MONASH University

ETC5521: Exploratory Data Analysis

Extending beyond the data, what can and cannot be inferred more generally, given the data collection

Lecturer: *Emi Tanaka*

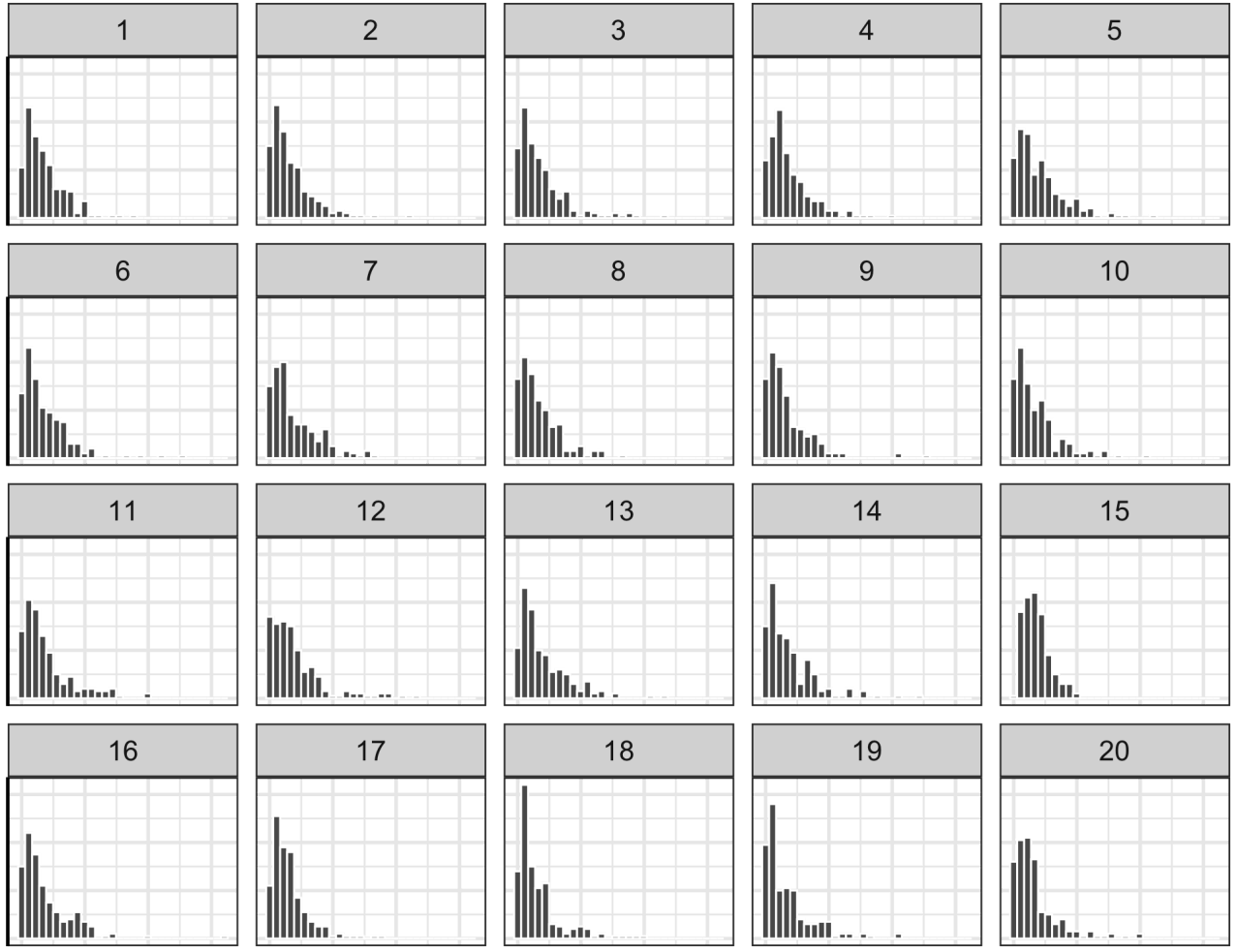
ETC5521.Clayton-x@monash.edu

Week 12 - Session 2

Sample size calculation

How many people should you survey?

