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*
Department of Econometrics and Business Statistics
ETC5521.Clayton-x@monash.edu
Week 11 - Session 1

Revisiting hypothesis testing

Testing coin bias Part 1/2

- Suppose I have a coin that I'm going to flip 
- If the coin is unbiased, what is the probability it will show heads?
- Yup, the probability should be 0.5.
- So how would I test if a coin is biased or unbiased?
- We'll collect some data.
- **Experiment 1**: I flipped the coin 10 times and this is the result:

- The result is 7 head and 3 tails. So 70% are heads.
- Do you believe the coin is biased based on this data?

Testing coin bias Part 2/2

- **Experiment 2**: Suppose now I flip the coin 100 times and this is the outcome:

- We observe 70 heads and 30 tails. So again 70% are heads.
- Based on this data, do you think the coin is biased?

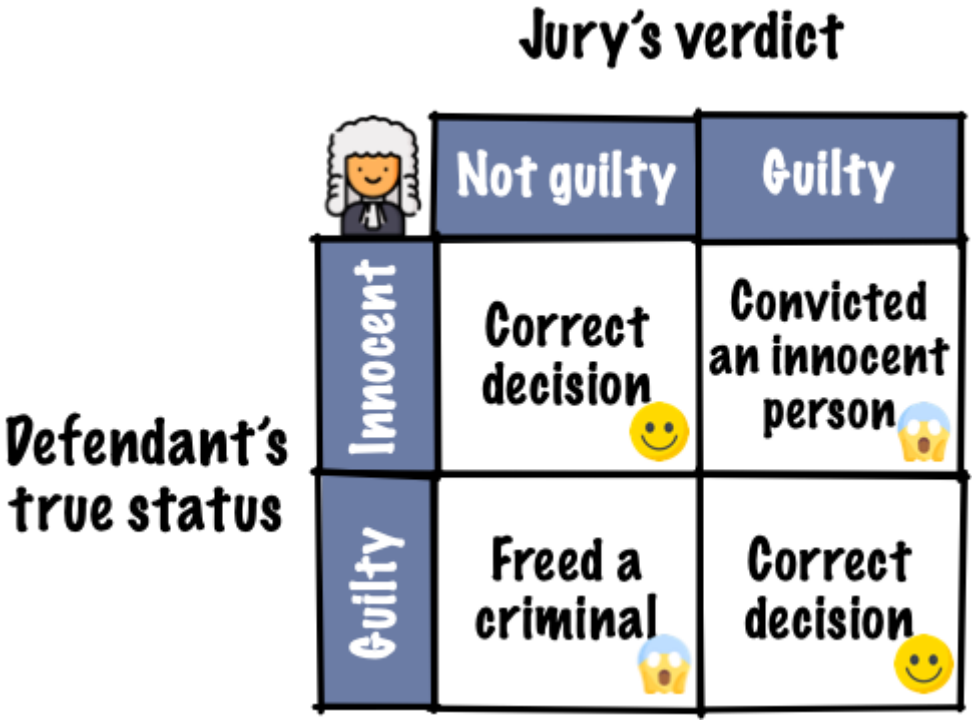
(Frequentist) hypotheses testing framework

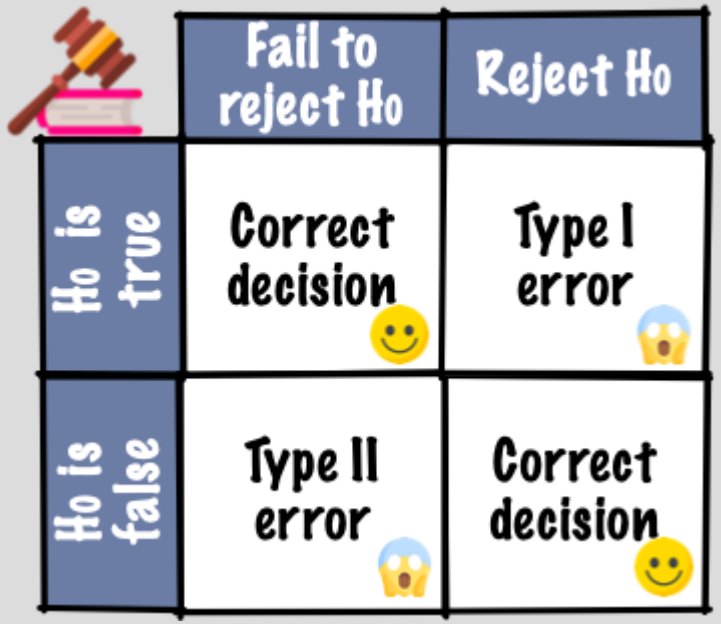
- Suppose X is the number of heads out of n independent tosses.
- Let p be the probability of getting a head for this coin.

