

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 1

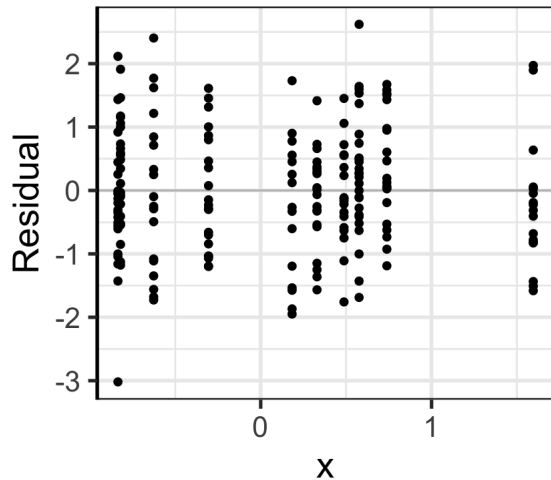


Today we are going to use surveys in visual inference as a way to think further about what can and cannot be inferred more generally given the data collection

Recap: Visual inference

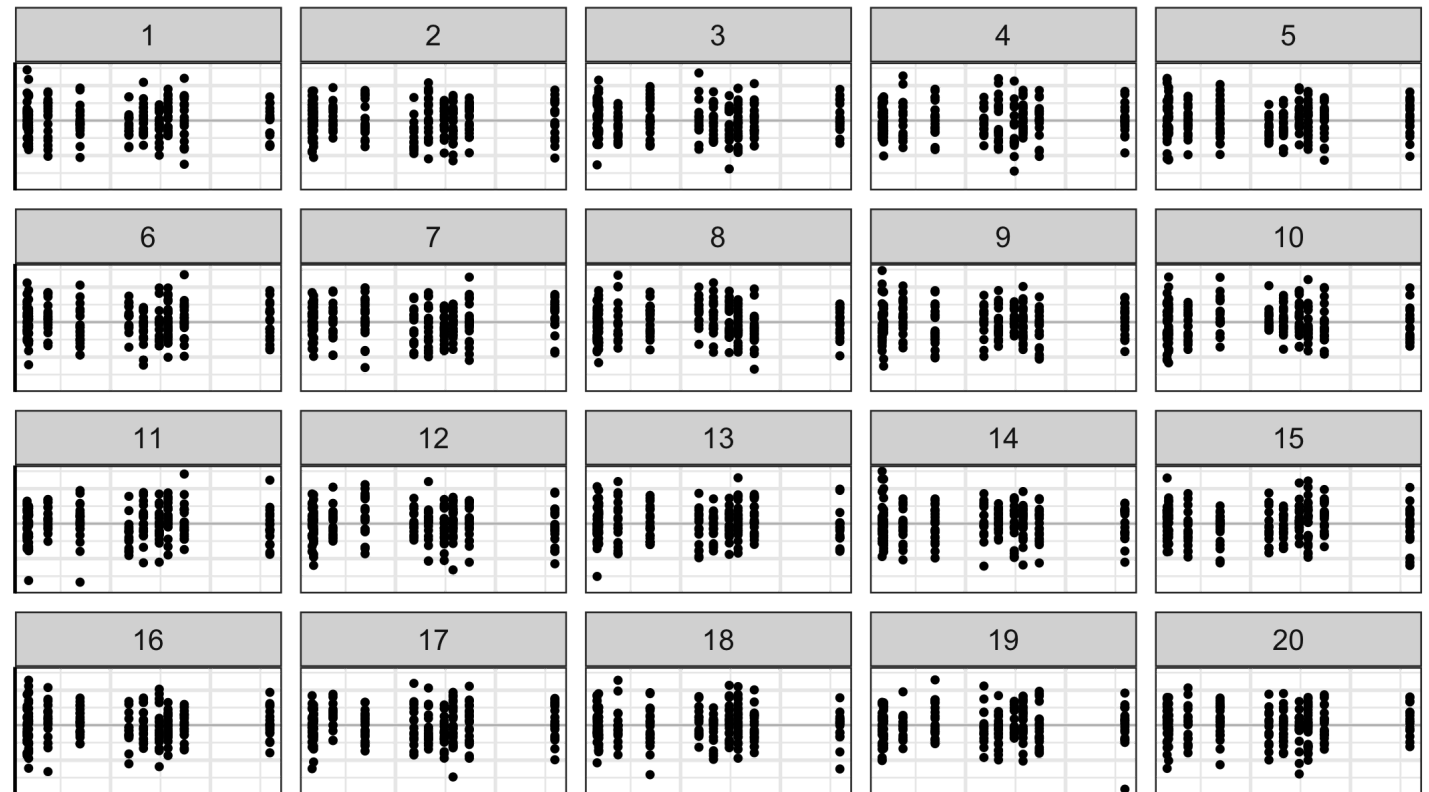
- Last week we studied how to use a hypothesis testing framework to assess a feature of a plot by treating a **plot as a test statistic**.