data R

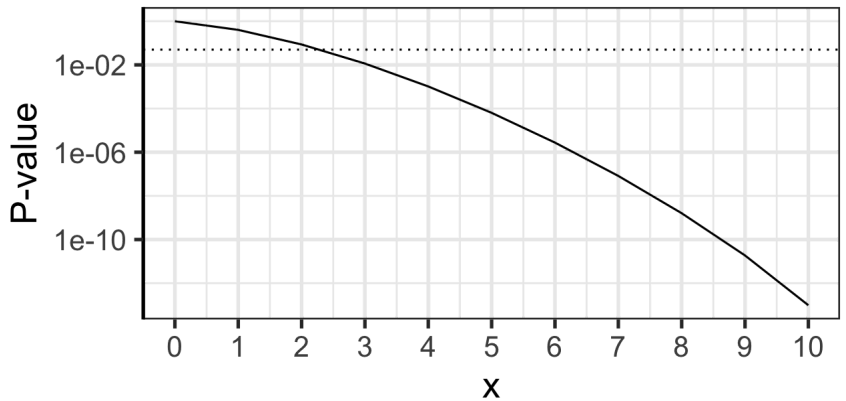


- Here we are testing $H_0 : Y \sim exp(\lambda)$.
- Suppose we only have one person to assess the lineup.
- If there is only a single response, then there are only two scenarios possible:
  - **Scenario 1**: the person detects the data plot
  - **Scenario 2**: the person does *not* detect the data plot
- The visual inference p-value under:
  - **Scenario 1** is 0.05
  - **Scenario 2** is 1
- Neither scenario yield p-values < 0.05!

Power of a binary hypothesis test

The statistical **power** of a binary hypothesis test is the probability that the test correctly rejects the null hypothesis (H_0) when a *specific* alternative hypothesis (H_1) is true.

- Since $m \geq 2$, i.e. under H_0, $0 < p = 1/m \leq 0.5$.
- Recall visual inference p-value is $P(X \geq x) = \sum_{k=x}^{n} \binom{n}{k} (1/m)^k (1 - 1/m)^{n-k}$.
- So for $m = 20$ and $n = 10$,



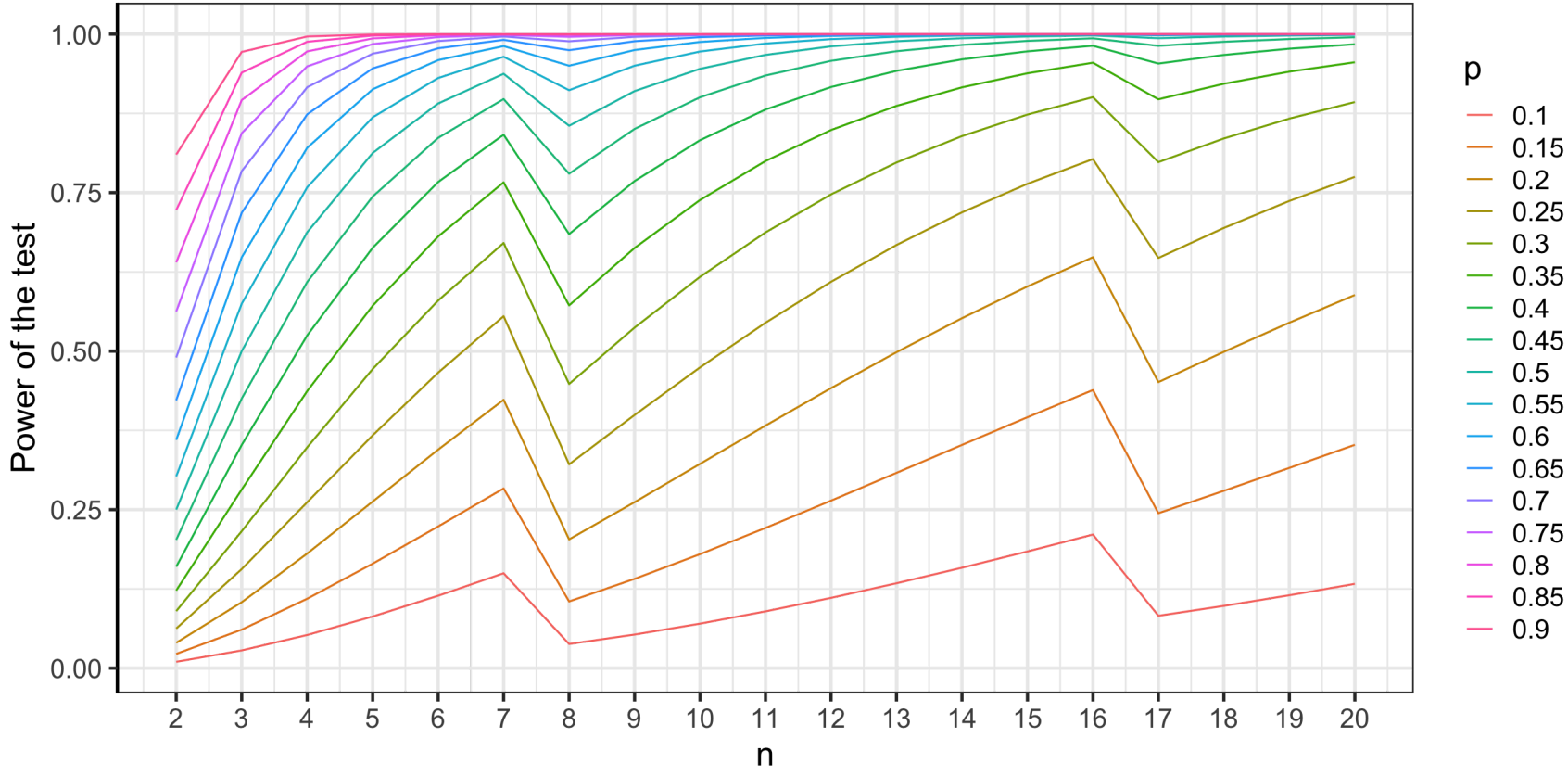
- So if we have $X > 2$, then p-value < 0.05.
- Suppose then the true detection probability is 0.9, therefore H_1 is true.
- Under $p = 0.9$,