Hypotheses	$H_0 : p = 0.5$ vs. $H_1 : p \neq 0.5$
Assumptions	Each toss is independent with equal chance of getting a head.
Test statistic	$X \sim B(n, p)$. Recall $E(X) = np$. The observed test statistic is denoted x.
P-value (or critical value or confidence interval)	$P(\mid X - np \mid \geq \mid x - np \mid)$
Conclusion	Reject null hypothesis when the p-value is less than some significance level α. Usually $\alpha = 0.05$.

- The p-value for experiment 1 is $P(|X - 5| \geq 2) \approx 0.34$.
- The p-value for experiment 2 is $P(|X - 50| \geq 20) \approx 0.00008$.

Judicial system





- Evidence by test statistic
- Judgement by p-value, critical value or confidence interval

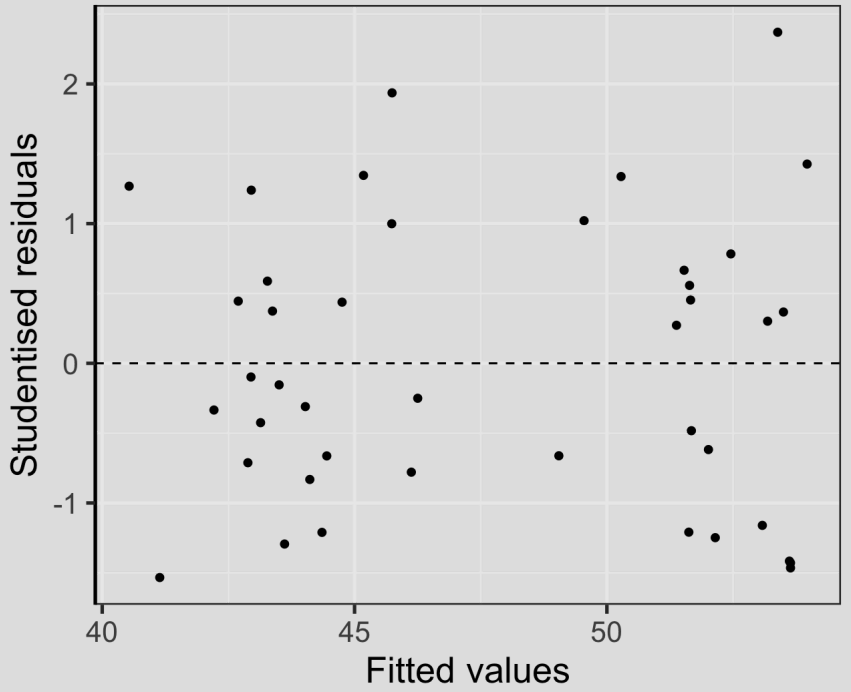
Does the test statistic have to be a *numerical summary statistics*?

Visual inference

Visual inference

- Hypothesis testing in visual inference framework is where:
  - 🔍 the *test statistic is a plot* and
  - 🔨 judgement is by human perceptions.
- You (and many other people) actually do visual inference many times but generally in an informal fashion.
- Here, we are making an inference on whether the residual plot has any patterns based on a single data plot.

From Exercise 4 in week 9 tutorial: a residual plot after modelling high-density lipoprotein in human blood.



- Data plots tend to be over-interpreted
- Reading data plots require calibration

Visual inference more formally

1. State your null and alternate hypotheses.
2. Define a **visual test statistic**, $V(.)$, i.e. a function of a sample to a plot.
3. Define a **null generating method** to generate **null data**, $\boldsymbol{y}_0$.
4. $V(\boldsymbol{y})$ maps the actual data, $\boldsymbol{y}$, to the plot. We call this the **data plot**.
5. $V(\boldsymbol{y}_0)$ maps a null data to a plot of the same form. We call this the **null plot**. We repeat this $m-1$ times to generate $m-1$ null plots.
6. A **lineup** displays these m plots in a random order.
7. Ask n human viewers to select a plot in the lineup that looks different to others without any context given.

