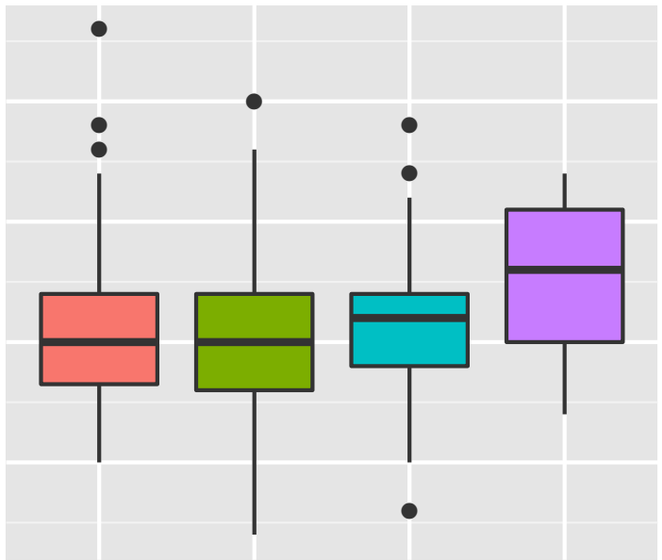
Statistics on Street Corners

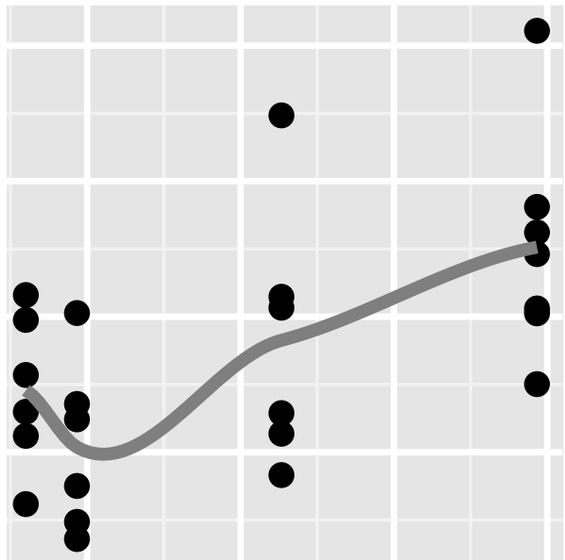
Using Crowd-Sourcing to Conduct Statistical Inference with Exploratory Data Analysis

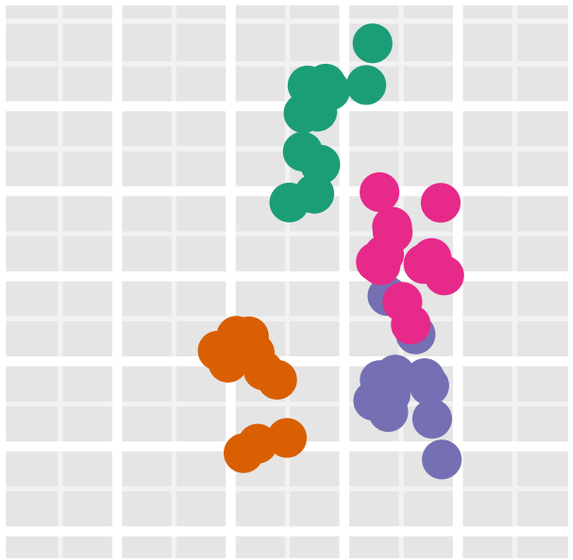
Di Cook
Econometrics and Business Statistics,
Monash University

Joint work with Heike Hofmann, Mahbub Majumder, Niladri Roy Chowdhury, Hadley Wickham, Andreas Buja, Debby Swayne, Eun-kyung Lee, Yifan Zhao, Tengfei Yin, Lendie Follett, Eric Hare, Adam Loy, Susan Vanderplas, Nathaniel Tomasetti

NZSA/ORSNZ AUCKLAND NOV 2016







What do you see?