Data plot = Residual plot



Lineup

One plot is a data plot and the other are null plots



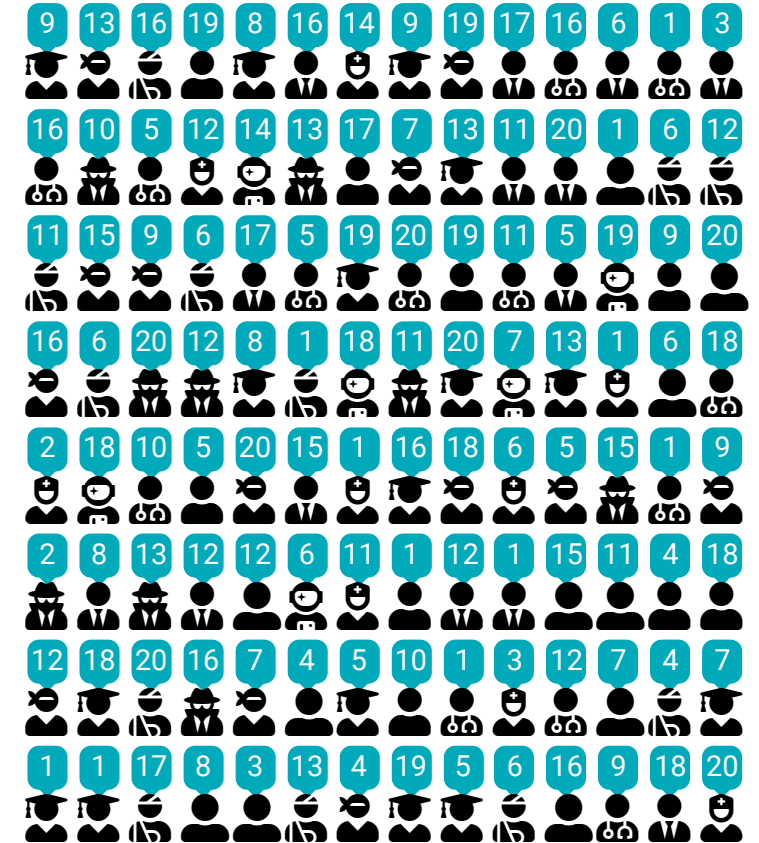
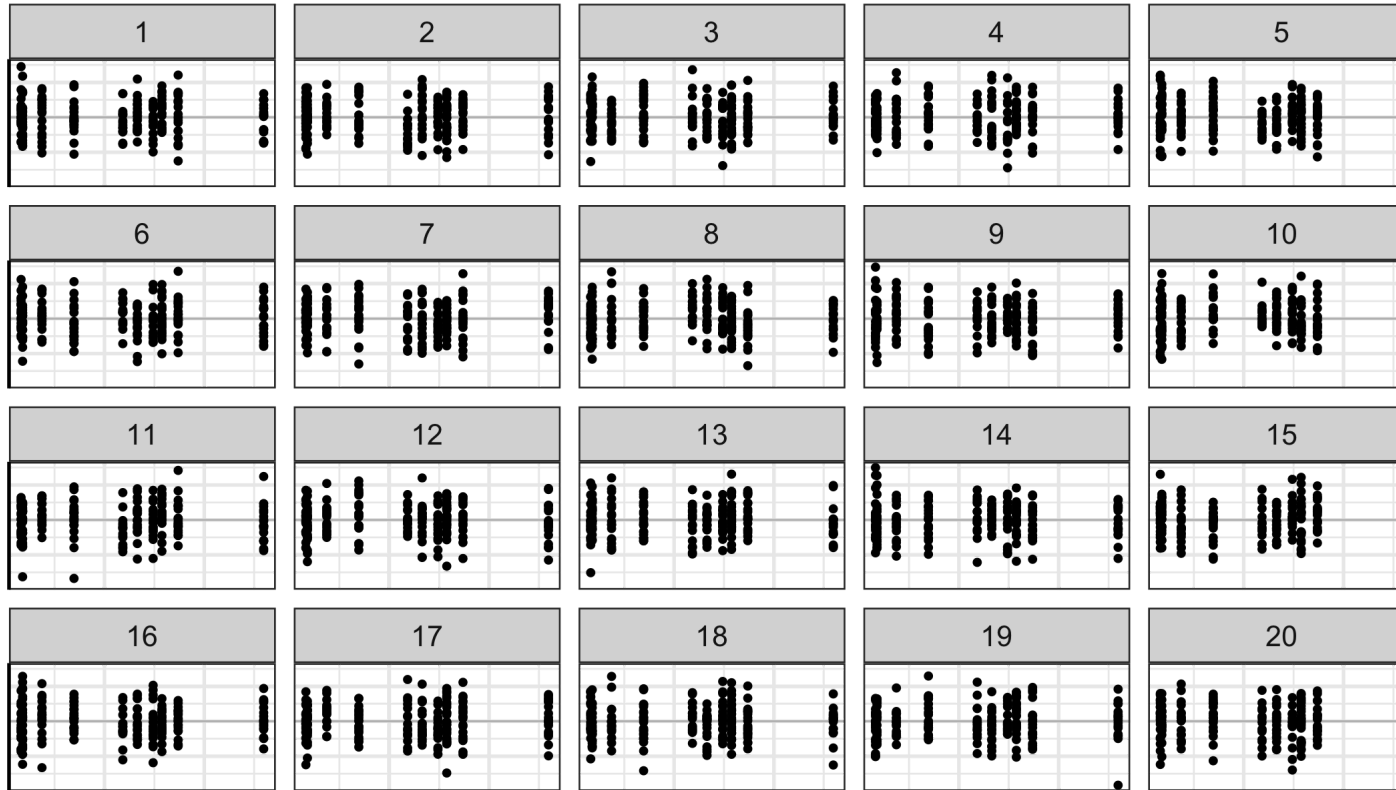
Visual inference test notations

We will use the following notation for the remaining lecture:

- There are n independent participants.
- We assume that each participant have the same detection probability (p), i.e. the probability of selecting the data plot in a lineup with m plots.
- Let X be the number of participants who detect the data plot out of n participants.
- We denote x to be the observed value of X .
- For visual inference test, we test the hypothesis: $H_0 : p = 1/m$ vs. $H_1 : p > 1/m$.
- Under H_0 , $X \sim B(n, 1/m)$.
- Therefore, the visual inference p-value is $P(X \geq x) = 1 - P(X \leq x - 1) = \sum_{k=x}^n \binom{n}{k} \frac{(m-1)^k}{m^n}$.
- Recall `pbinom(x, n, p)` is $P(X \leq x)$.

Lineup

One plot is a data plot and the other are null plots



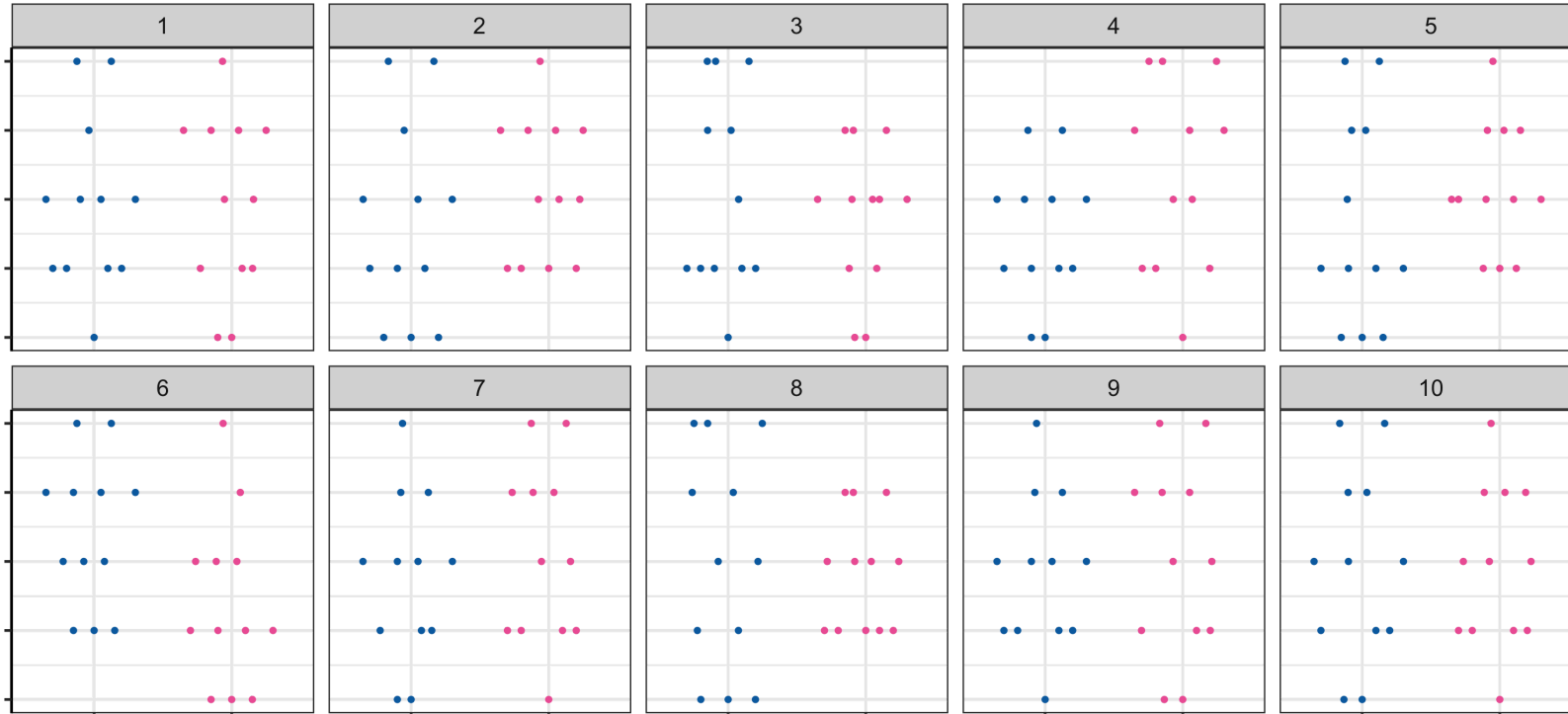
Choices	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Total
Frequency	11	2	3	4	7	8	5	4	6	3	6	8	6	2	4	8	4	7	6	8	112

The data plot is Plot 13 and visual inference p-value is $P(X \geq 6) = 0.491$ where $X \sim B(112, 0.05)$.