$$P(X > 2) = \sum_{k=3}^{10} 0.9^k 0.1^{(10-k)} = 0.9999996$$

- Therefore the power of the test is 0.9999996 if $p = 0.9$.

Power analysis

- Let's suppose H_1 is true and that specifically $p = 0.9$.
- Let's fix $m = 20$ and reject H_0 if p-value $< \alpha = 0.05$.



Estimating the detection probability p

- For a fixed power $(1-\beta)$, the minimum sample size n we need depends on the detection probability p
- Generally if p is larger, less n is sufficient to get equivalent or larger power.
- But we don't know what the true p is! (If we did, we don't need to test for it.)
- Either you will need to make a guess from past experience or run a pilot test.
- If you find in the pilot test, x_p out of n_p participants detected the data plot then an estimate of $\hat{p} = x_p/n_p$.

Sample size calculation

- The sample size calculation depends on:
  - the selected false positive rate (α)
  - the detection probability p
  - the number of plots in the lineup m
  - the minimum power $(1 - \beta)$ desired
  - the expected dropout rate d (i.e. proportion of samples that cannot be used due to incomplete results or other quality issues)
- Say if $\alpha = 0.05, p = 0.1, m = 20, d = 0.95$ and at least 80% power is desired then at least 178 samples is required.

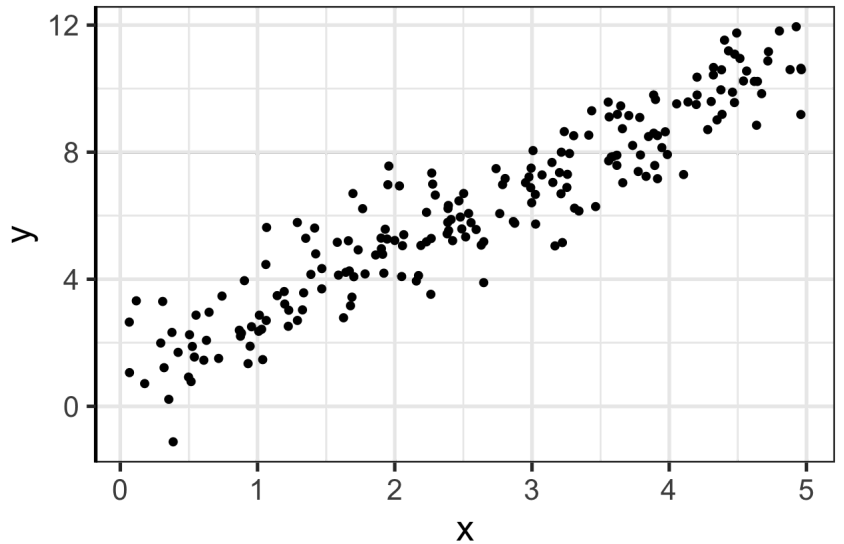
```
p <- 0.1
m <- 20
d <- 0.95
power_df <- tibble(n = 2:200) %>%
  mutate(power = map_dbl(n, function(n) {
      x <- 1:n
      pval <- map_dbl(x, ~1 - pbinom(.x - 1, n, 1/m))
      xmin <- x[which.max(pval < alpha)]
      1 - pbinom(xmin - 1, n, p)
    }))

power_df %>%
  filter(power > 0.8) %>%
  pull(n) %>%
  min() %>%
  magrittr::divide_by(d) %>%
  ceiling()


## [1] 178
```

Simulating from the null distribution

Recap: Simulating data from parametric models

- Recall in lecture 8, we studied how to simulate data from parametric models.