Purple group has generally bigger values than the others

Relationship between the two variables is positive

The four groups are generally different, green and orange are separated from the other two

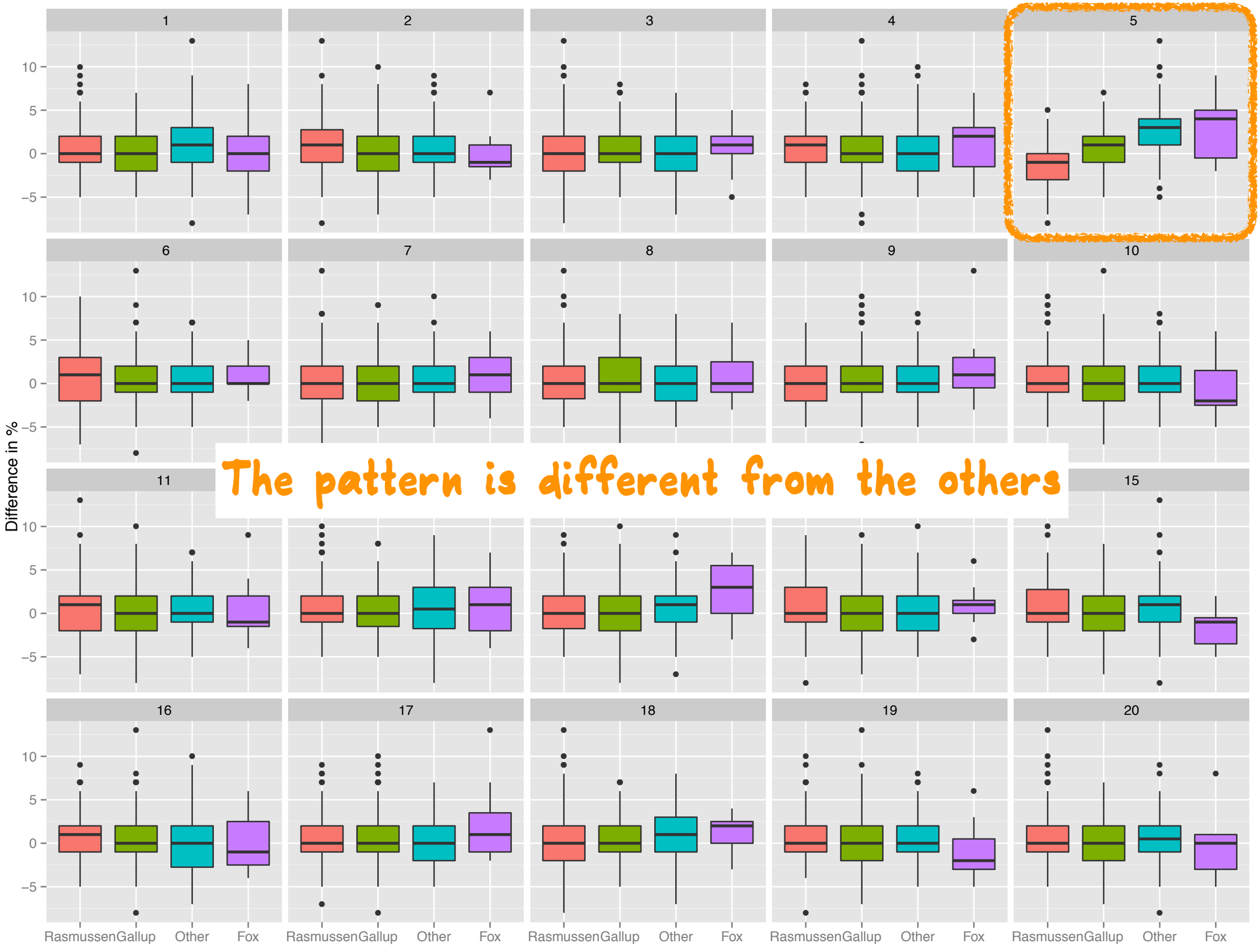
Statistical graphics

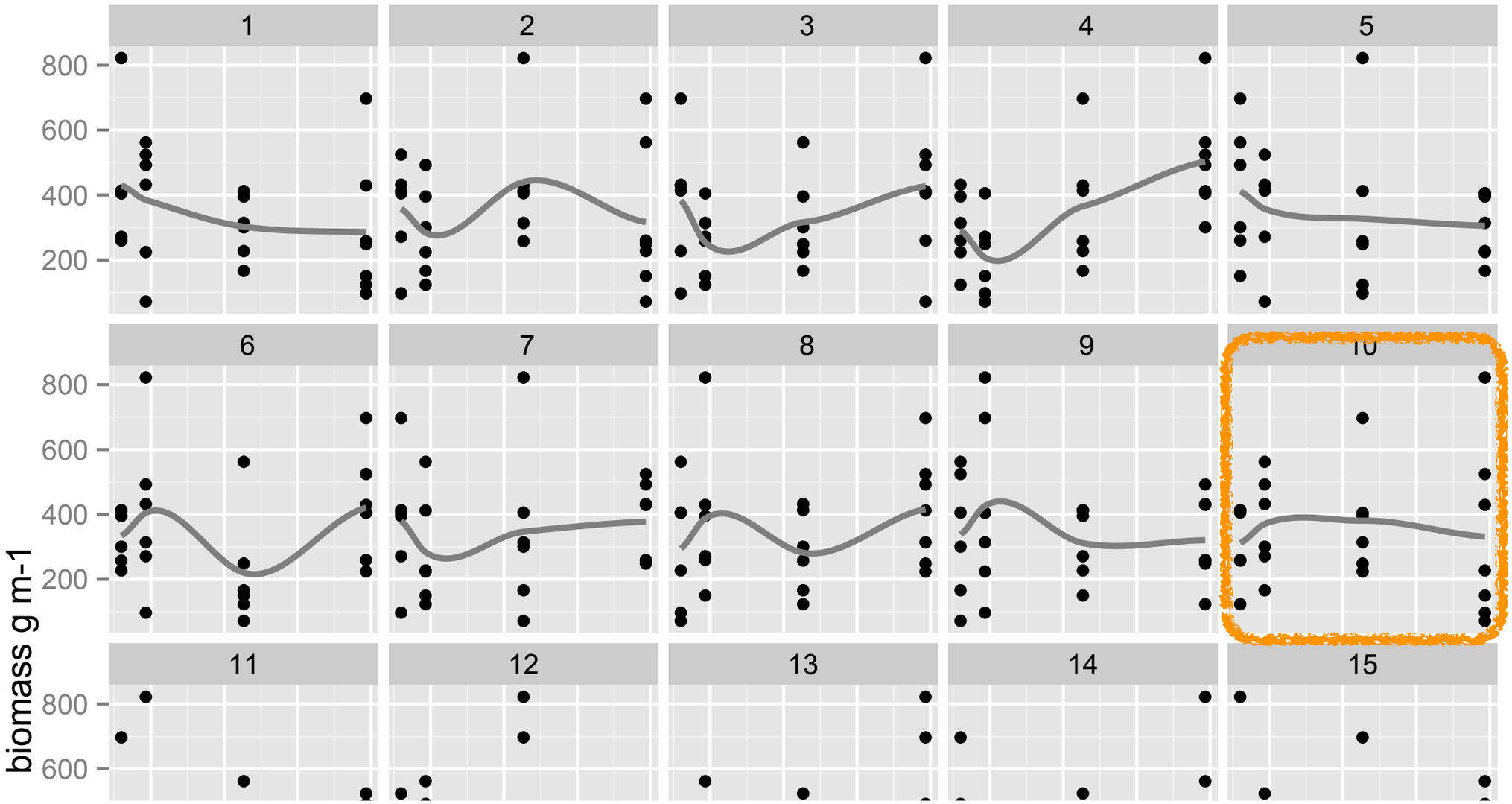
- Plots allow us to see patterns that numerical statistics cannot show.
- Plots encourage curiosity and discovery when most of statistics is about being skeptical.
- We use plots, e.g. for model diagnosis, but we have no rigorous quantification of making a decision about the model fit.
- This new work provides a way to evaluate whether what we see in plots is statistically significant, bridging curiosity with skepticism.