Data collection

Let's use last week's survey results to discuss about data collection issues

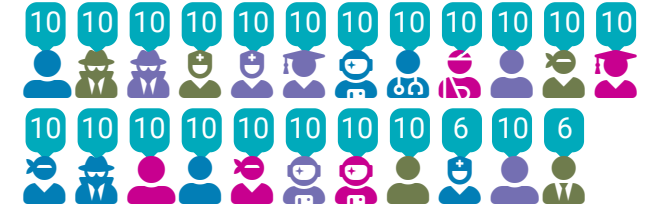
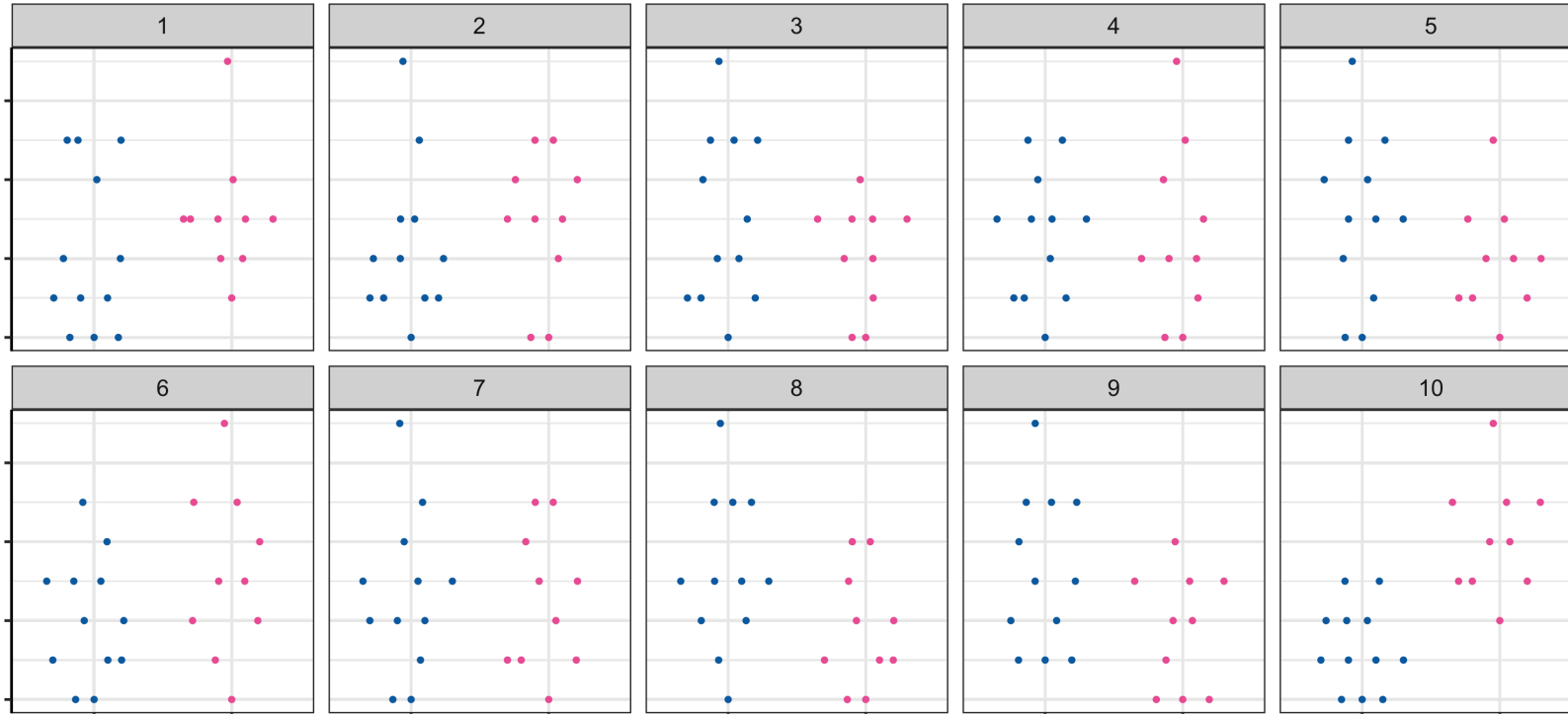
Lineup 1 In which plot is the pink group higher than the blue group?



- A person is indexed by the combination of the icon and the color
- The above colored icons are ordered by the time of response with those that answered it earlier appearing first

Choices	1	2	3	4	5	6	7	8	9	10	Total
Frequency	0	2	0	19	3	4	0	1	4	1	34

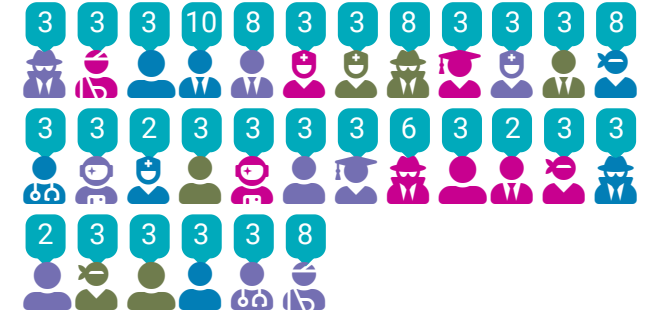
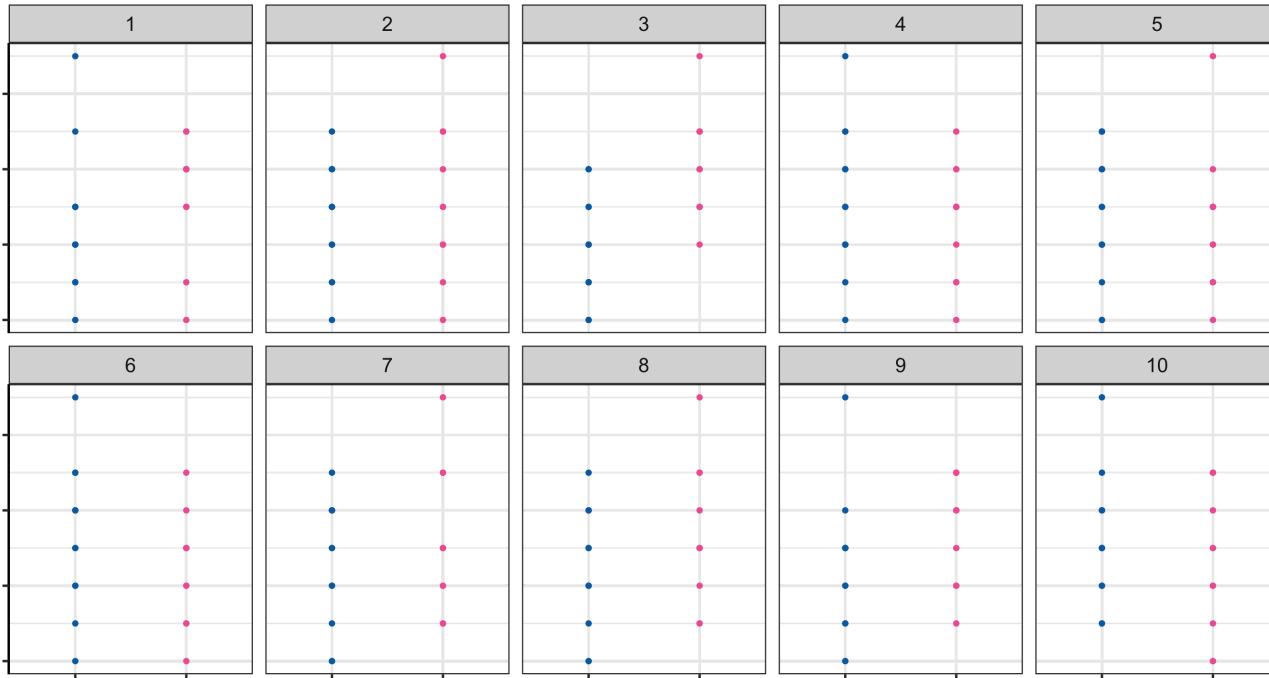
Lineup 2 In which plot is the pink group higher than the blue group?