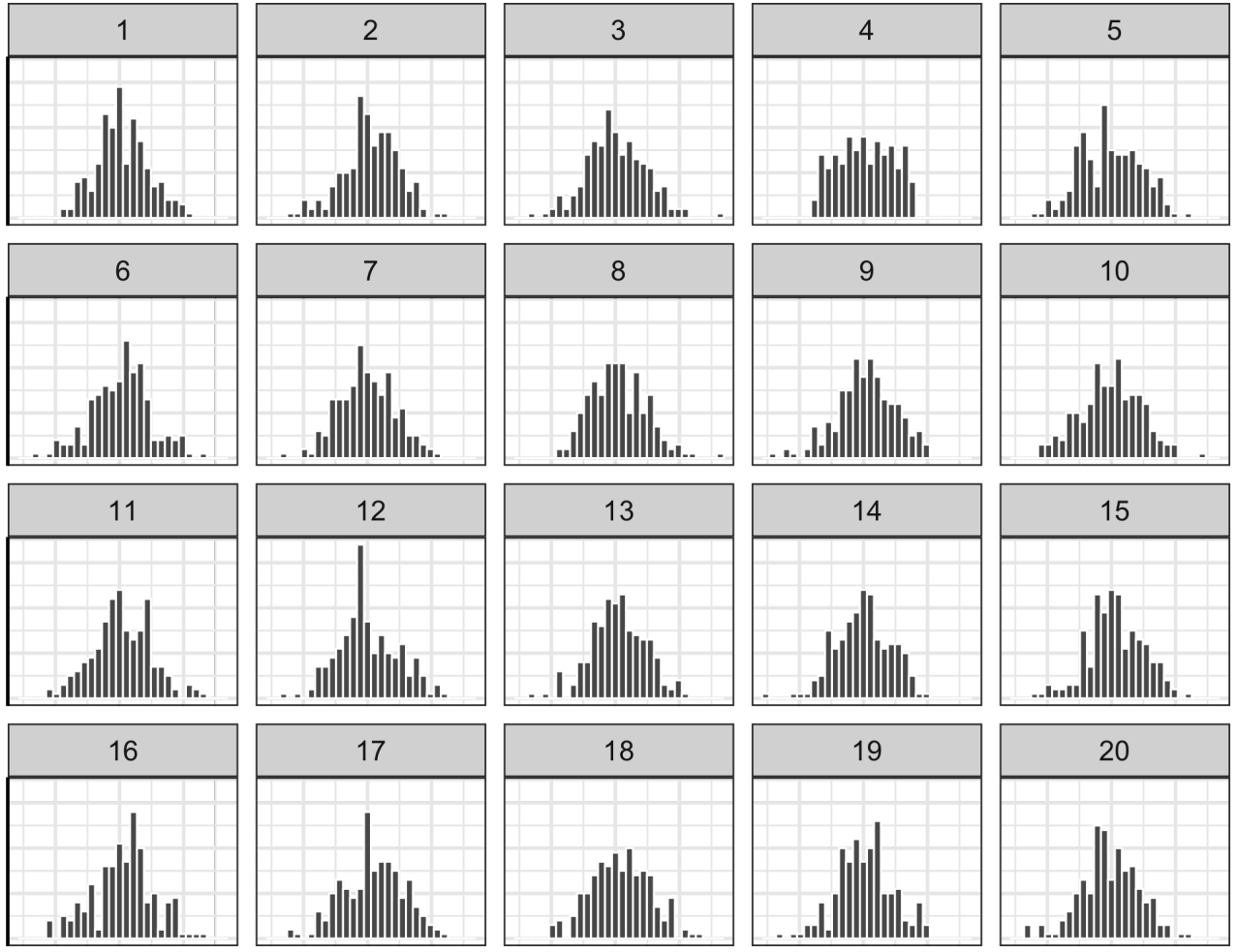
```
set.seed(1)
df1 <- tibble(id = 1:200) %>%
  mutate(x = runif(n(), 0, 5),
         y = 2 * x + 1 + rnorm(n()))

ggplot(df1, aes(x, y)) + geom_point()
```



- We also briefly discussed how to simulate data from the null distribution in lecture 11.

