MONASH University

ETC5521: Exploratory Data Analysis

Using computational tools to determine whether what is seen in the data can be assumed to apply more broadly

Lecturer: *Di Cook*

ETC5521.Clayton-x@monash.edu

Week 4 - Session 2

These slides cover

- Why is a data plot a statistic?
- Determining the null hypothesis
- Generating null samples
- Computing the power


Graphical inference

Why is a data plot a statistic?

- The concept of tidy data matches elementary statistics
- Tabular form puts variables in columns and observations in rows
- Not all tabular data is in this form
- This is the point of tidy data

$$X = \begin{bmatrix} X_1 & X_2 & \dots & X_p \end{bmatrix}$$
$$= \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{np} \end{bmatrix}$$

- We might even make assumptions about the distribution of each variable, e.g. $X_1 \sim N(0,1),\ X_2 \sim \text{Exp}(1) \dots$

Why is a data plot a statistic?

- A statistic is a function on the values of items in a sample, e.g. for n iid random variates $\bar{X}_1 = \sum_{i=1}^{n} X_{i1}$, $s_1^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_{i1} - \bar{X}_1)^2$
- We study the behaviour of the statistic over all possible samples of size n.
- The grammar of graphics is the mapping of (random) variables to graphical elements, making plots of data into statistics

```
ggplot(threept_sub,
       aes(x=angle, y=r)) +
  geom_point(alpha=0.3)
```

`angle` is mapped to the x axis

`r` is mapped to the y axis

```
ggplot(penguins,
      aes(x=bill_length_mm,
          y=flipper_length_mm,
          colour=species)) +
  geom_point()
```

`bill_length_mm` is mapped to the x axis

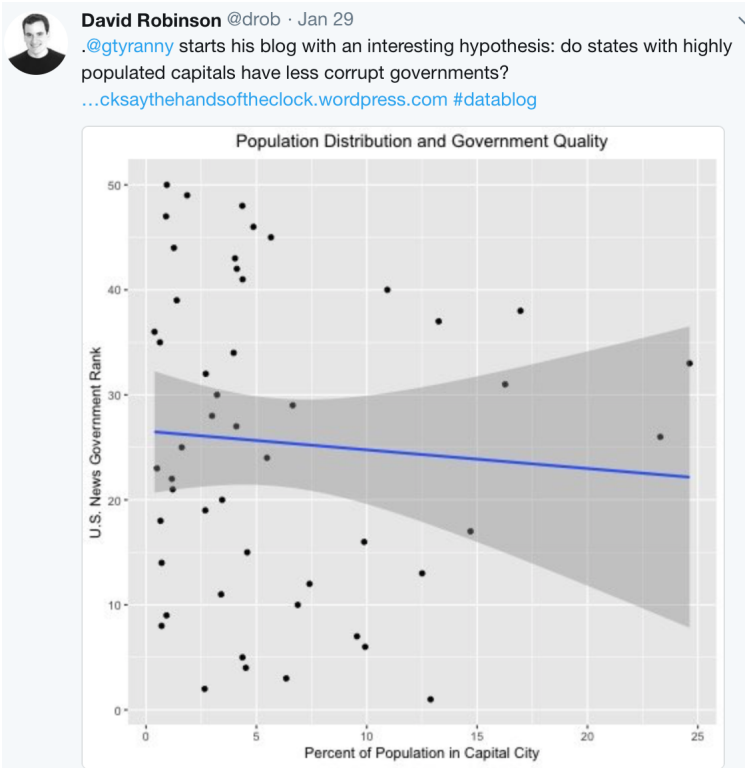
`flipper_length_mm` is mapped to the y axis

`species` is mapped to colour

What is inference?

Inferring that what we see in the data at hand holds more broadly in life, society and the world.

Here's an example tweeted by David Robinson, commenting on an analysis in Tick Tock blog, by Graham Tierney



> *Below is a simple scatterplot of the two variables of interest. A slight negative slope is observed, but it does not look very large. There are a lot of states whose capitals are less than 5% of the total population. The two outliers are Hawaii (government rank 33 and capital population 25%) and Arizona (government rank 26 and capital population 23%). Without those two the downward trend (an improvement in ranking) would be much stronger.*
>
> *... I'm not convinced ...*

To do *statistical* inference

You need a:

- null hypothesis, and alternative
- statistic computed on the data
- reference distribution on which to measure the statistic: if its extreme on this scale you would reject the null

Inference with *data plots*

You need a:

- plot description, as provided by the grammar (which is a statistic)
- plot description prescribing the null hypothesis (not the data plot itself)
- null generating mechanism, e.g. permutation, simulation from a distribution or model
- human visual system, to examine array of plots and decide if any are different from the others

Some examples

Here are several plot descriptions. What would be the null hypothesis in each?