Suppose x out of n people correctly identified the data plot from a lineup, then

- the **visual inference p-value** is given as

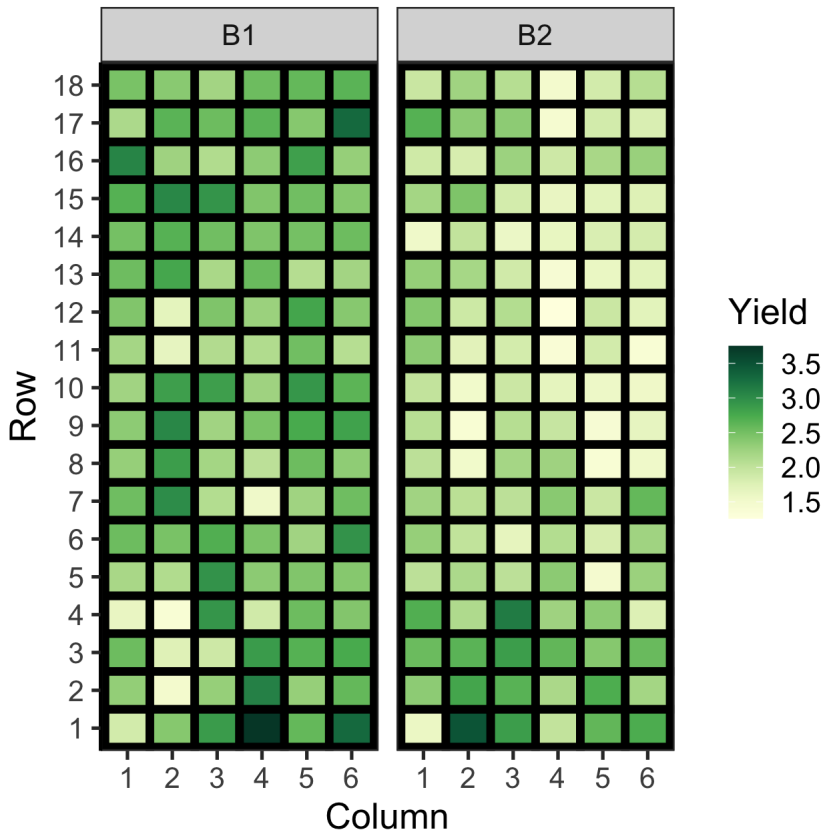
$$P(X \geq x)$$

where $X \sim B(n, 1/m)$, and

- the **power of a lineup** is estimated as x/n.

Case study 1 Uniformity trial of peanuts Part 1/8

data R



- Same peanut variety planted in Alabama field in two blocks of rectangular array of 18 rows by 6 plots.
- The yield (in pounds) is measured.
- Are the yields from the two blocks similar?