Very useful for data science

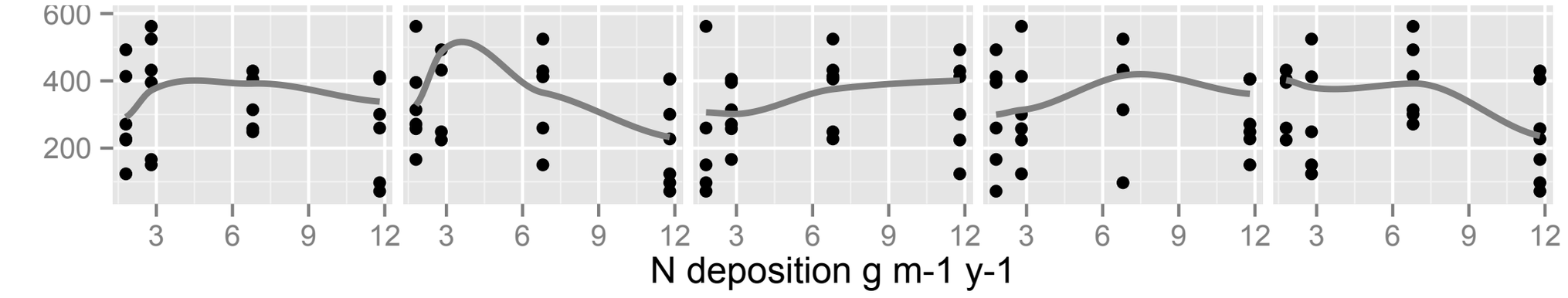
Idea

Instead of looking at the data plot in isolation, place it in the context of plots of data where there is nothing happening.