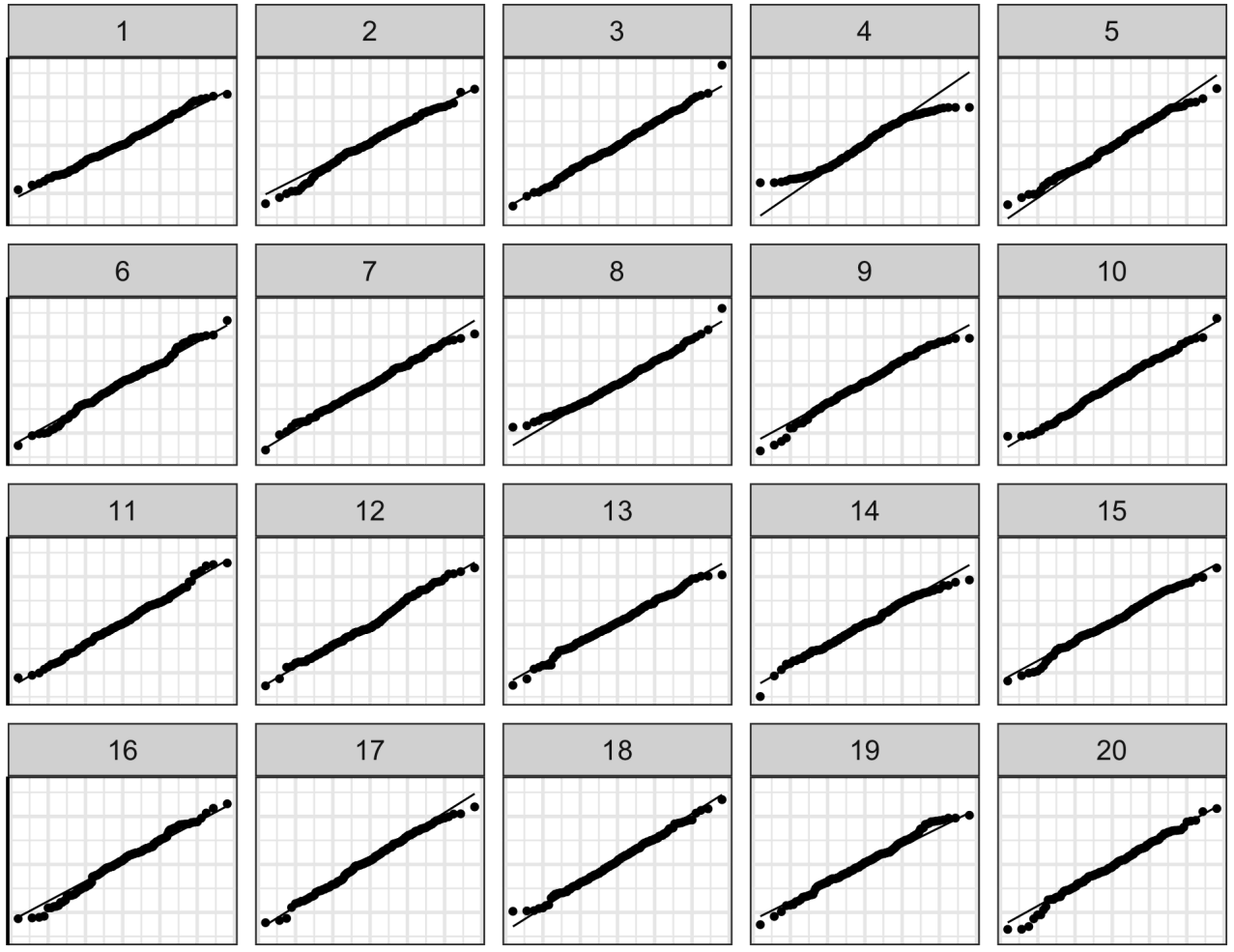
Case study 1 Testing for normality

data R



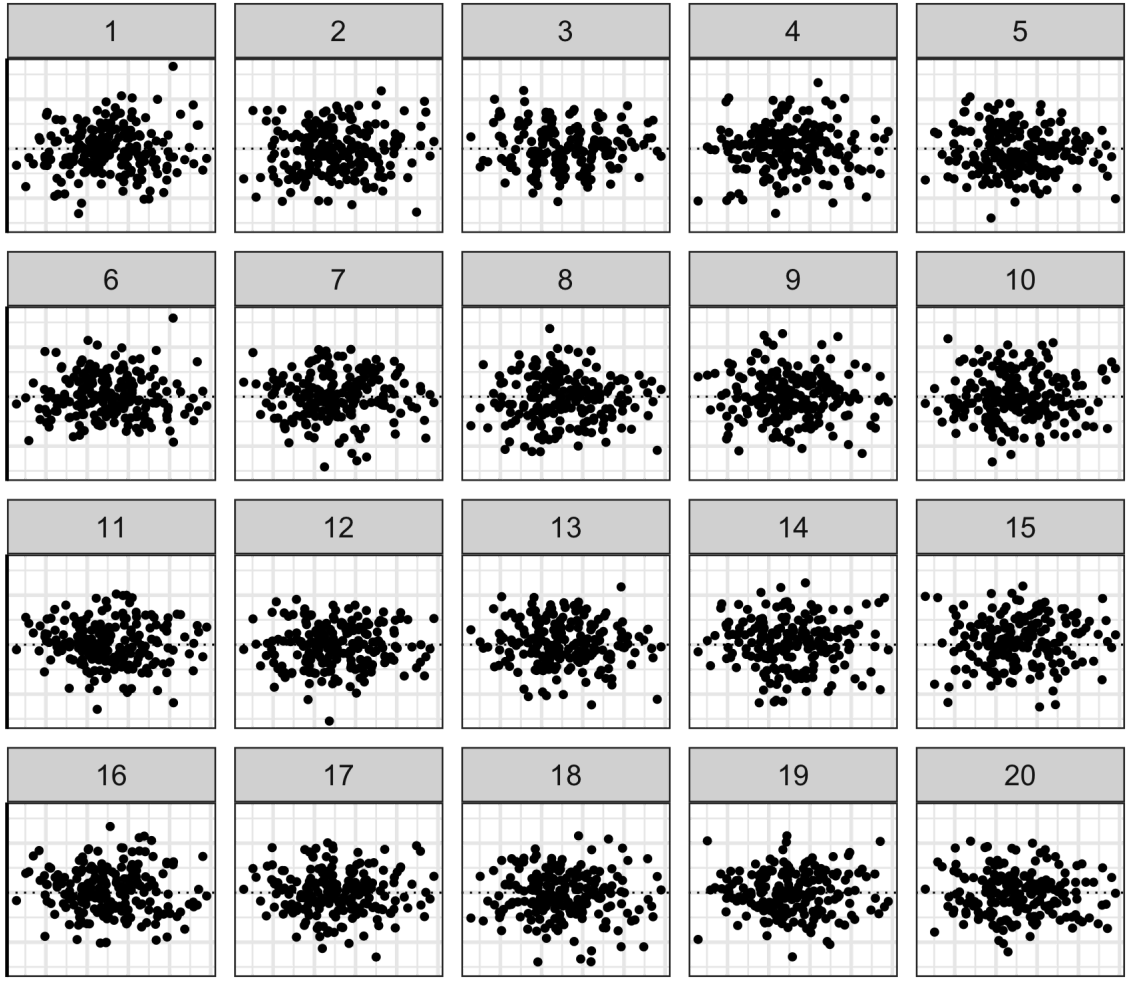
- We are testing $H_0 : Y \sim N(\mu, \sigma^2)$.
- An estimate of $\hat{\mu} = \bar{Y}$ is estimated the sample mean
- An estimate of $\hat{\sigma} = sd(Y)$ is estimated the sample standard deviation