Block	Mean	Std. Dev
B1	2.458	0.392
B2	2.049	0.422

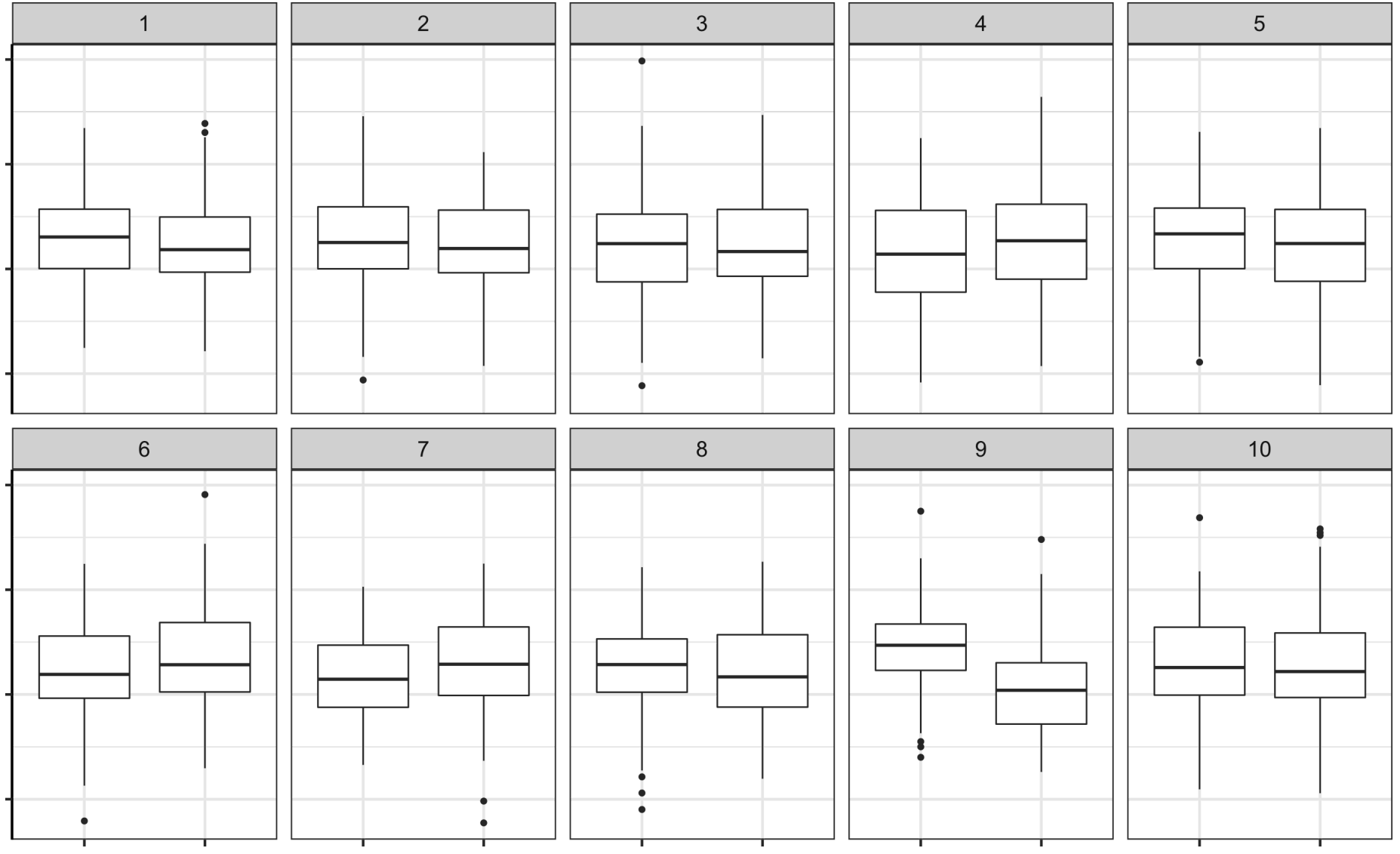
```
##
##     Welch Two Sample t-test
##
## data:  yield[block == "B1"] and yield[block == "B2"]
## t = 7.3979, df = 212.85, p-value = 3.149e-12
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.3004162 0.5186579
## sample estimates:
```

Case study 1 Uniformity trial of peanuts Part 2/8

1. $H_0 : \mu_1 = \mu_1$ vs. $H_1 : \mu_1 \neq \mu_2$ where μ_1 and μ_2 are yield means of block 1 and 2, respectively.
2. We choose our visual test statistic V_1 as a side-by-side boxplots of yield by block.
3. We generate the null data from $N(\bar{y}, s^2)$ where $\bar{y}$ and s are the sample mean and sample standard deviation of the yields.

Now we construct the lineup...

Case study 1 Uniformity trial of peanuts Part 3/8



Case study 1 Uniformity trial of peanuts Part 4/8

- So x out of n of you chose the data plot.
- So the visual inference p-value is $P(X \geq x)$ where $X \sim B(n, 1/10)$.
- In R, this is