Choices	1	2	3	4	5	6	7	8	9	10	Total
Frequency	0	0	0	0	0	2	0	0	0	21	23

The data plot is Plot 10 and visual inference p-value is $P(X \geq 21) = 0$ where $X \sim B(23, 0.1)$.

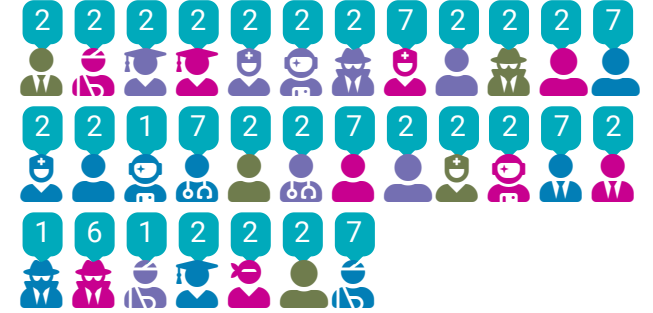
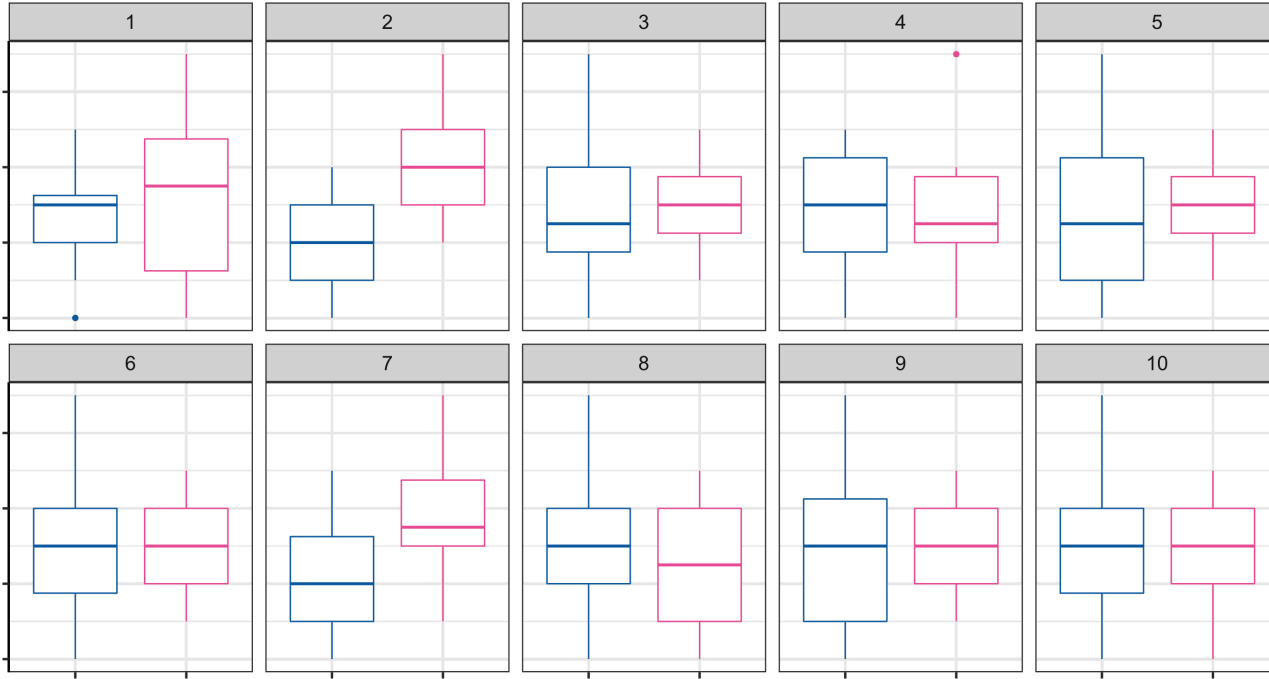
Lineup 3 In which plot is the pink group higher than the blue group?



Choices	1	2	3	4	5	6	7	8	9	10	Total
Frequency	0	3	21	0	0	1	0	4	0	1	30

The data plot is Plot 3 and visual inference p-value is $P(X \geq 21) = 0$ where $X \sim B(30, 0.1)$.

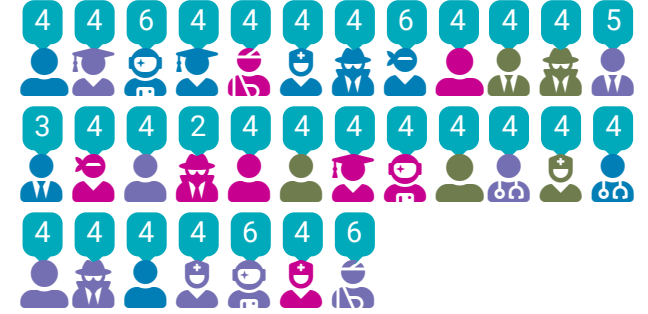
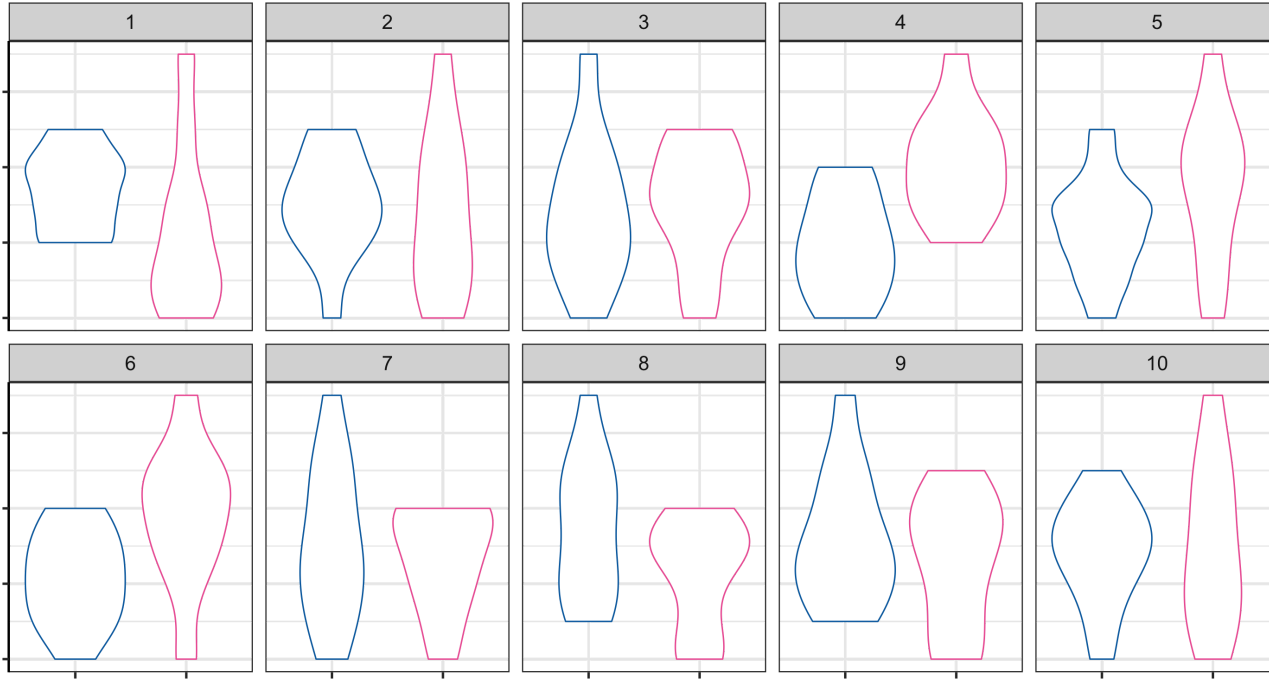
Lineup 4 In which plot is the pink group higher than the blue group?



