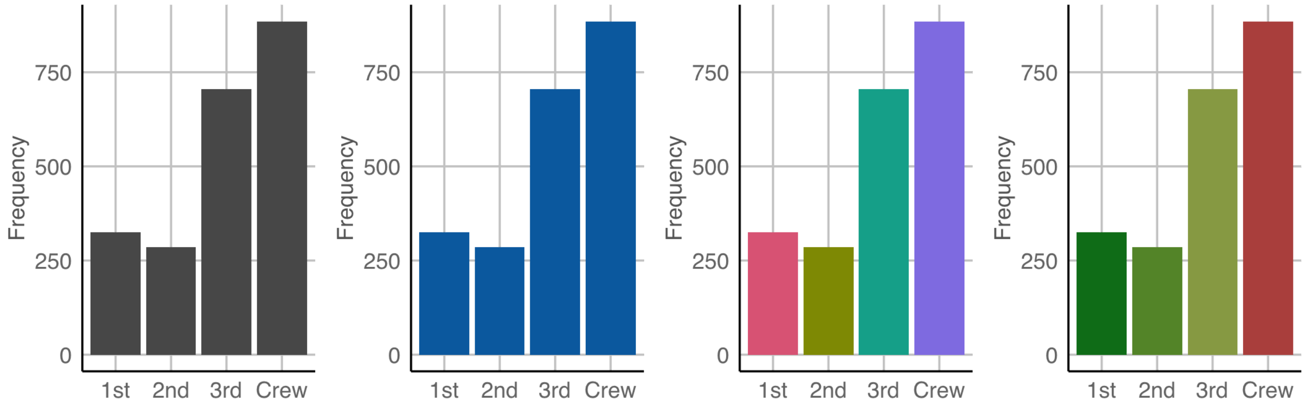


What does the graphs for each categorical variable tell us?

- There were more crews than 1st to 3rd class passengers
- There were far more males on ship; possibly because majority of crew members were male. You can further explore this by constructing two-way tables or graphs that consider both variables.
- Most passengers were adults.
- More than two-thirds of passengers died.

British Board of Trade (1990), Report on the Loss of the 'Titanic' (S.S.). British Board of Trade Inquiry Report (reprint). Gloucester, UK: Allan Sutton Publishing

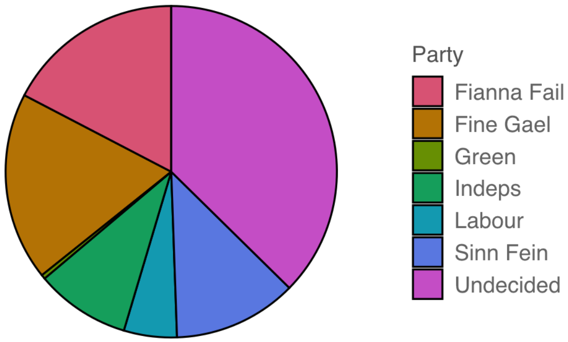
Coloring bars

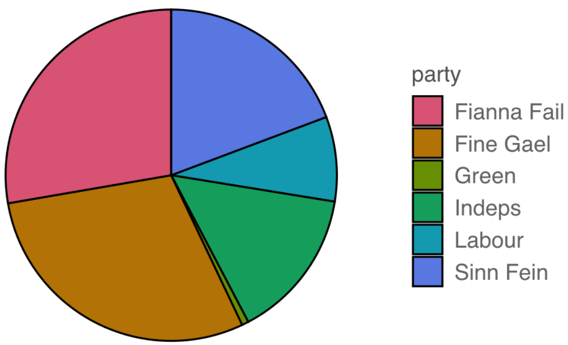


- Colour here doesn't add information as the x-axis already tells us about the categories, but colouring bars can make it more visually appealing, and provide subtle clues.
- If you have too many categories colour won't work well to differentiate the categories.

Case study 9 Opinion poll in Ireland Aug 2013

data R

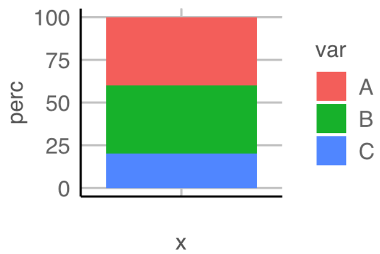




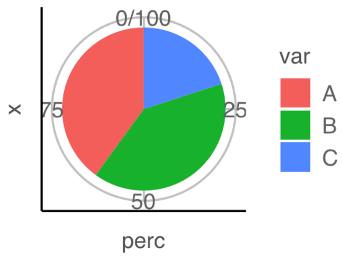
- Pie chart is popular in mainstream media but are not generally recommended as people are generally poor at comparing angles.
- 3D pie charts should definitely be avoided!
- Here you can see that there are many people that are "Undecided" for which political party to support and failing to account for this paints a different picture.