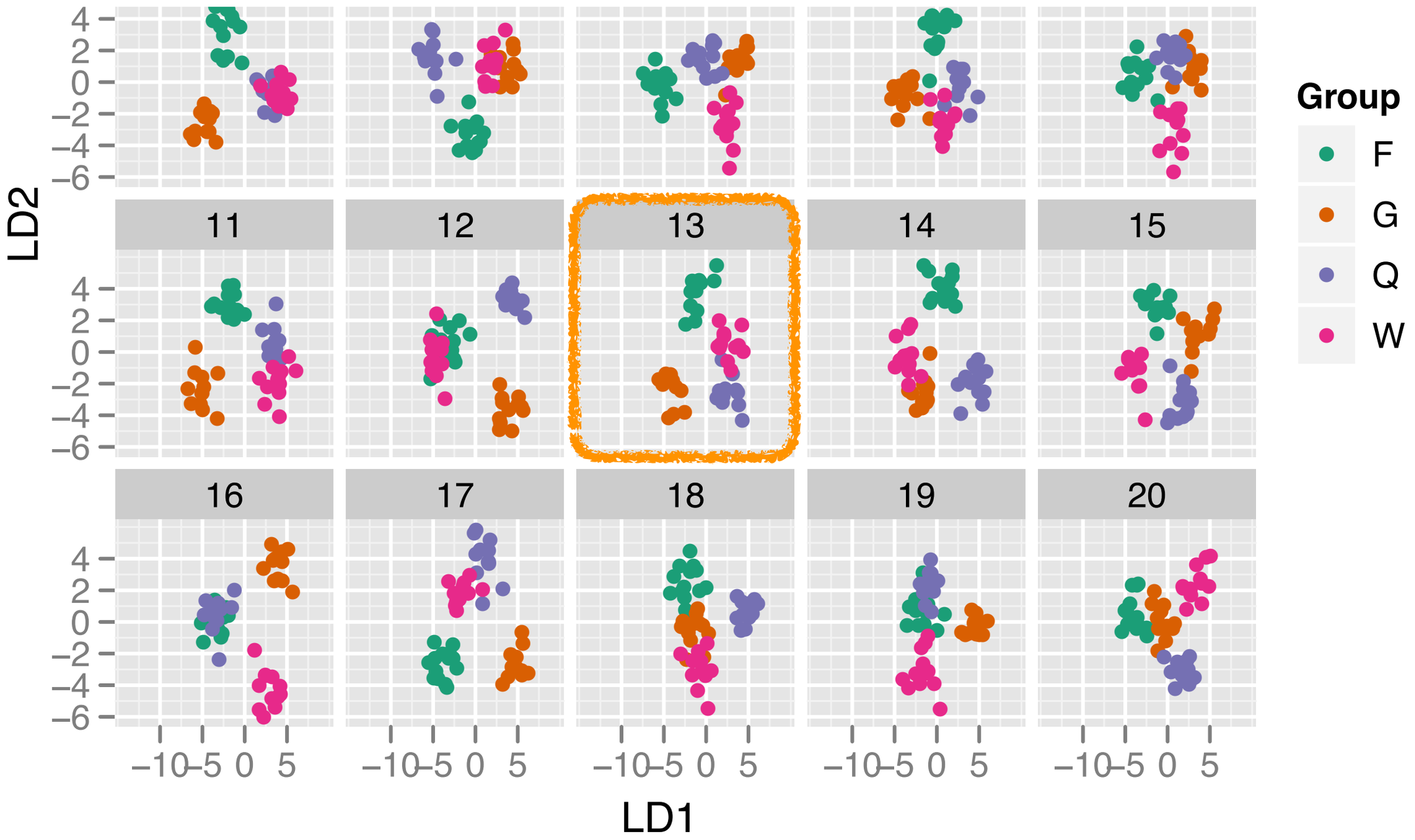

1
2
3
4
5
6
7
8
9
10
11
15
16
17
18
19
20
10
5
0
−5
Difference in %
The pattern is different from the others
RasmussenGallup
Other
Fox



The pattern is not any more different than in other plots. The plot I showed earlier is actually a NULL plot, where I know the pattern is due to randomness.



The groups are no more separated than in other plots. The separation in the data is purely by chance, actually because it is a projection from EMPTY high-dimensional space, there are many ways to find gaps.

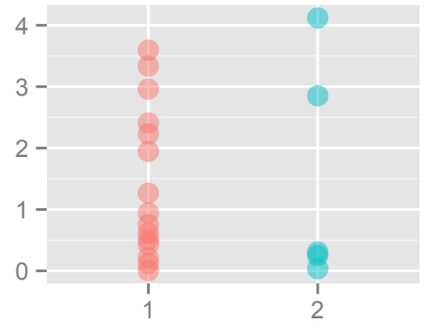


Outline

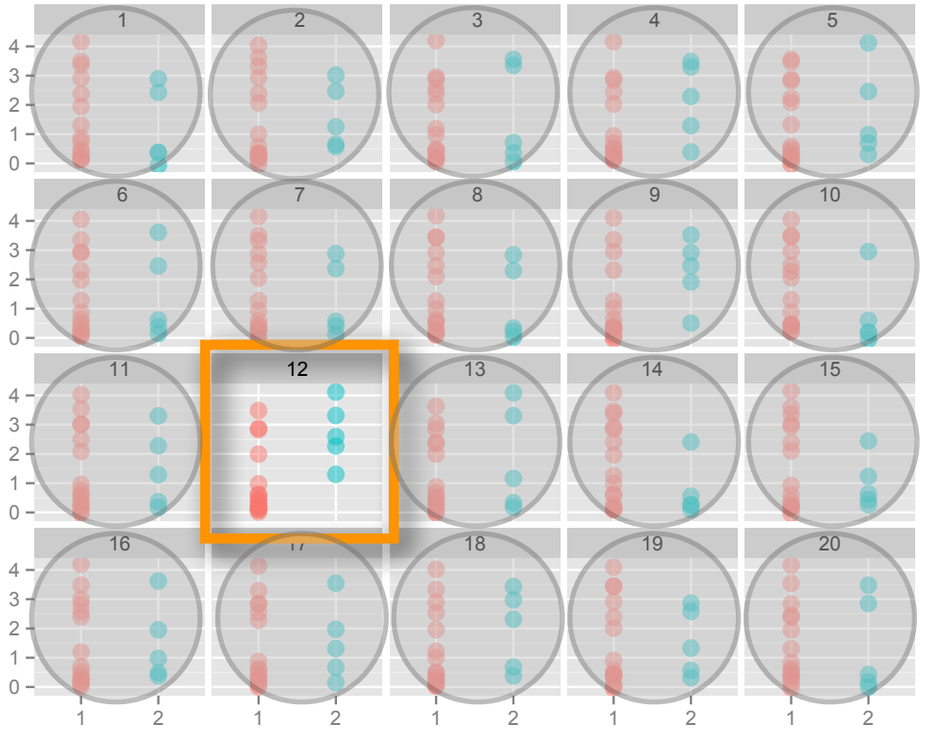
- Review of hypothesis testing, protocols, concepts
- Validation study, computing p-values and power
- Follow-up: Developing metrics, eye-tracking
- The value of the human significance machine: learning, age, gender, education