- A null data here is simply simulated from $N(\hat{\mu}, \hat{\sigma})$.

Case study 2 Testing for a distribution



- It is easier to compare a distribution using Q-Q plot
- Plot 4 is in indeed the data plot.
- In fact the data is generated from a uniform distribution.

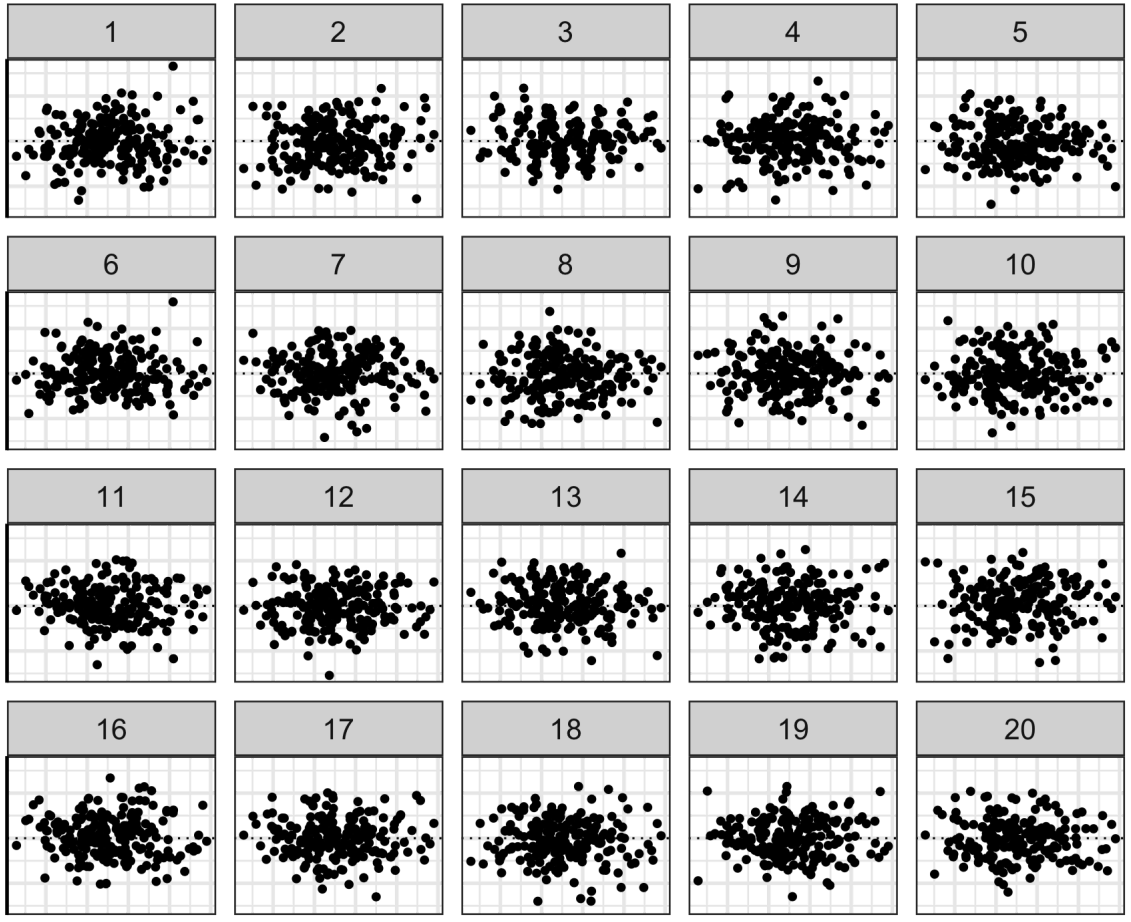
Case study 3 Checking if there is a pattern in residual plot