Choices	1	2	3	4	5	6	7	8	9	10	Total
Frequency	3	21	0	0	0	1	6	0	0	0	31

The data plot is Plot 2 and visual inference p-value is $P(X \geq 21) = 0$ where $X \sim B(31, 0.1)$.

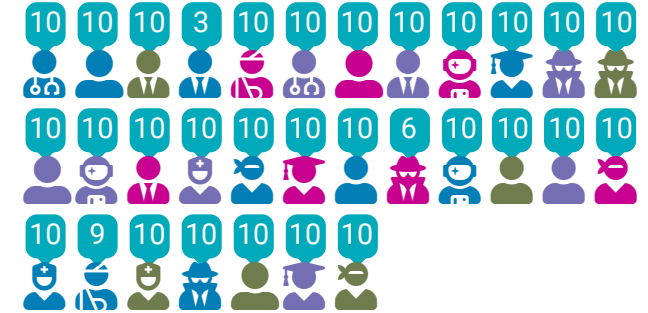
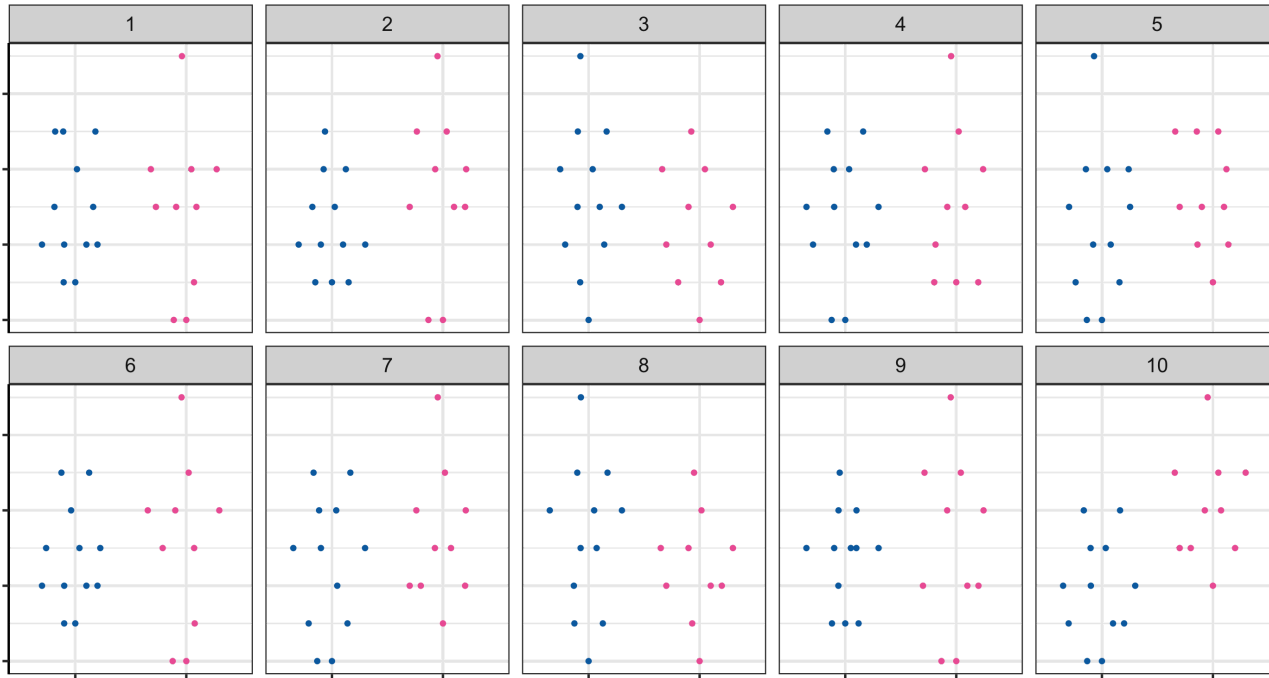
Lineup 5 In which plot is the pink group higher than the blue group?



Choices	1	2	3	4	5	6	7	8	9	10	Total
Frequency	0	1	1	24	1	4	0	0	0	0	31

The data plot is Plot 4 and visual inference p-value is $P(X \geq 24) = 0$ where $X \sim B(31, 0.1)$.



















































Lineup 6 In which plot is the pink group higher than the blue group?




Choices	1	2	3	4	5	6	7	8	9	10	Total
Frequency	0	0	1	0	0	1	0	0	1	28	31

The data plot is Plot 10 and visual inference p-value is $P(X \geq 28) = 0$ where $X \sim B(31, 0.1)$.

Do you notice anything from the results?

Poll	Choice
1	2  
	4                
	5   
	6    
	8 
	9    
	10 
2	6  
	10                 

- Let's have a closer look at . This person chose Plot 6 for 4 lineups out of 5 answered.
- Is this likely to happen by chance? What do you think about the data quality?
- Are there any other results you would question about?

02:00

Pre-emptive data quality checks in mass-scale visual inference surveys

The survey may be designed so it:

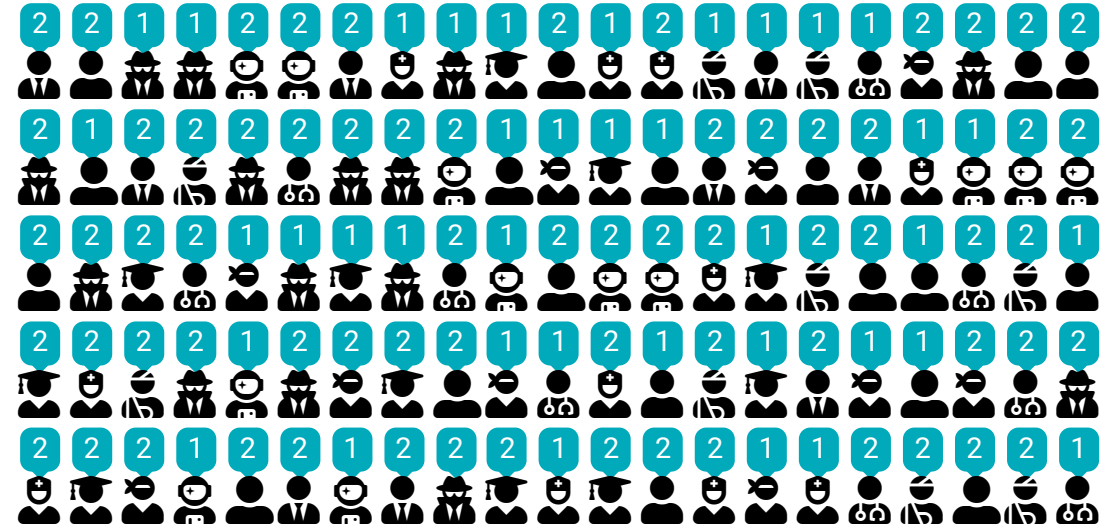
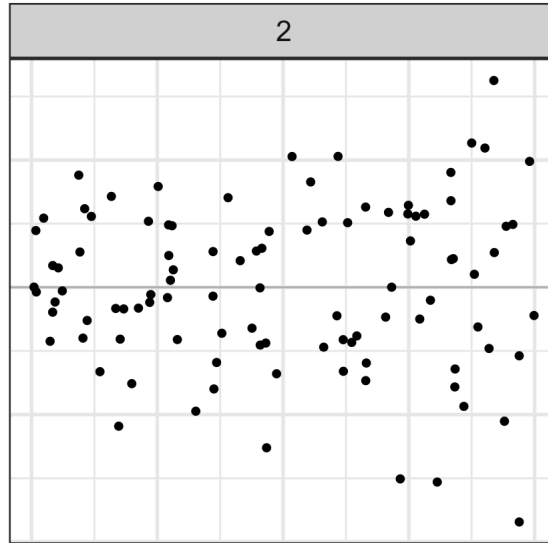
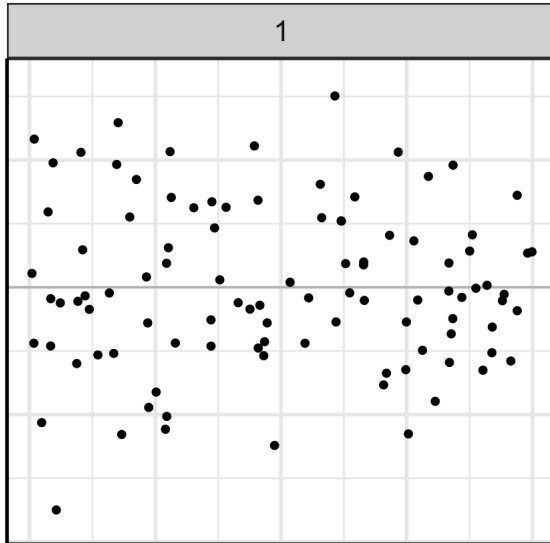
1. records the demographic of participants (e.g. gender, age and education),
2. records the choices of multiple lineups from each participant,
3. records the reaction time for selecting their choice, and
4. includes some lineups with an "obvious" data plot.

Post data collection then you can check:

- if the survey was representative of the population by checking the demographic information,
- if participant fails to detect the data plot in the "obvious" lineups, it means that they may not be answering sincerely or they did not understand the instructions, which means you may want to remove their data,
- if participant appears to be selecting too quickly, they may not actually be processing the plots appropriately.

**How do choices in the lineup
design effect the inference?**

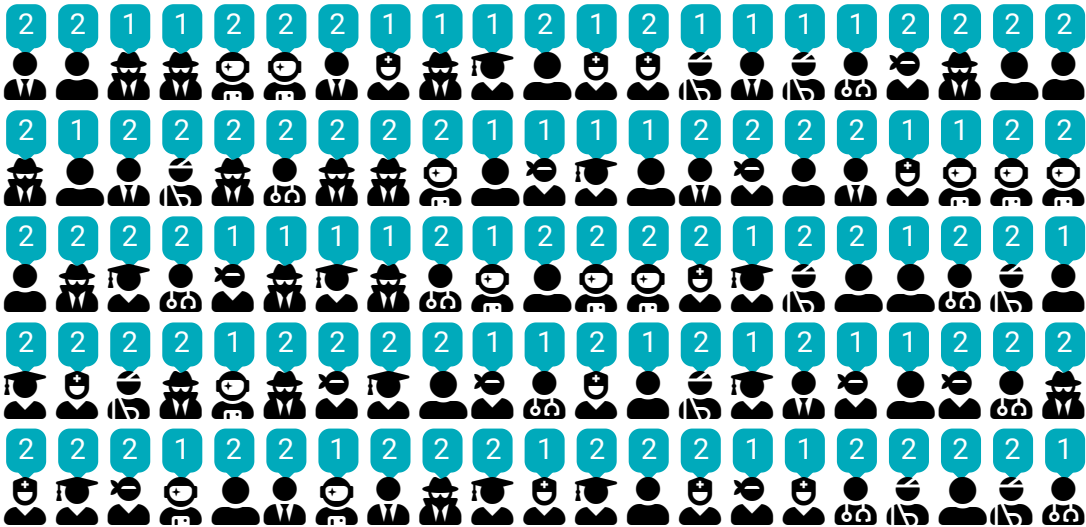
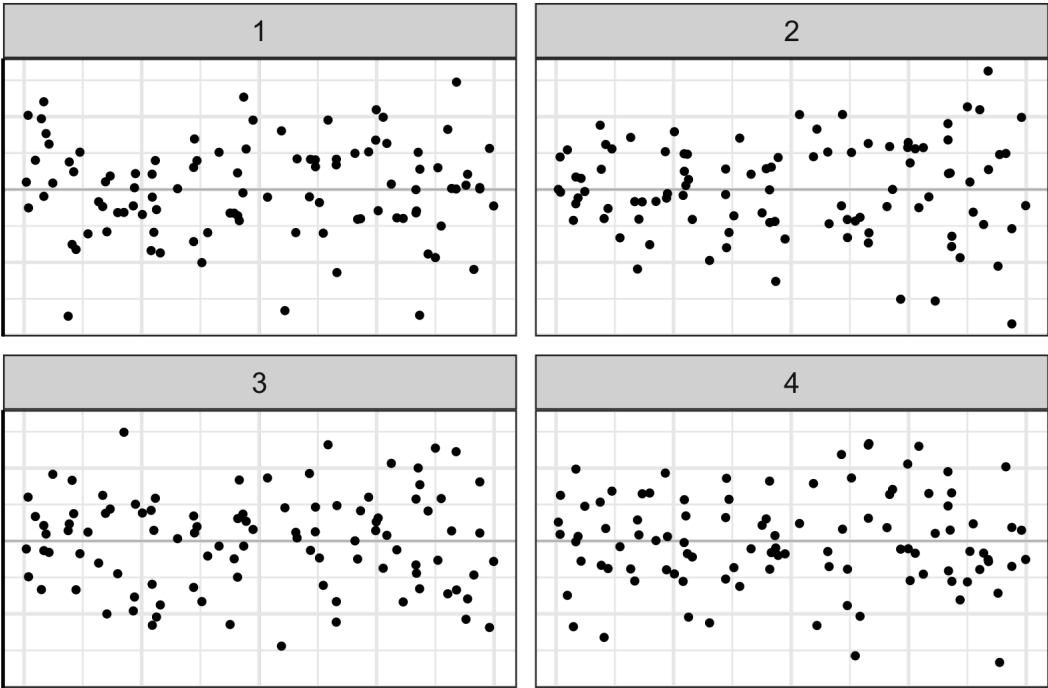
Lineup with $m = 2$ plots



Choices	1	2	Total
Frequency	38	67	105

The data plot is Plot 2 and visual inference p-value is $P(X \geq 67) = 0.003$ where $X \sim B(105, 0.5)$.