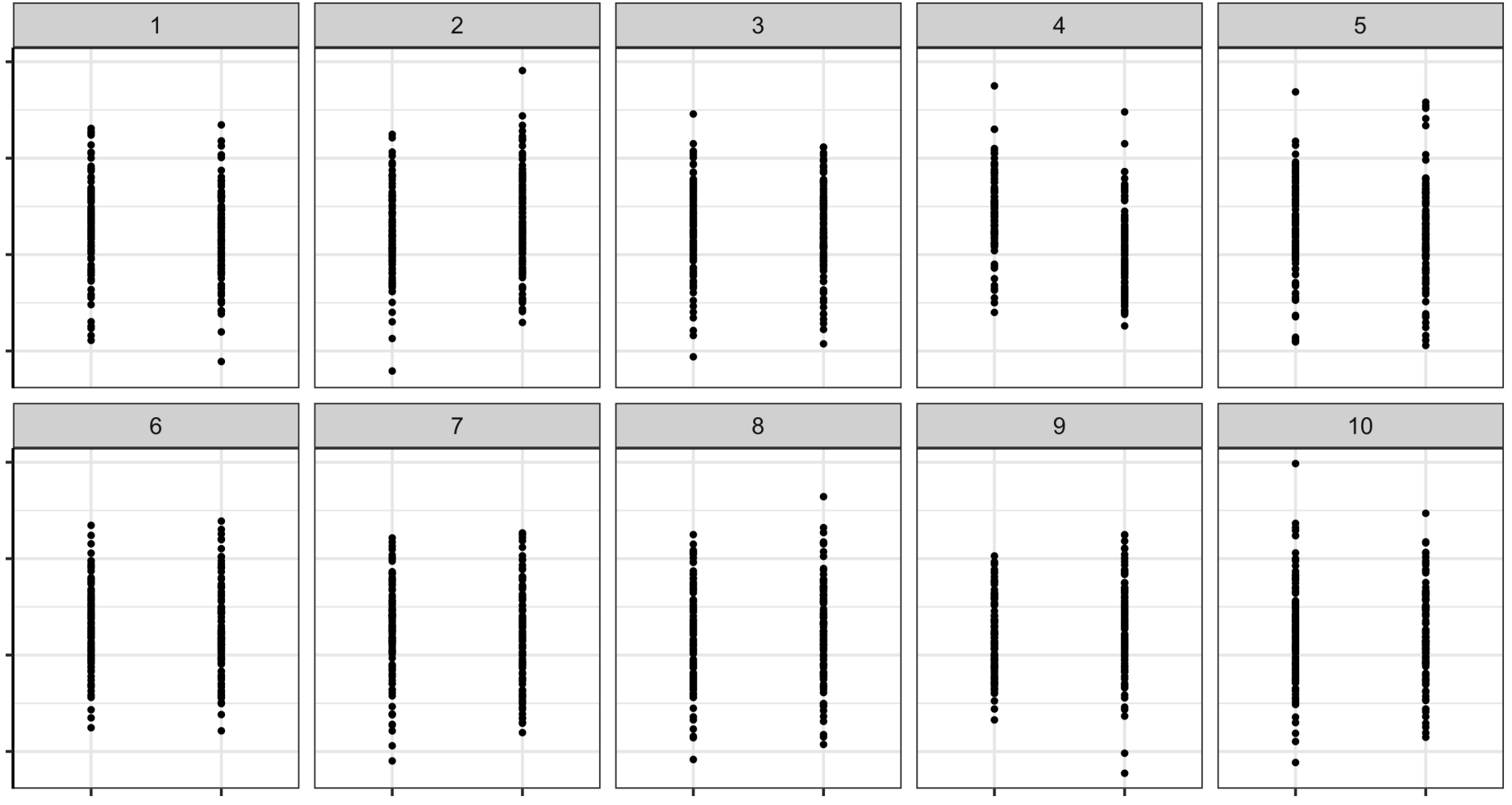
```
1 - pbinom(x - 1, n, 1/10)
# OR
nullabor::pvisual(x, n, 10)
```

- The power of the lineup is given as

```
x/n
# OR (need to use development one)
# install.packages("devtools")
# devtools::install_github("dicook/nullabor")
tibble(pic_id = 1, id = 1:n,
       response = 1, # dummy responses
       detected = rep(c(0, 1), c(x, n - x))) %>%
nullabor::visual_power(10)
```