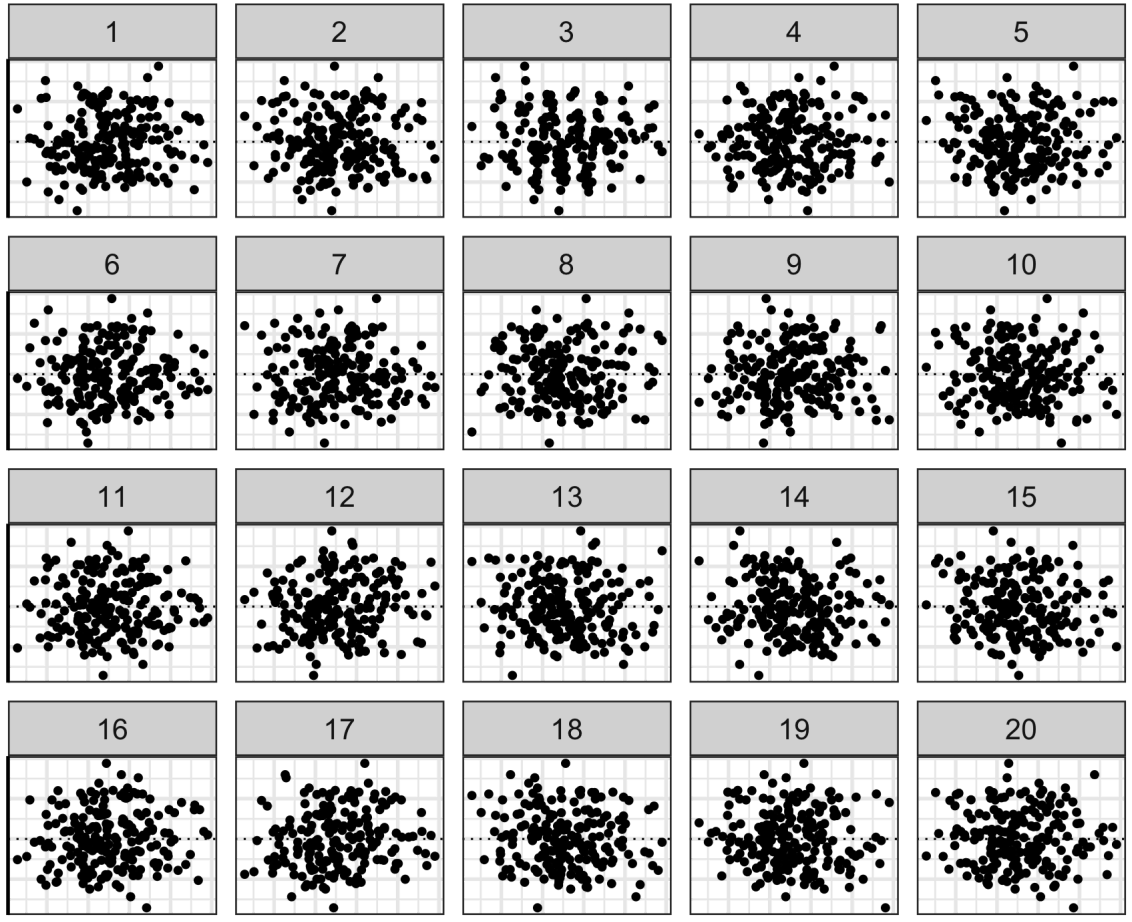
- In the left lineup, we are testing the data plot to see if there is any pattern.
- When the null distribution is imprecise, for example in search of a pattern in residual plot, you need to choose a null generation method that mimics an appropriate distribution under the null.