Lineup with $m = 4$ plots



Choices	1	2	3	4	Total
Frequency	38	67	0	0	105

The data plot is Plot 2 and visual inference p-value is $P(X \geq 67) = 0$ where $X \sim B(105, 0.25)$.

Lineup with $m = 4$ plots

- In fact in visual inference if x people detect out of n people from the same lineup, the frequency distribution for selection of null plots does *not* change the visual inference p -value nor the power of the lineup
- So all the outcomes below yield the *same* visual inference p -values and power estimate

Choices	1	2	3	4	Total
Experment 1	52	53	0	0	105
Experment 2	20	53	20	12	105
Experment 3	4	53	36	12	105

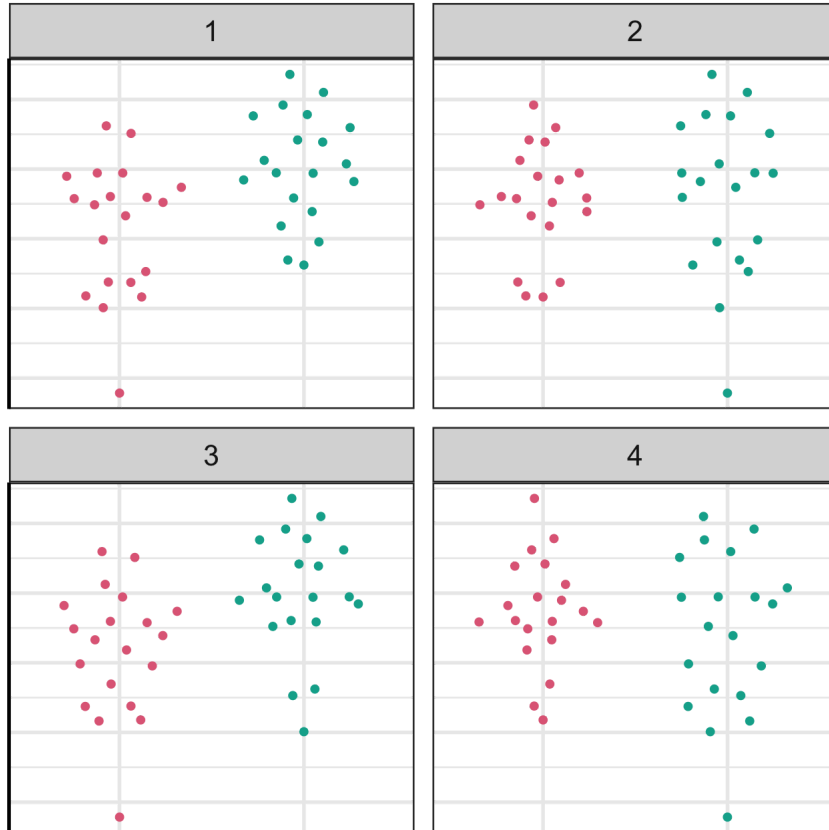
- Is that an issue?

When null data has features under alternative hypothesis



data

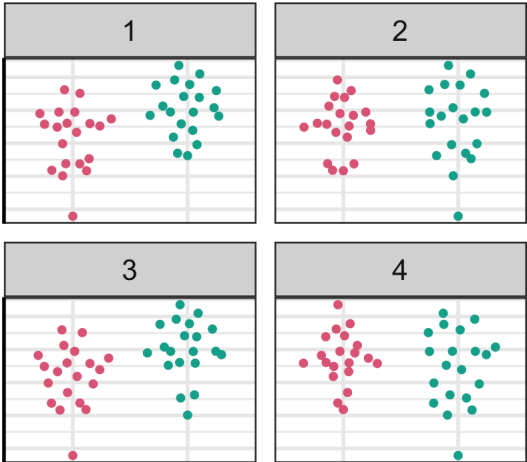
R



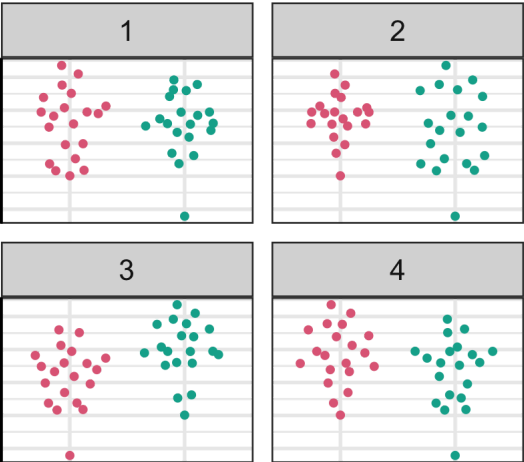
- Occasionally, null data demonstrates features that are more aligned with data generated from the alternative hypothesis
- For example, the data plot is in Plot 3 for the lineup on the left but Plot 1 demonstrates a bigger difference in the mean (and smaller standard deviation) of the two groups
- There is indeed a significant mean difference between the two groups for the data in Plot 3, but this can be overshadowed by null data in Plot 1
- This case is extremely rare; in fact, I cheated by generating 100 lineups of the same dimension and took the extreme case.
- While this case is rare, it can happen so we need to be careful in generalising the results based on a test on one lineup

