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: *Emi Tanaka*

ETC5521.Clayton-x@monash.edu

Week 11 - Session 2

Visual inference with the nullabor 

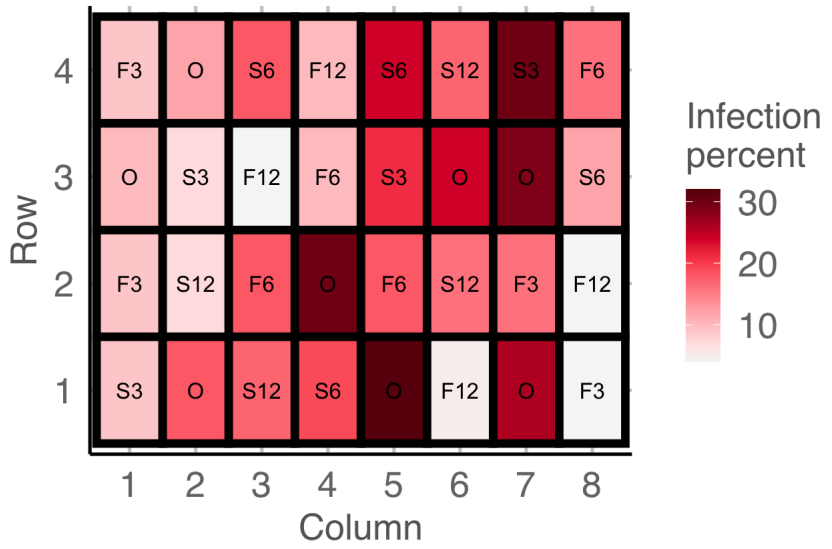
`nullabor` + `ggplot2`

- You can construct the null data "by hand" as you have done for Exercise 4 (d) in tutorial 9.
- You will then need to create null plots and then randomly place the data plot to present the lineup.
- You'll need to know which one is the data plot so you can tell if viewer's chose the data plot or not.
- The `nullabor` package makes it easy to create the data for the lineup and you can use `ggplot2` to construct the lineup.

```
library(nullabor)
library(tidyverse) # which includes ggplot2
```

Case study 2 Potato scab infection Part 1/4

data R



- Experiment was conducted to investigate the effect of sulfur on controlling scab disease in potatoes.
- There were seven treatments in total: control plus spring and fall application of 300, 600 or 1200 lbs/acres of sulfur.
- Employs a completely randomised design with 8 replications for control and 4 replications for other treatments.

W.G. Cochran and G. Cox, 1957. Experimental Designs, 2nd ed. John Wiley, New York.

Case study 2 Potato scab infection Part 2/4

- We are testing $H_0 : \mu_1 = \mu_2 = \ldots = \mu_7$ vs. H_1 : at least one mean is different to others.
- Here we don't have to many observation per treatment so we can use a dotplot.
- For the method to generate null, we consider permuting the treatment labels.

```
method <- null_permute("trt")
```

- Then we generate the null data, also embedding the actual data in a random position. Make sure to `set.seed` to get the same random instance.

```
set.seed(1)
line_df <- lineup(method, true = cochran.crd, n = 10)

## decrypt("bhMq KJPJ 62 sSQ6P6S2 ua")
```