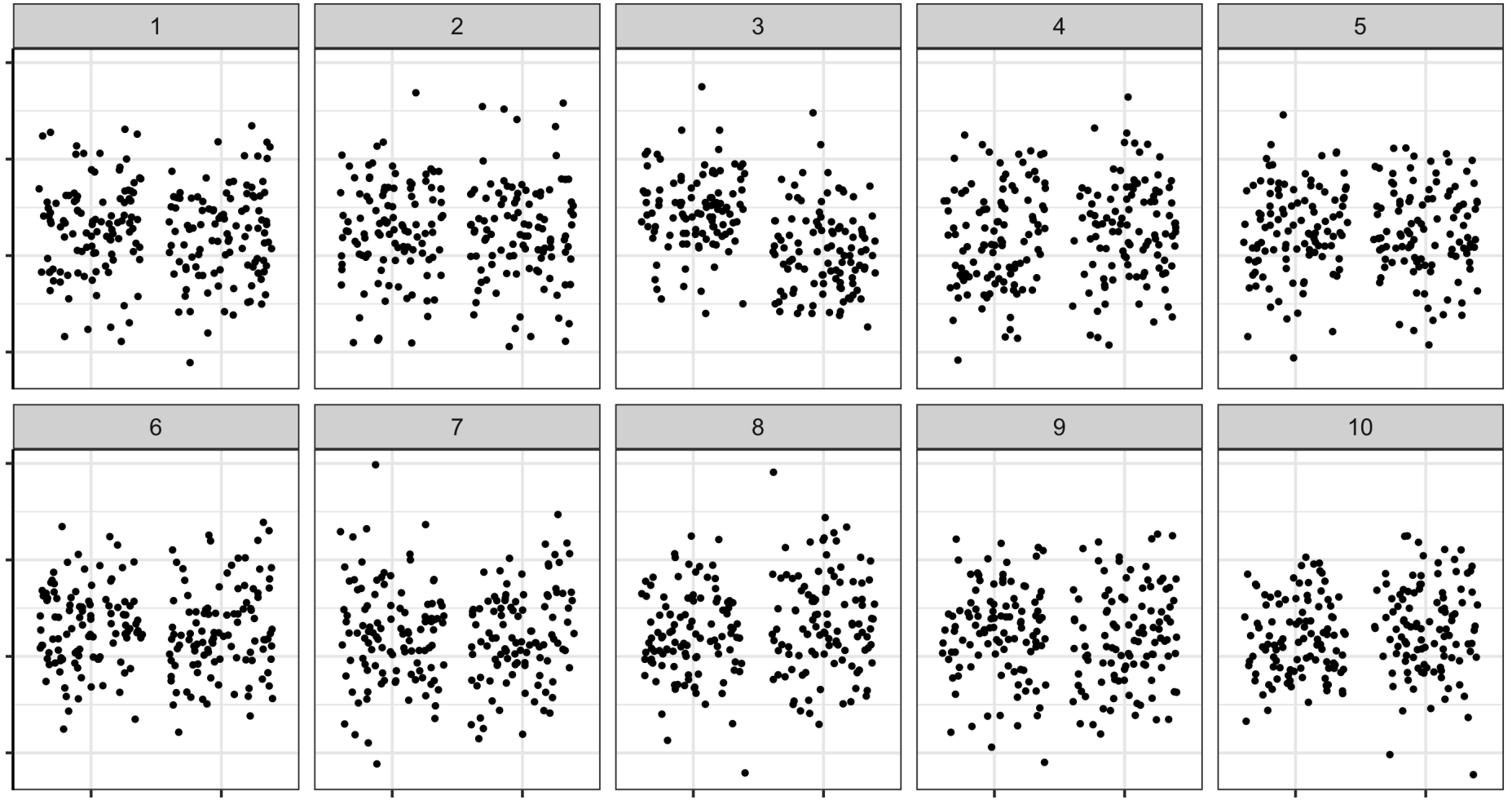
Case study 1 Uniformity trial of peanuts Part 5/8

What about if we change the visual test statistic?



Case study 1 Uniformity trial of peanuts Part 6/8

What about this one?



Case study 1 Uniformity trial of peanuts Part 7/8

- For the scatter plot x_2 out of n_2 of you chose the data plot.
- So the visual inference p-value is $P(X_2 \geq x_2)$ where $X_2 \sim B(n_2, 1/10)$.