Multiple lineups for the same data plot

Lineup 1



Lineup 2



Lineup 3



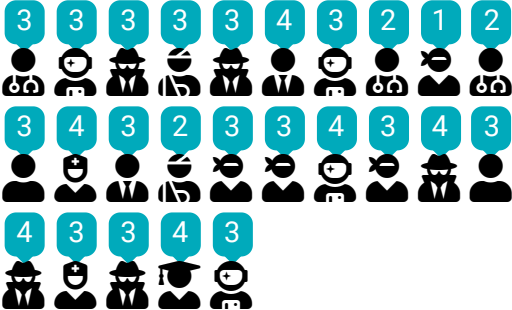
Lineup 4



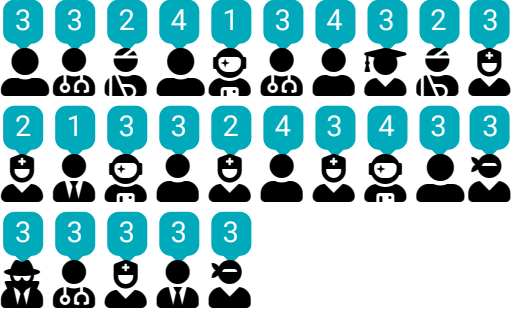
Lineup 1



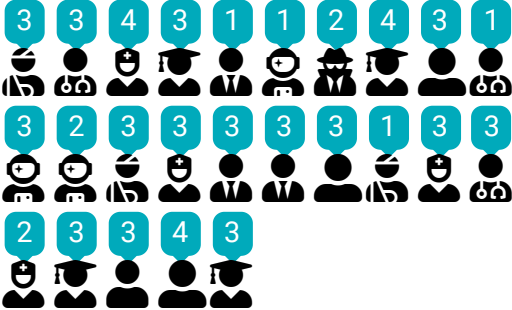
Lineup 2



Lineup 3



Lineup 4



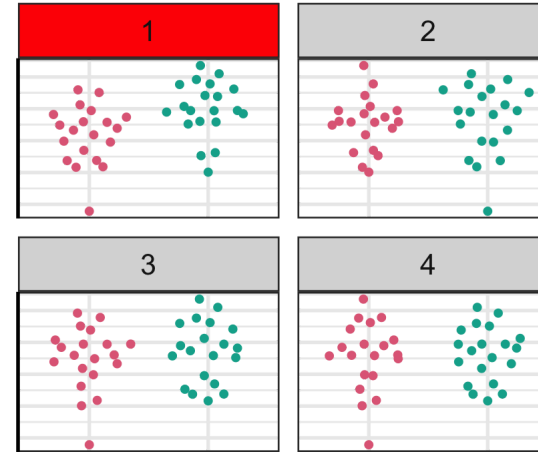
Lineup 1

Choices	1	2	3	4	Total
---------	---	---	---	---	-------

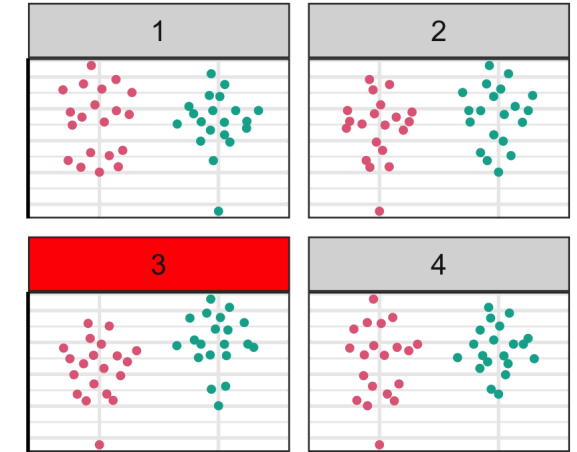
What if the positions of the plots has an effect on detection?

- If the position of the plots has an effect on detection, we should take this into account as a potential source of variation and consider using position as a "block" in randomisation
- In another words, we should randomise the position of the data plot in the lineup such that each position appears roughly equally number of times for the data plot
- For example, in the lineups on the right, the data plot is highlighted with the red color and you can see each position appears exactly once

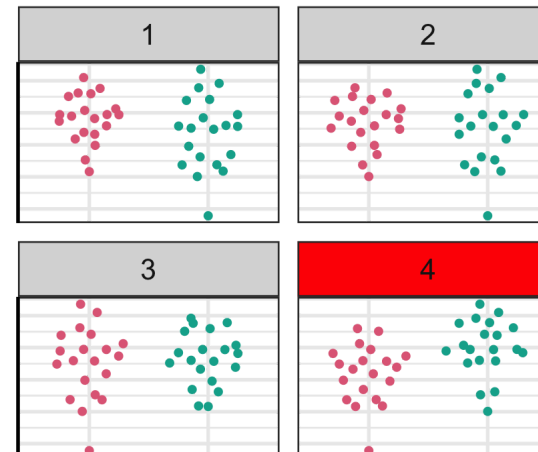
Lineup 1



Lineup 2



Lineup 3



Lineup 4



Statistical significance and practical significance

```
library(nulllabor)
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"]))

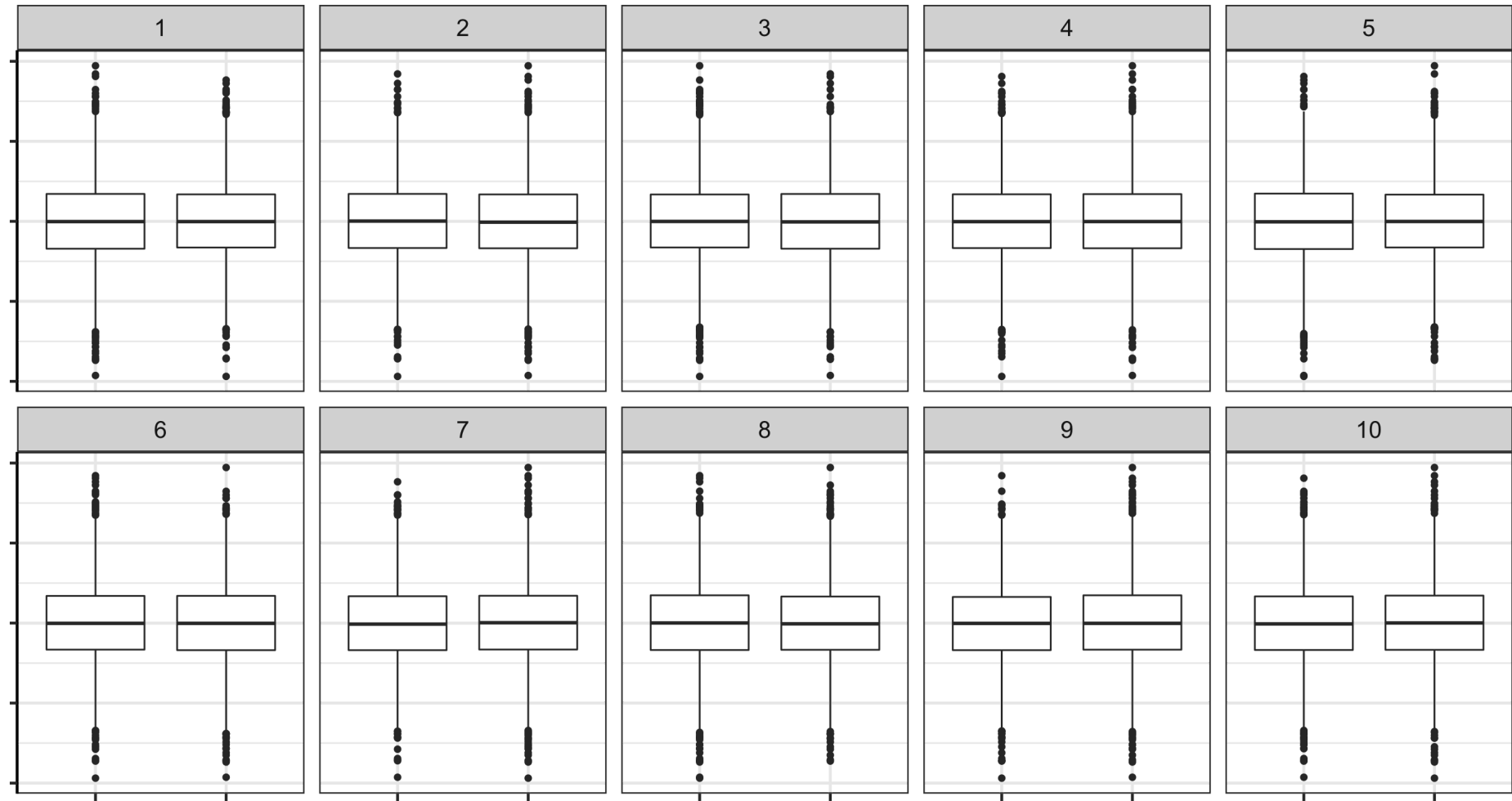
## [1] -0.001443504

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
## 95 percent confidence interval:
##  -0.0026833804 -0.0002036271
## sample estimates:
##      mean of x      mean of y
```

- 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.

Lineup of small effect difference



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

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 null generation method,
 - the number of n observers,
 - the size m of the lineup, and
 - the effect size (or detection probability p).
- 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.
- If you reject the null hypothesis, you can find out *why* that might be the case



This work is licensed under a [Creative Commons Attribution-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-sa/4.0/).

Lecturer: *Emi Tanaka*

✉ ETC5521.Clayton-x@monash.edu

📅 Week 12 - Session 1