Selecting an appropriate null generation method

Parametric bootstrap

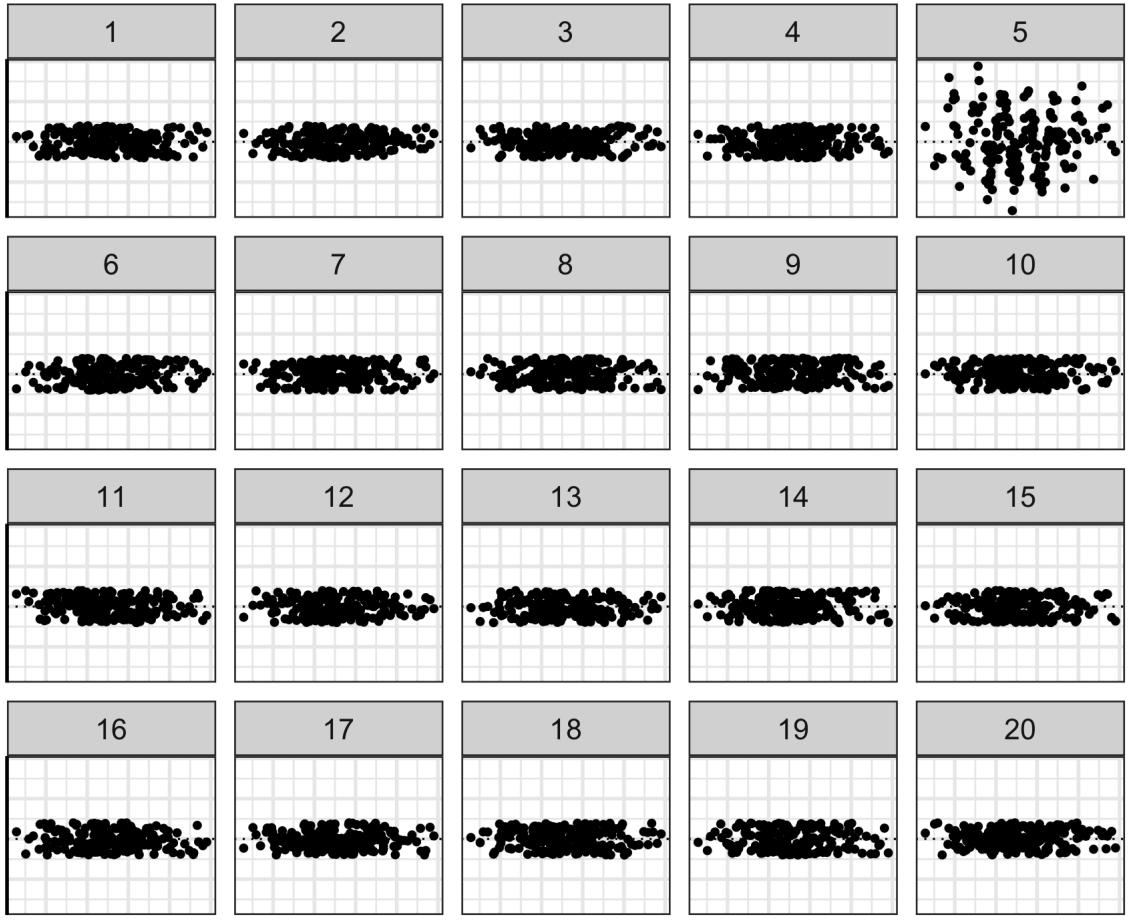


(Non-parametric) Bootstrap



Mis-specifying the null distribution

Null distribution is set to a uniform distribution



- If the null distribution is mis-specified, this can make the detection probability larger.
- This however can result in an incorrect conclusion.

While today's focus was on data collection from visual inference surveys, concepts such as data quality checks and sufficient sample size to draw inference is applicable to other data collection.

There's always more to learn but **stay curious** and make sure you **plot your data** before rushing off to fitting some models!

MONASH University

Lecturer: *Emi Tanaka*

ETC5521.Clayton-x@monash.edu

Week 12 - Session 2