A

```
ggplot(data) +
  geom_point(aes(x=x1, y=x2))
```

B

```
ggplot(data) +
  geom_point(aes(x=x1,
    y=x2, colour=cl))
```

C

```
ggplot(data) +
  geom_histogram(aes(x=x1))
```

D

```
ggplot(data) +
  geom_boxplot(aes(x=cl, y=x1))
```

Which plot definition would most match to a null hypothesis stating **there is no difference in the distribution between the groups?**

Some examples

Here are several plot descriptions. What would be the null hypothesis in each?

A

H_o : no association between `x1` and `x2`

B

H_o : no difference between levels of `cl`

C

H_o : the distribution of `x1` is XXX

D

H_o : no difference in the distribution of `x1` between levels of `cl`

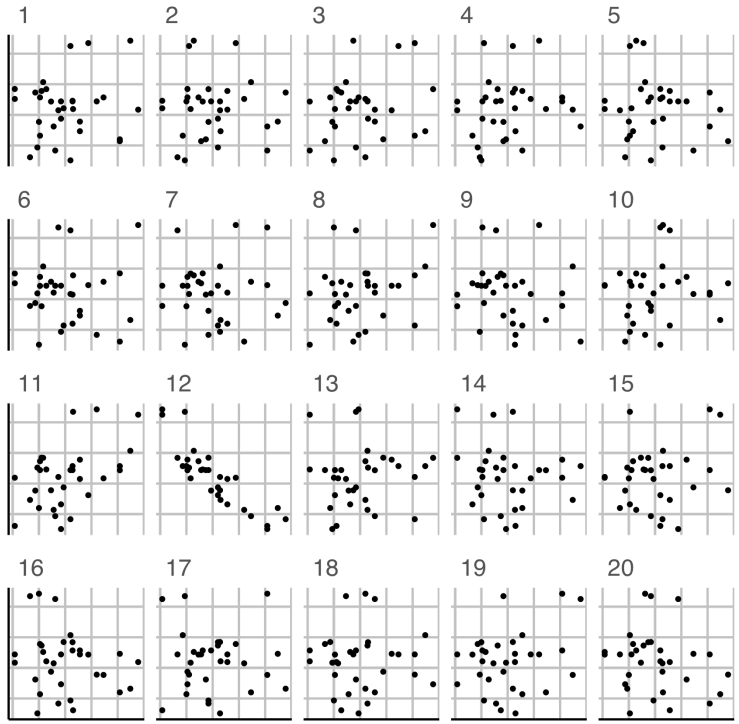
Visual inference with the nullabor

Example from the nullabor package. The data plot is embedded randomly in a field of null plots, this is a **lineup**. Can you see which one is different?

When you run the example yourself, you get a `decrypt` code line, that you run after deciding on a plot to print the location of the data plot amongst the nulls.

- plot is a scatterplot, null hypothesis is *there is no association between the two variables mapped to the x, y axes*
- null generating mechanism: permutation

```
# Make a lineup of the mtcars data, 20 plots, one is the dat
# and the others are null plots. Which one is different?
set.seed(20190709)
ggplot(lineup(null_permute('mpg'), mtcars), aes(mpg, wt)) +
  geom_point() +
  facet_wrap(~ .sample) +
  theme(axis.text=element_blank(),
        axis.title=element_blank())
```



Lineup

Embed the data plot in a field of null plots

```
library(nullabor)
pos <- sample(1:20, 1)
df_null <- lineup(
  null_permute('v1'),
  df, pos=pos)
ggplot(df_null,
       aes(x=v2,
           y=v1,
           fill=v2)) +
  geom_boxplot() +
  facet_wrap(~.sample,
             ncol=5) +
  coord_flip()
```

Ask: Which plot is the most different?

Null-generating mechanisms

- Permutation: randomizing the order of one of the variables breaks association, but keeps marginal distributions the same
- Simulation: from a given distribution, or model. Assumption is that the data comes from that model

Evaluation

- Compute p-value
- Power = signal strength

Case study 1 Temperatures of stars Part 1/2

- The data consists of the surface temperature in Kelvin degrees of 96 stars.
- We want to check if the surface temperature has an exponential distribution.
- We use histogram with 30 bins as our visual test statistic.
- For the null data, we will generate from an exponential distribution.