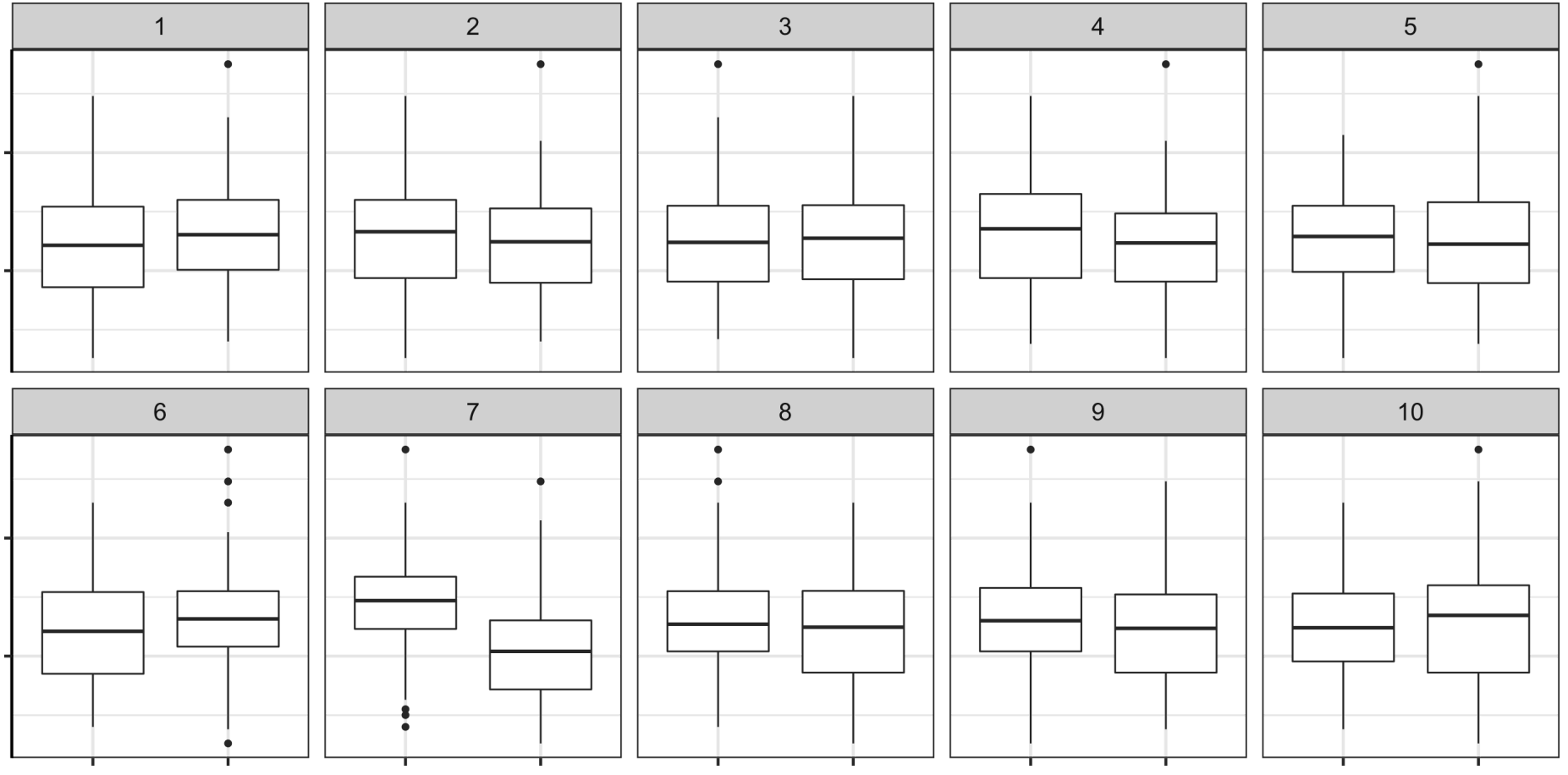
- For the jittered plot x_3 out of n_3 of you chose the data plot.
- So the visual inference p-value is $P(X_3 \geq x_3)$ where $X_3 \sim B(n_3, 1/10)$.

So the power difference is $100 \times \left| \frac{x_2}{n_2} - \frac{x_3}{n_3} \right| \%$.

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.

Case study 1 Uniformity trial of peanuts Part 8/8

What if we change the null generating method so that instead we permute the block labels?