Protocols

- Rorschach: Show many pictures of data with "nothing" happening, pictures from a null distribution



- Lineup: Embed the plot of the data among plots of data generated from the null distribution

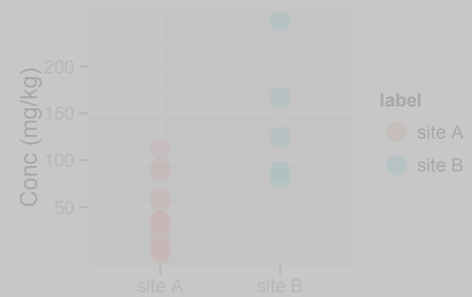
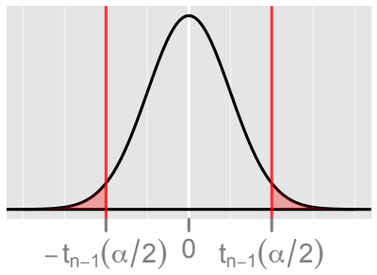
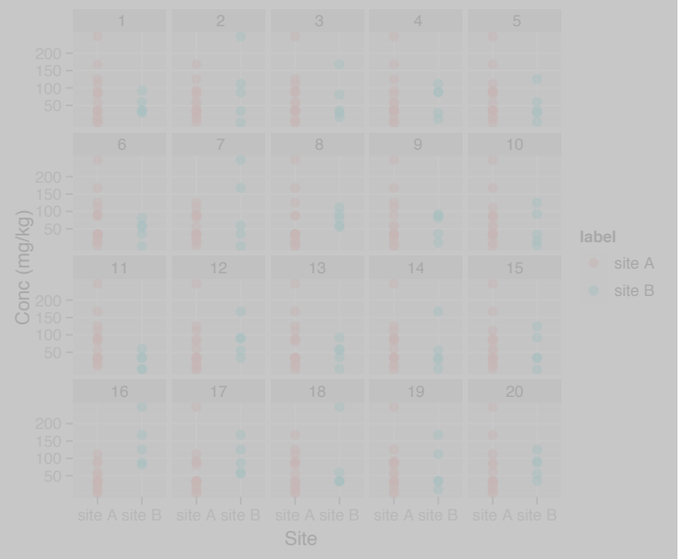


Data plot

Null plots

Source: Buja et al (2009) RSPT(A)

Hypothesis testing

	Mathematical Inference	Visual Inference
Hypothesis	$H_0 : \mu_1 = \mu_2$ vs $H_a : \mu_1 \neq \mu_2$	$H_0 : \mu_1 = \mu_2$ vs $H_a : \mu_1 \neq \mu_2$
	↓	↓
Test Statistic	$T(y) = \frac{\bar{y}_1 - \bar{y}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	$T(y) =$ 
	↓	↓
Sampling Distribution	$f_{T(y)}(t)$; 	$f_{T(y)}(t)$; 
	↓	↓
Reject H_0 if	observed T is extreme	observed plot is identifiable

Key points

- Plot of data is a test statistic
- Type of plot used typically indicates null/alternative hypothesis, eg scatterplot suggests null hypothesis "no association between two variables"
- Null hypothesis suggests null generating mechanism
- Generate draws from the null, plot, show uninvolved observer
- Data plot detected equivalent to rejection of null, it is extreme relative to the sampling distribution