Case study 2 Potato scab infection Part 3/4

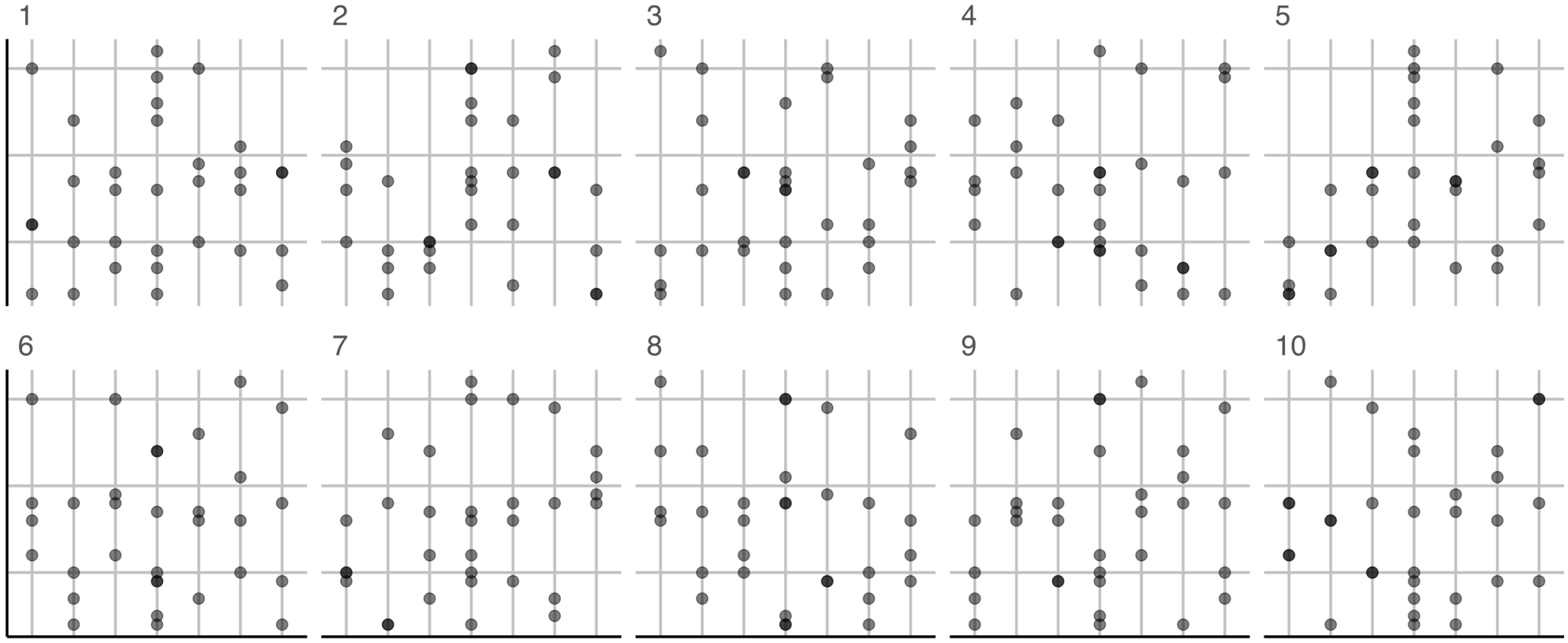
```
glimpse(line_df)
```

```
## Rows: 320
## Columns: 5
## $ inf     <int> 9, 12, 18, 10, 24, 17, 30, 16, 10, 7, 4, 10, 21, 24, 29, 12, 9, 7, 18, 30, 18
## $ trt     <fct> S3, F12, S3, F3, 0, F3, F12, 0, S12, F6, 0, F6, S3, F3, 0, F12, 0, 0, S6, S12
## $ row     <int> 4, 4, 4, 4, 4, 4, 4, 4, 3, 3, 3, 3, 3, 3, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1,
## $ col     <int> 1, 2, 3, 4, 5, 6, 7, 8, 1, 2, 3, 4, 5, 6, 7, 8, 1, 2, 3, 4, 5, 6, 7, 8, 1, 2,
## $ .sample <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
```

- The `.sample` variable has information of which sample it is.
- One of the `.sample` number belongs to the real data.

```
line_df %>%
  ggplot(aes(trt, inf)) +
  geom_point(size = 3, alpha = 1/2) +
  facet_wrap(~.sample, nrow = 2) +
  theme(axis.text = element_blank(), # remove data context
        axis.title = element_blank())
```

Case study 2 Potato scab infection Part 4/4

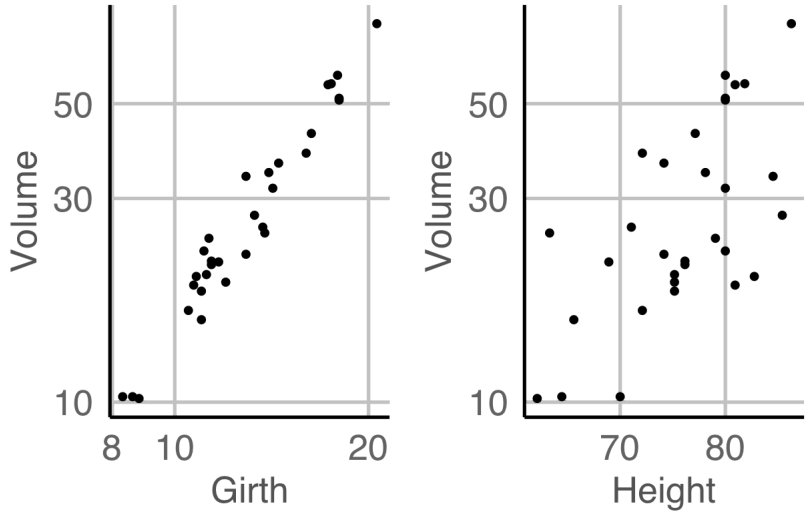



```
decrypt("bhMq KJPJ 62 sSQ6P6S2 ua")

## [1] "True data in position  5"
```