Statistical significance and practical significance

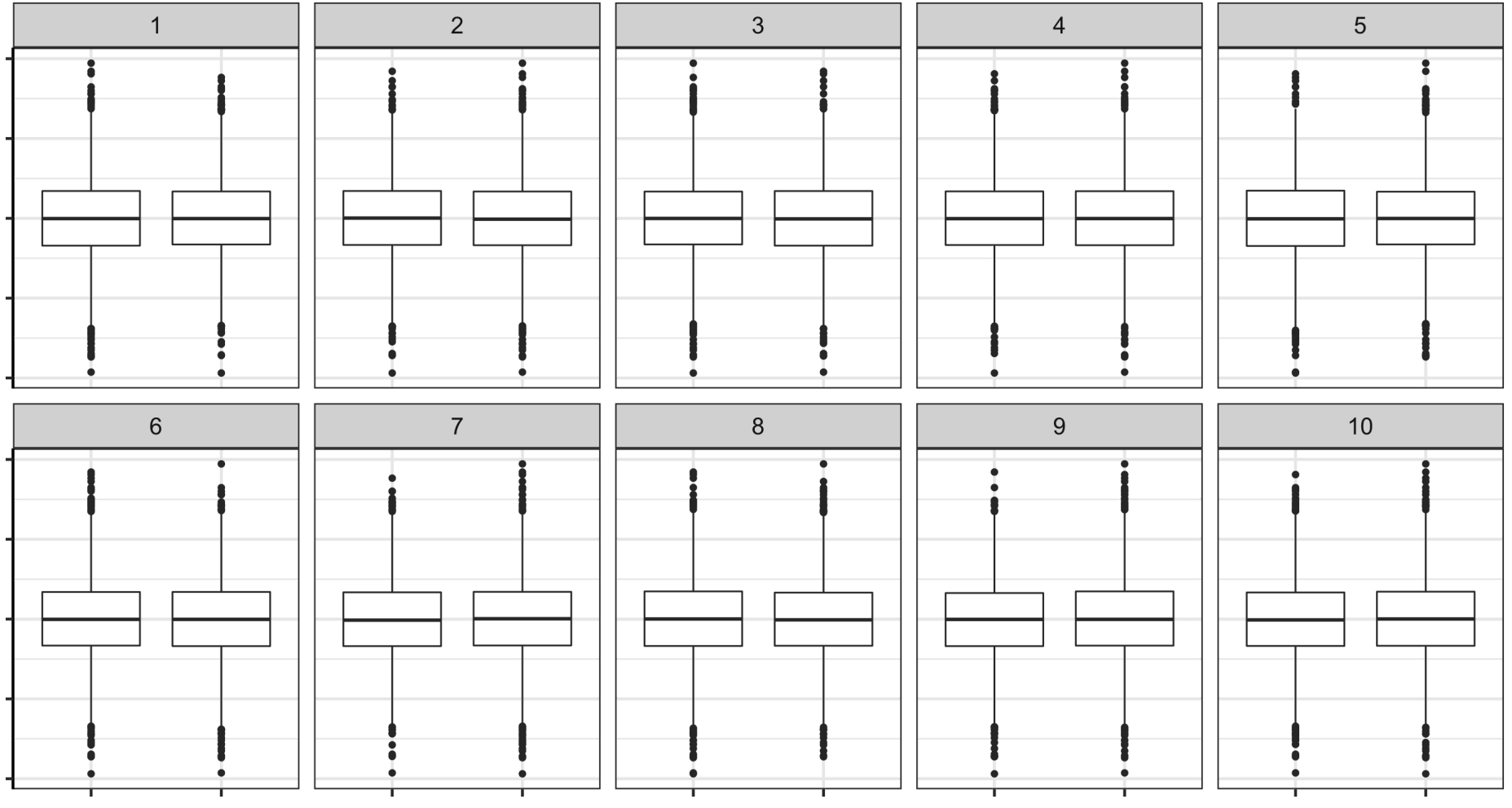
```r
set.seed(1)
sim <- tibble(id = 1:10000000) %>%
         mutate(y = c(rnorm(n()/2), rnorm(n()/2, 0.001)),
                group = rep(c("A", "B"), each = n()/2))
with(sim, mean(y[group=="A"]) - mean(y[group=="B"]))

## [1] -0.001443504

with(sim, t.test(y[group=="A"], y[group=="B"]))

##
##     Welch Two Sample t-test
##
## data:  y[group == "A"] and y[group == "B"]
## t = -2.2819, df = 1e+07, p-value = 0.0225
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.0026833804 -0.0002036271
## sample estimates:
##    mean of x    mean of y
## 0.0001819234 0.0016254271
```

- Notice here the real difference in the two groups is small (0.001) here.
- The two groups have a slightly different but the true difference is small, you might not care.
- The **practical significance** takes into account the effect size.

Statistical significance of the data plot

- Unlike conventional hypothesis testing, visual inference p-value depends on:
  - the visual test statistic V,
  - the individuals' visual perceptions,
  - the number of K observers,
  - the size m of the lineup, and
  - the effect size.
- The concept of conventional p-value is difficult for those that are not trained in statistics.
- The lineup is easier to understand to both novices and experts.

Lineup of small effect difference



For computational reasons, only 10,000 data points for each plot are used above.

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.
- If the data is for research purposes, then you may need ethics approval for publication.

That's it, for this lecture!

Lecturer: Emi Tanaka

Department of Econometrics and Business Statistics

ETC5521.Clayton-x@monash.edu