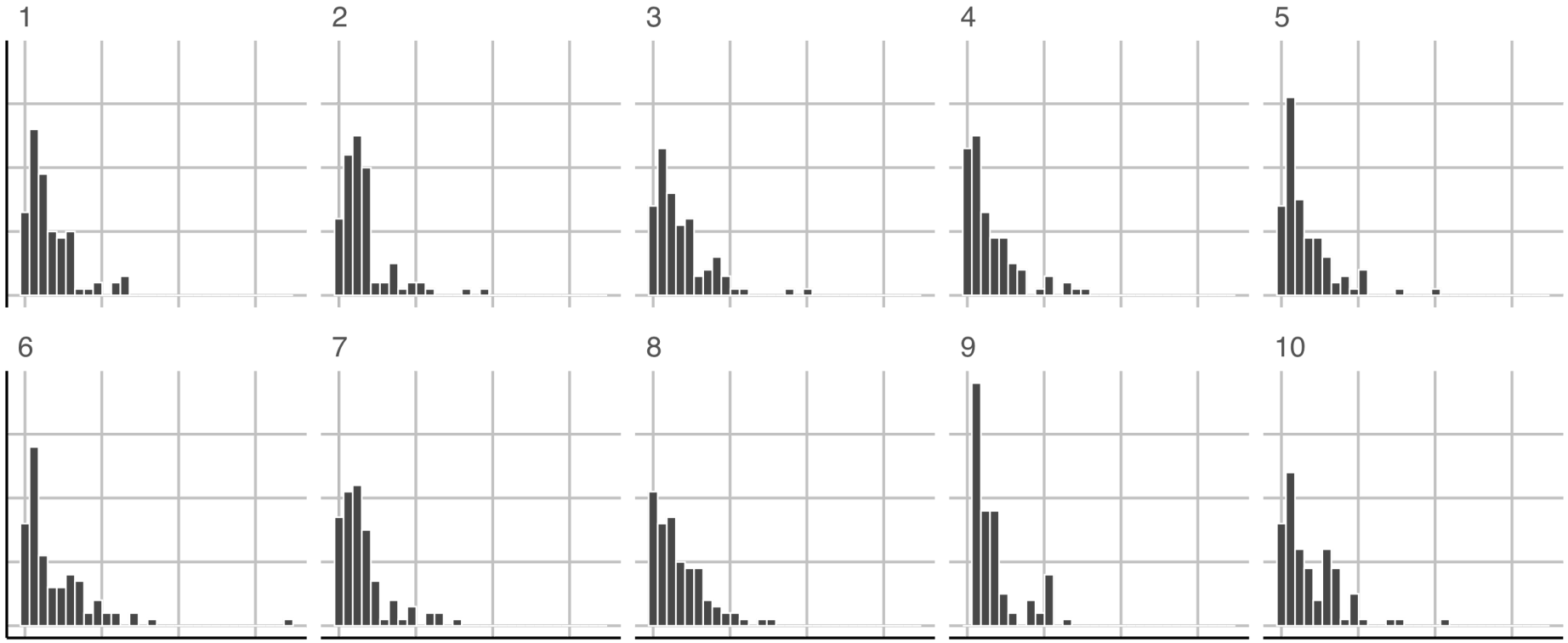
```
line_df <- lineup(null_dist("temp", "exp",
    list(rate = 1 / mean(dslabs::stars$temp))),
  true = dslabs::stars,
  n = 10
)

## decrypt("clZx bKhK oL 3OHohoOL BC")
```

- Note: the rate in an exponential distribution can be estimated from the inverse of the sample mean.

Case study 4 Temperatures of stars Part 2/2

R



Case study 2 Foreign exchange rate Part 1/2

- The data contains the daily exchange rate of 1 AUD to 1 USD between 9th Jan 2018 to 21st Feb 2018.
- Does the rate follow an ARIMA model?