Case study 3 Black Cherry Trees Part 1/4

data R



- Data measures the diameter, height and volume of timber in 31 felled black cherry trees.
- We fit the model

```
fit <- lm(log(Volume) ~ log(Girth) + log(Height),
          data = trees)

fit_df <- trees %>%
  # below are needed for lineup
  mutate(.resid = residuals(fit),
         .fitted = fitted(fit))
```

Case study 3 Black Cherry Trees Part 2/4

- We are testing H_0 : errors are $NID(0, \sigma^2)$ vs. H_1 : errors are not $NID(0, \sigma^2)$.
- We will use the residual plot as the visual statistic.
- For the method to generate null, we generate residuals from random draws from $N(0, \hat{\sigma}^2)$.

```
method <- null_lm(log(Volume) ~ log(Girth) + log(Height),
                  method = "pboot")
```

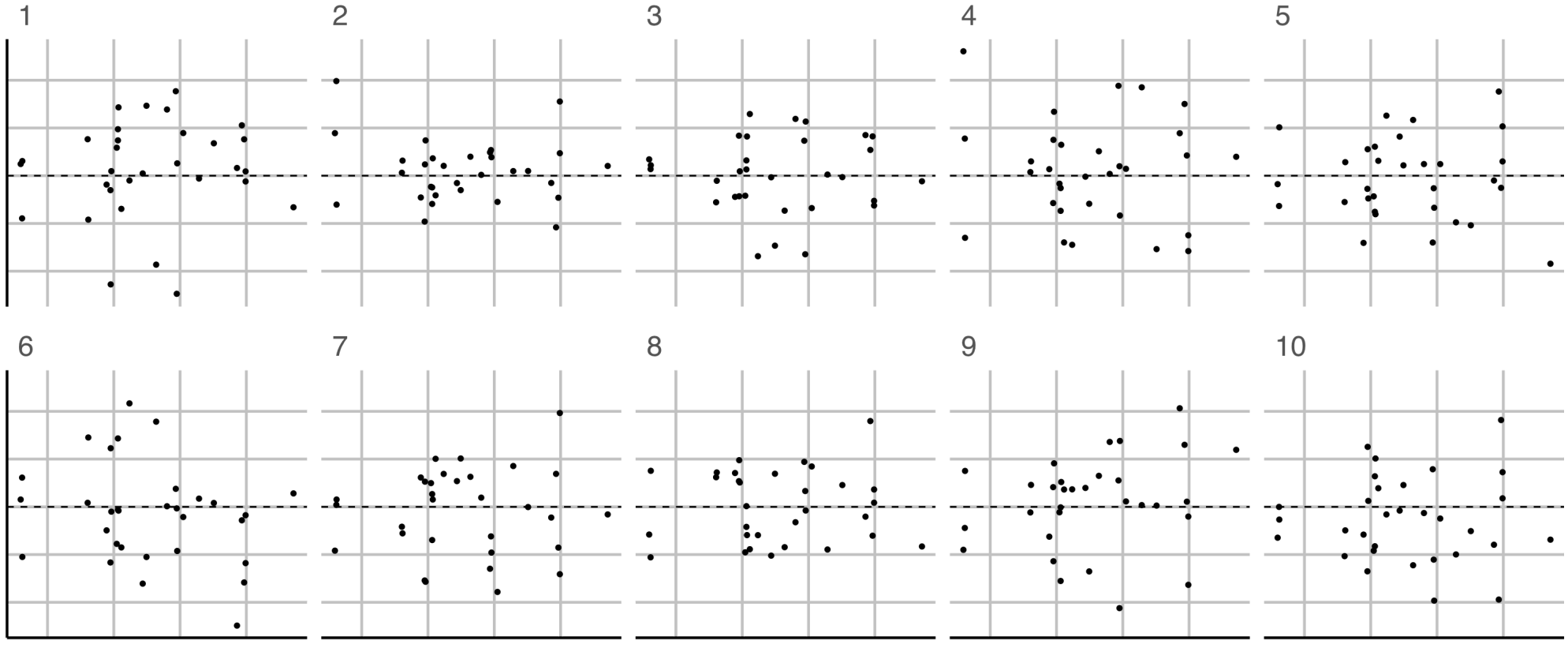
- Then we generate the lineup data.

```
set.seed(2020)
line_df <- lineup(method, true = fit_df, n = 10)

## decrypt("bhMq KJPJ 62 sSQ6P6S2 uT")
```