PURPOSE

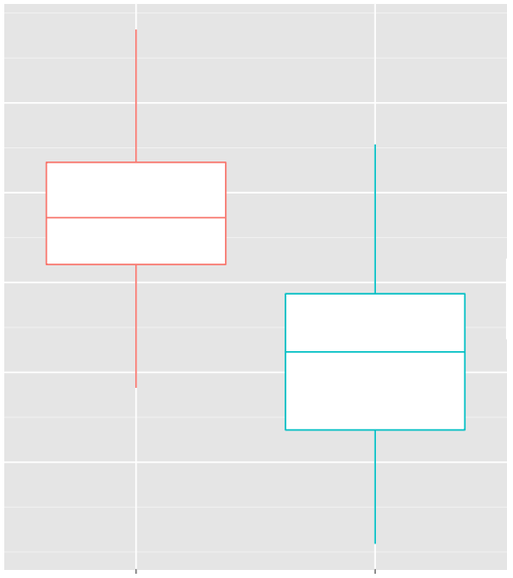
We do NOT replace EXISTING statistical tests.

The PURPOSE is to make tests for problems where NO CLASSICAL TEST EXISTS, and to provide exploratory visual methods with statistical significance.

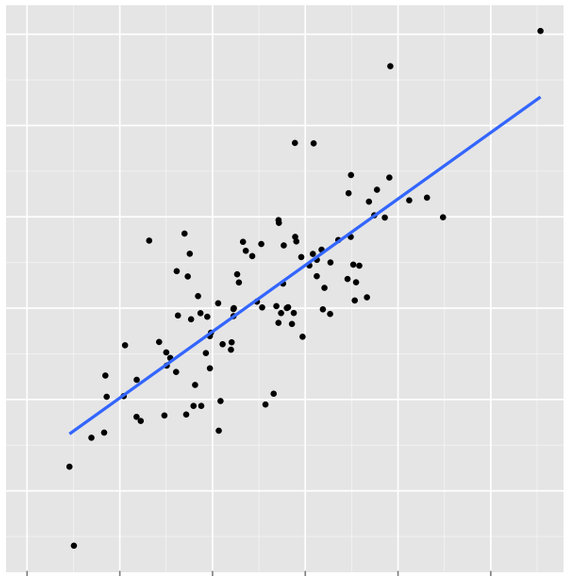
Validation experiments

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_1 X_{i2} + \beta_3 X_{i1} X_{i2} + \cdots + \epsilon_i$$

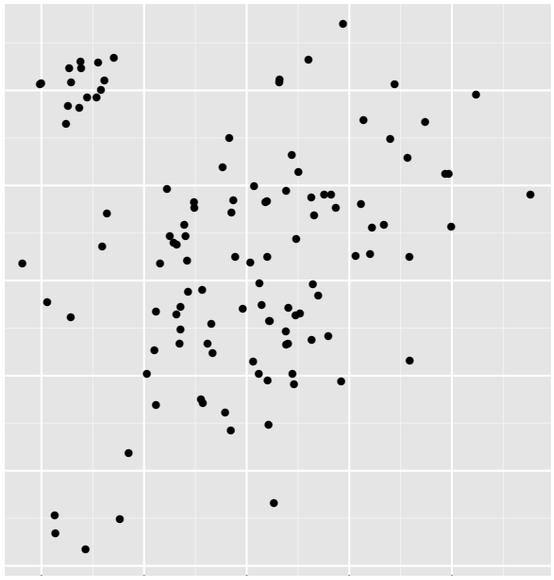
$$H_0 : \beta_k = 0 \text{ vs } H_1 : \beta_k \neq 0$$