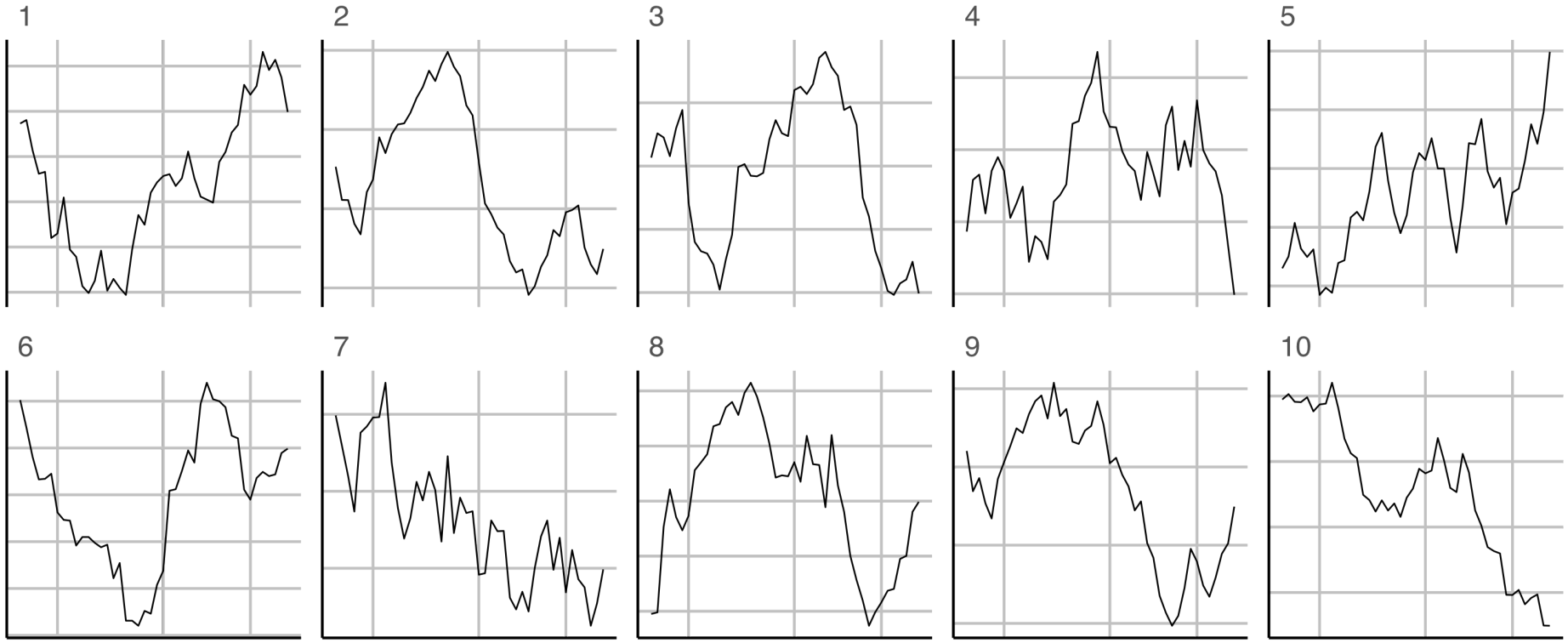
```
data(aud, package = "nullabor")
line_df <- lineup(null_ts("rate", forecast::auto.arima), true = aud, n = 10)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
## decrypt("clZx bKhK oL 3OHohoOL BY")
```

```
ggplot(line_df, aes(date, rate)) +
  geom_line() +
  facet_wrap(~.sample, scales = "free_y", nrow = 2) +
  theme(
    axis.title = element_blank(),
    axis.text = element_blank()
  )
```