Piechart is a stacked barplot just with a transformed coordinate system

```
df <- data.frame(var = c("A", "B", "C"), perc = c(40, 40, 20))
g <- ggplot(df, aes("", perc, fill = var)) +
  geom_col()
g
```

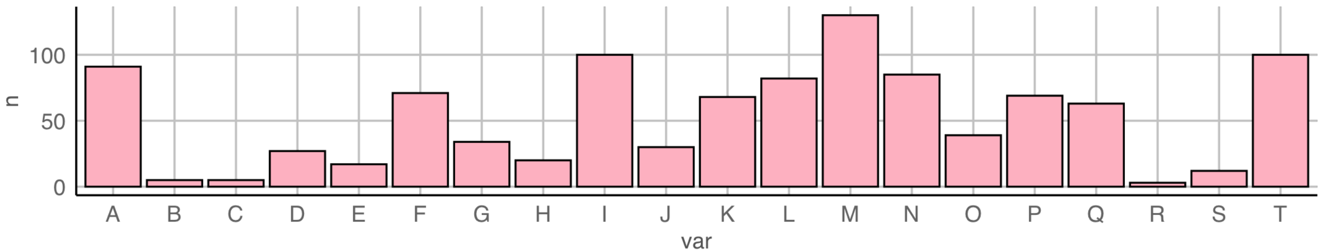



```
g + coord_polar("y")
```

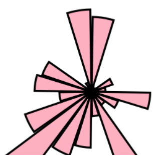


Roseplot is a barplot just with a transformed coordinate system

```
dummy <- data.frame(
  var = LETTERS[1:20],
  n = round(rexp(20, 1 / 100))
)
g <- ggplot(dummy, aes(var, n)) +
  geom_col(fill = "pink", color = "black")
g
```




```
g + coord_polar("x") + theme_void()
```



Visual inference

Typical plot description:

```
ggplot(data, aes(x=var1)) +
  geom_col()

ggplot(data, aes(x=var1)) +
  geom_bar()
```

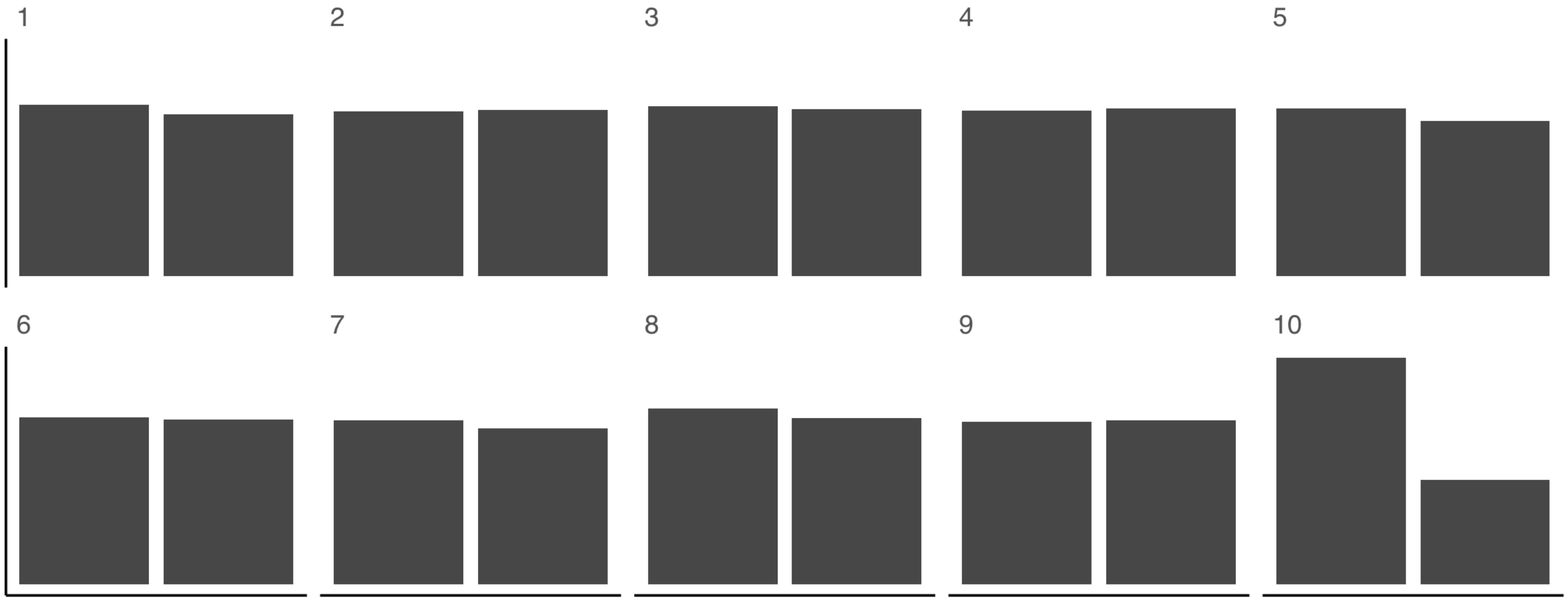
Potential simulation method from binomial

```
# Only one option
null_dist("var1", "binom",
  list(size=n, p=phat))
```

Is the distribution consistent with a sample from a binomial distribution with a given p?

Lineup of tuberculosis count between sexes

R



Note: This is nothing more than you can learn from a conventional hypothesis test of

$$H_0 : p = 0.5$$

. **Stay tuned for more interesting visual inference lineups in coming weeks!**

Take away messages

- Again, be prepared to do multiple plots
- Changing bins or binwidth/bandwidth in histogram, violin or density plots can paint a different picture
- Consider different representations of categorical variables (reordering meaningfully, lumping low frequencies together, plot or table, pie or barplot, missing categories)

Resources and Acknowledgement

- Slides originally written by Emi Tanaka and constructed with `xaringan`, remark.js, `knitr`, and R Markdown.

Lecturer: *Di Cook*

ETC5521.Clayton-x@monash.edu

Week 5 - Session 2