Case study 3 Black Cherry Trees Part 3/4

R



Case study 3 Black Cherry Trees Part 4/4

- We can have:
  - `method = "pboot"`,
  - `method = "boot"` or
  - `method = "rotate"`
    for different (and valid) methods to generate null data when fitting a linear model.

```
method <- null_lm(log(Volume) ~ log(Girth) + log(Height),
                  method = "boot")
```

- We can also consider using a different visual statisitc, e.g. QQ-plot to assess normality.

Case study 4 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.

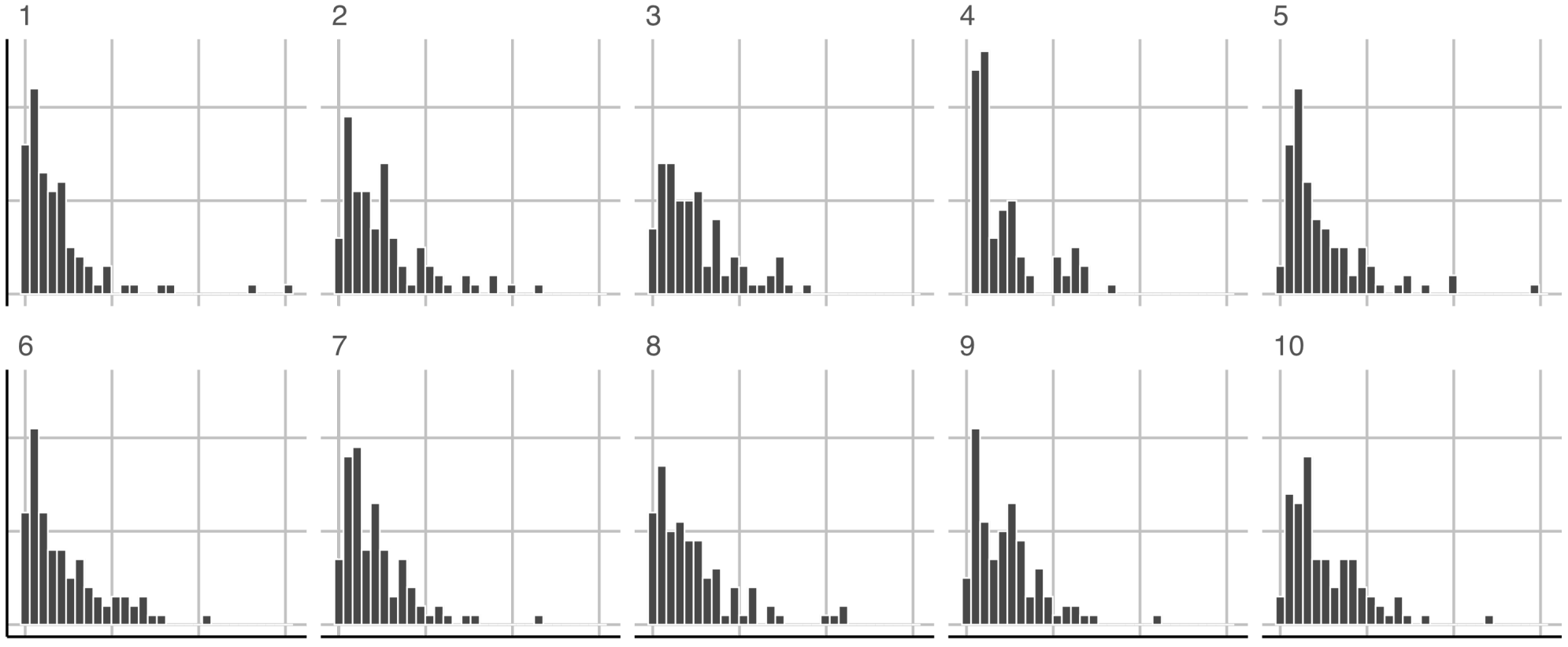
```
line_df <- lineup(null_dist("temp", "exp", list(rate = 1/mean(dslabs::stars$temp
                  true = dslabs::stars,
                  n = 10)

## decrypt("bhMq KJPJ 62 sSQ6P6S2 ug")
```

- 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 5 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?