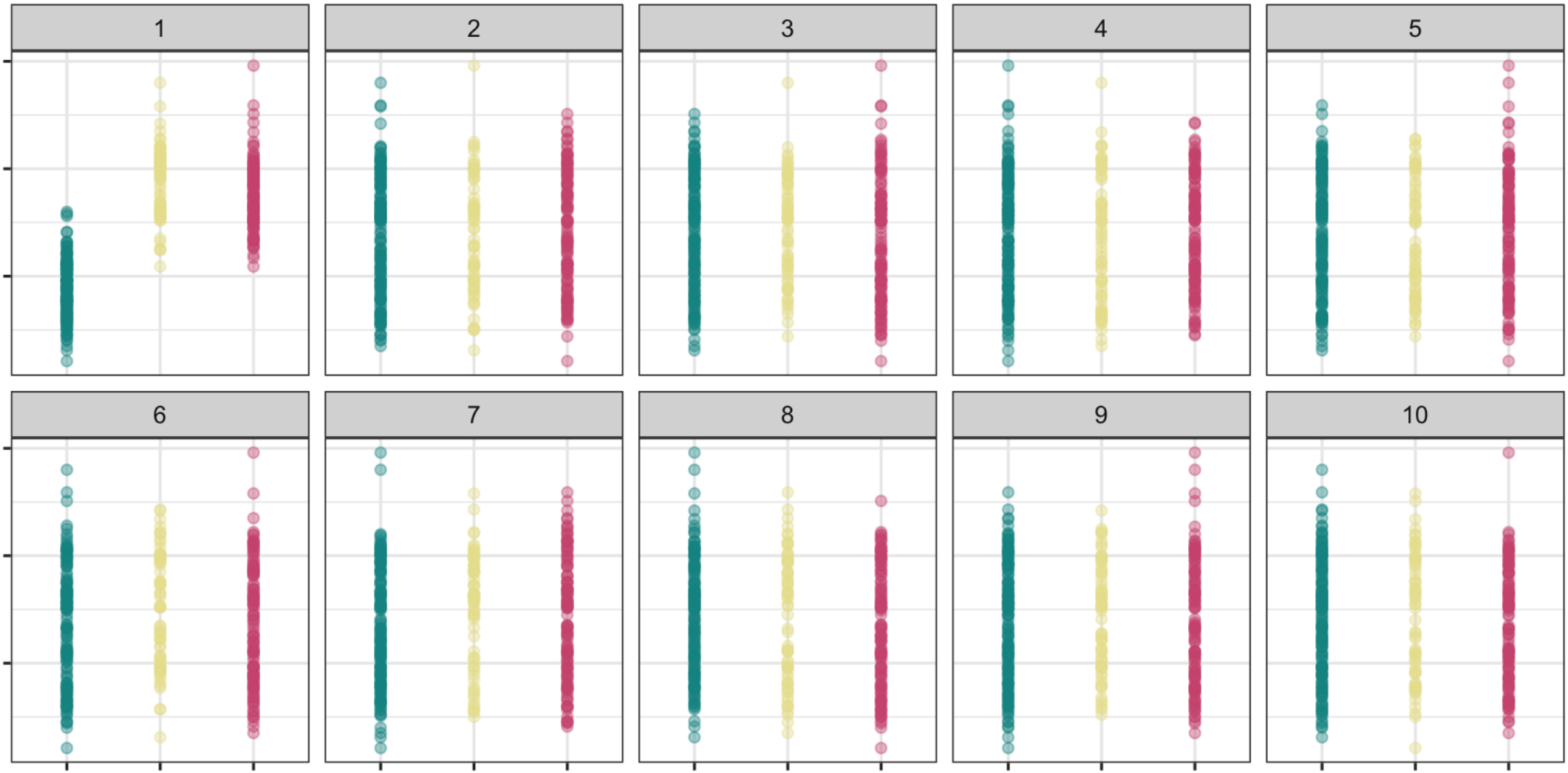
Case study 5 Foreign exchange rate Part 2/2

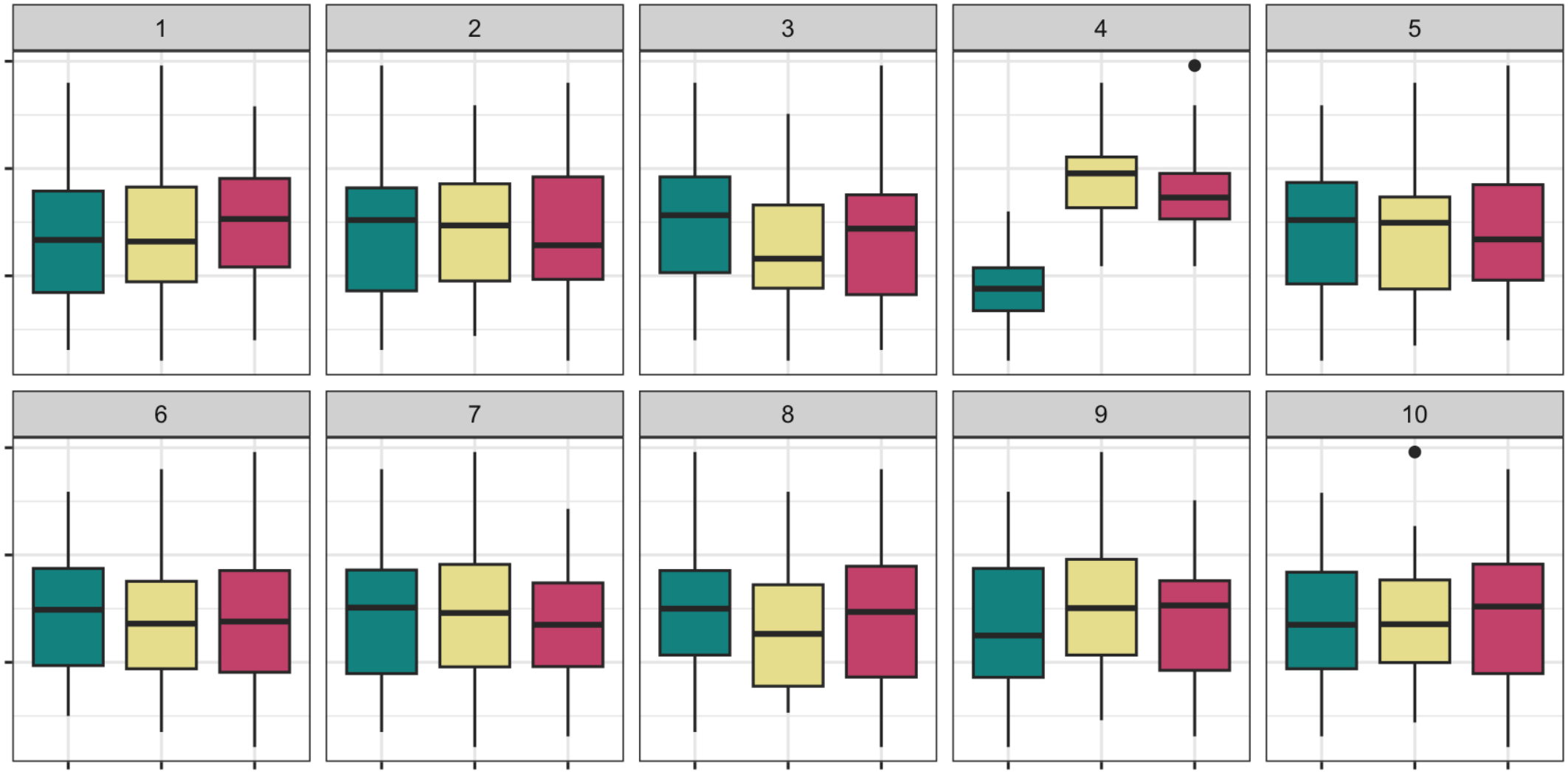


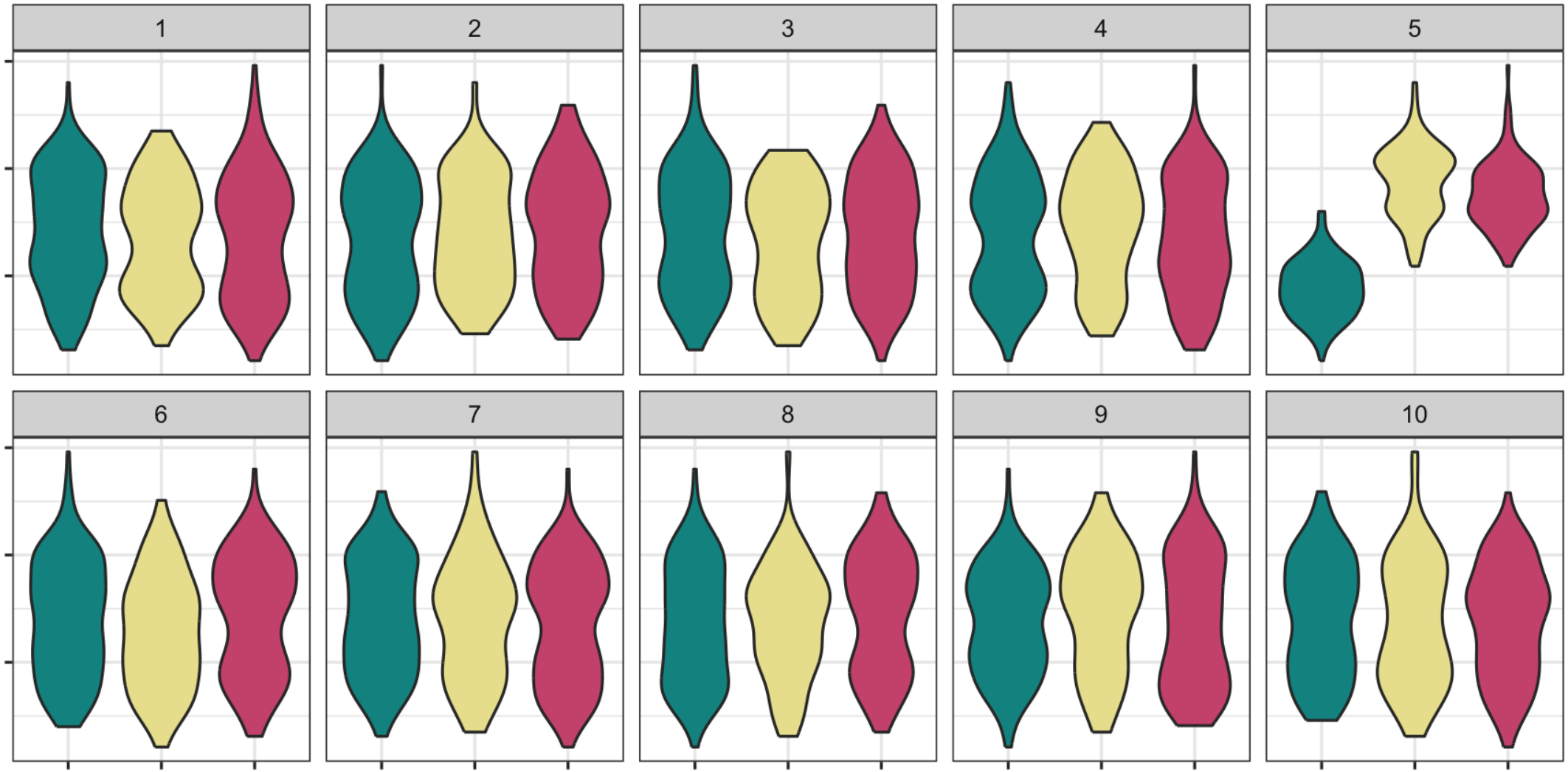
Power of a lineup

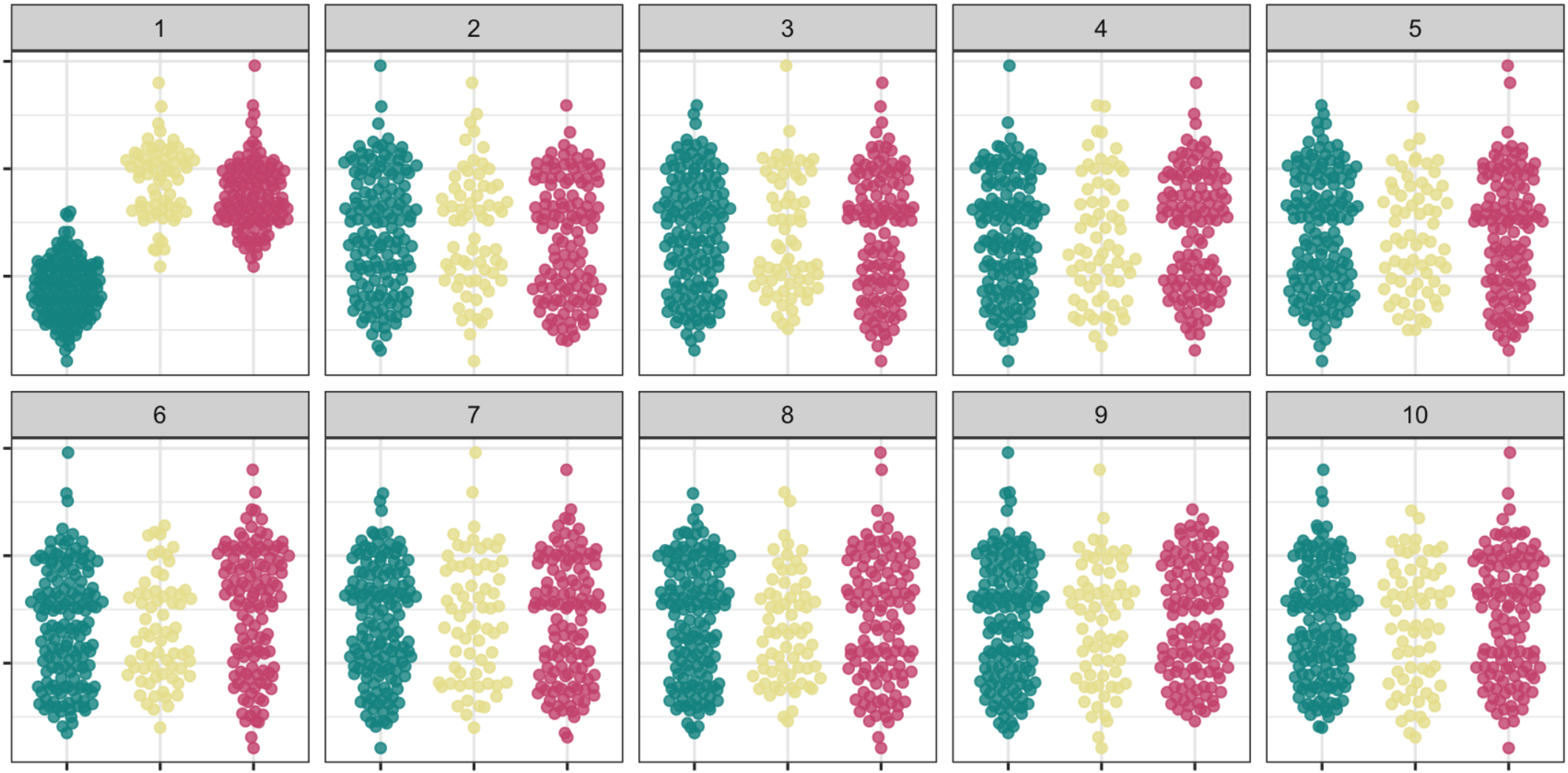
- The power of a lineup is calculated as x/n where x is the number of people who detected the data plot out of n people.
- This is useful if you want to decide which plot design is better.
- Show the same lineup made using different plots to observers (different sets of observers, the same person cannot see the same data more than once, else they may be biased).