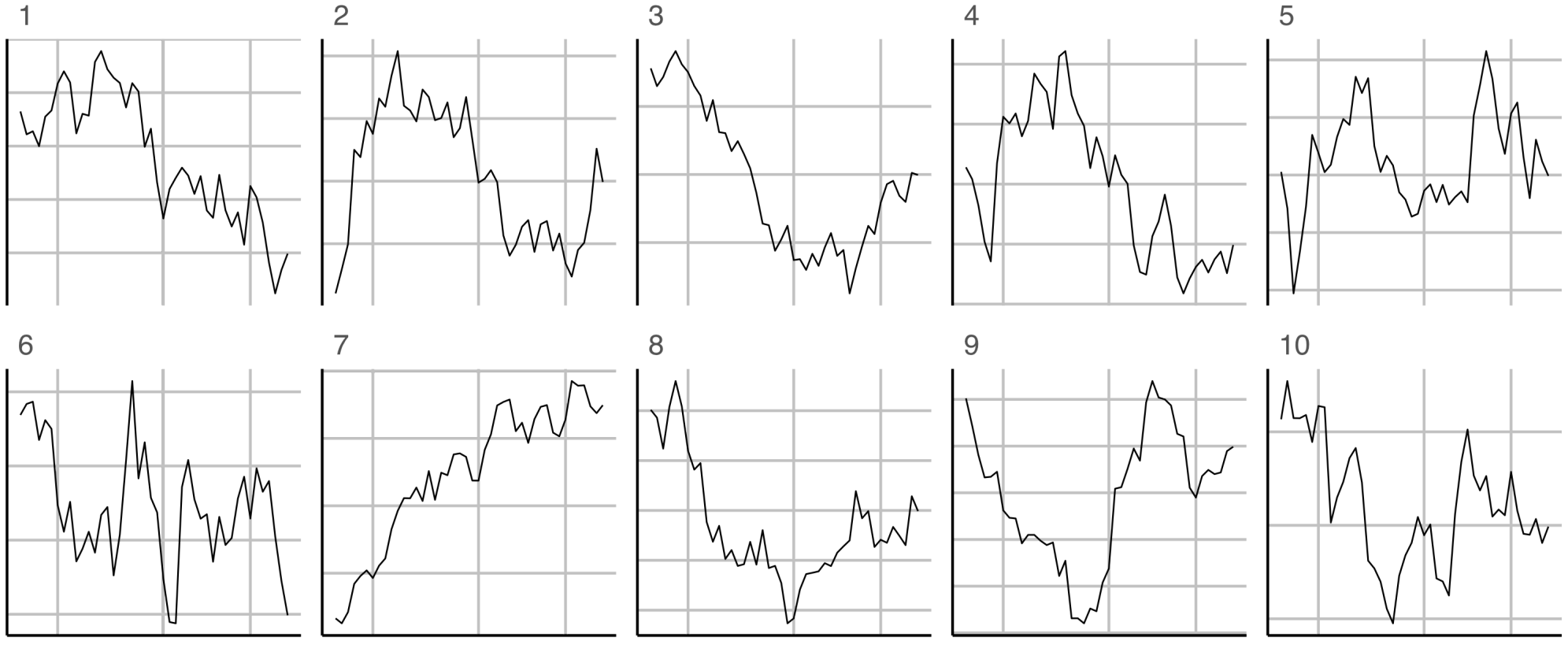
```
data(aud, package = "nullabor")
line_df <- lineup(null_ts("rate", forecast::auto.arima), true = aud, n = 10)

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

## decrypt("bhMq KJPJ 62 sSQ6P6S2 um")

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



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 constructed with `xaringan`, remark.js, `knitr`, and R Markdown.

MONASH University

Lecturer: *Emi Tanaka*

ETC5521.Clayton-x@monash.edu

Week 11 - Session 2