Hofmann, H., L. Follett, M. Majumder, and D. Cook. 2012. “Graphical Tests for Power Comparison of Competing Designs.” IEEE Transactions on Visualization and Computer Graphics 18 (12): 2441–48.


1
2
3
4
5
6
7
8
9
10


1
2
3
4
5
6
7
8
9
10


1
2
3
4
5
6
7
8
9
10


1
2
3
4
5
6
7
8
9
10

Plot type	x	n	Power
`geom_point`	$x_1 = 4$	$n_1 = 23$	$x_1/n_1 = 0.174$
`geom_boxplot`	$x_2 = 5$	$n_2 = 25$	$x_2/n_2 = 0.185$
`geom_violin`	$x_3 = 6$	$n_3 = 29$	$x_3/n_3 = 0.206$
`ggbeeswarm::geom_quasirandom`	$x_4 = 8$	$n_4 = 24$	$x_4/n_4 = 0.333$

- The plot type with a higher power is preferable
- You can use this framework to find the optimal plot design

Some considerations in visual inference

- In practice you don't want to bias the judgement of the human viewers so for a proper visual inference:
  - you should *not* show the data plot before the lineup
  - you should *not* give the context of the data
  - you should remove labels in plots
- You can crowd source these by paying for services like:
  - Amazon Mechanical Turk,
  - Appen (formerly Figure Eight) and
  - LABVANCED.
  - prolifico.
- If the data is for research purposes, then you may need ethics approval for publication.

Resources and Acknowledgement

- Buja, Andreas, Dianne Cook, Heike Hofmann, Michael Lawrence, Eun-Kyung Lee, Deborah F. Swayne, and Hadley Wickham. 2009. “Statistical Inference for Exploratory Data Analysis and Model Diagnostics.” Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences 367 (1906): 4361–83.
- Wickham, Hadley, Dianne Cook, Heike Hofmann, and Andreas Buja. 2010. “Graphical Inference for Infovis.” IEEE Transactions on Visualization and Computer Graphics 16 (6): 973–79.
- Hofmann, H., L. Follett, M. Majumder, and D. Cook. 2012. “Graphical Tests for Power Comparison of Competing Designs.” IEEE Transactions on Visualization and Computer Graphics 18 (12): 2441–48.
- Majumder, M., Heiki Hofmann, and Dianne Cook. 2013. “Validation of Visual Statistical Inference, Applied to Linear Models.” Journal of the American Statistical Association 108 (503): 942–56.

- Data coding using `tidyverse` suite of R packages
- 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 4 - Session 2