Categorical X



Quantitative X



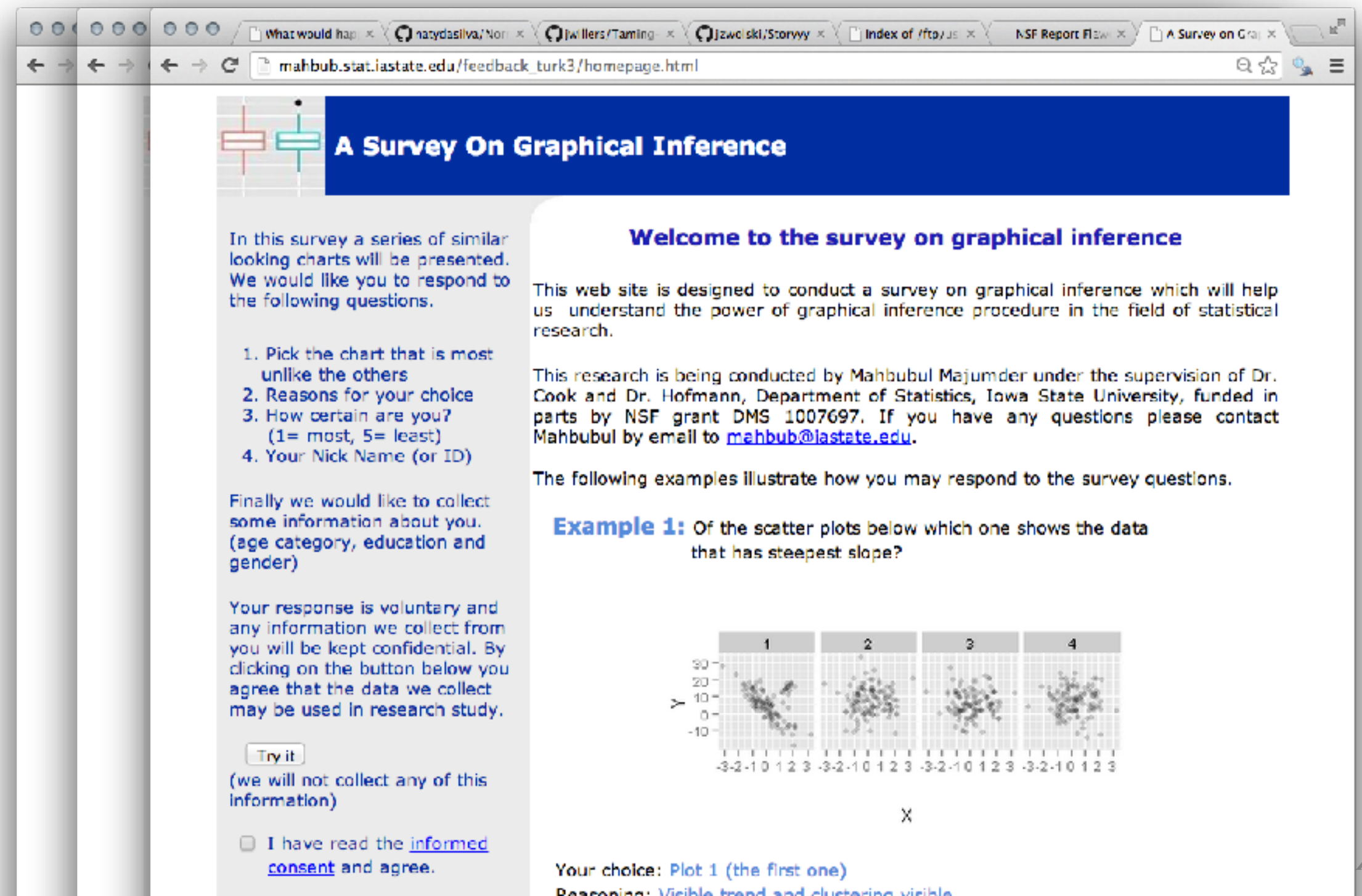
Contamination

Engaging observers

- Independent observers are engaged from Amazon's Mechanical Turk (https://www.mturk.com/mturk/welcome)
- Service named after "The Turk", middle ages mechanical chess machine that turned out to be a human
- Service for jobs where people can do better than machines

Experiments 1, 2, 3

http://mahbub.stat.iastate.edu/feedback_turk/homepage.html



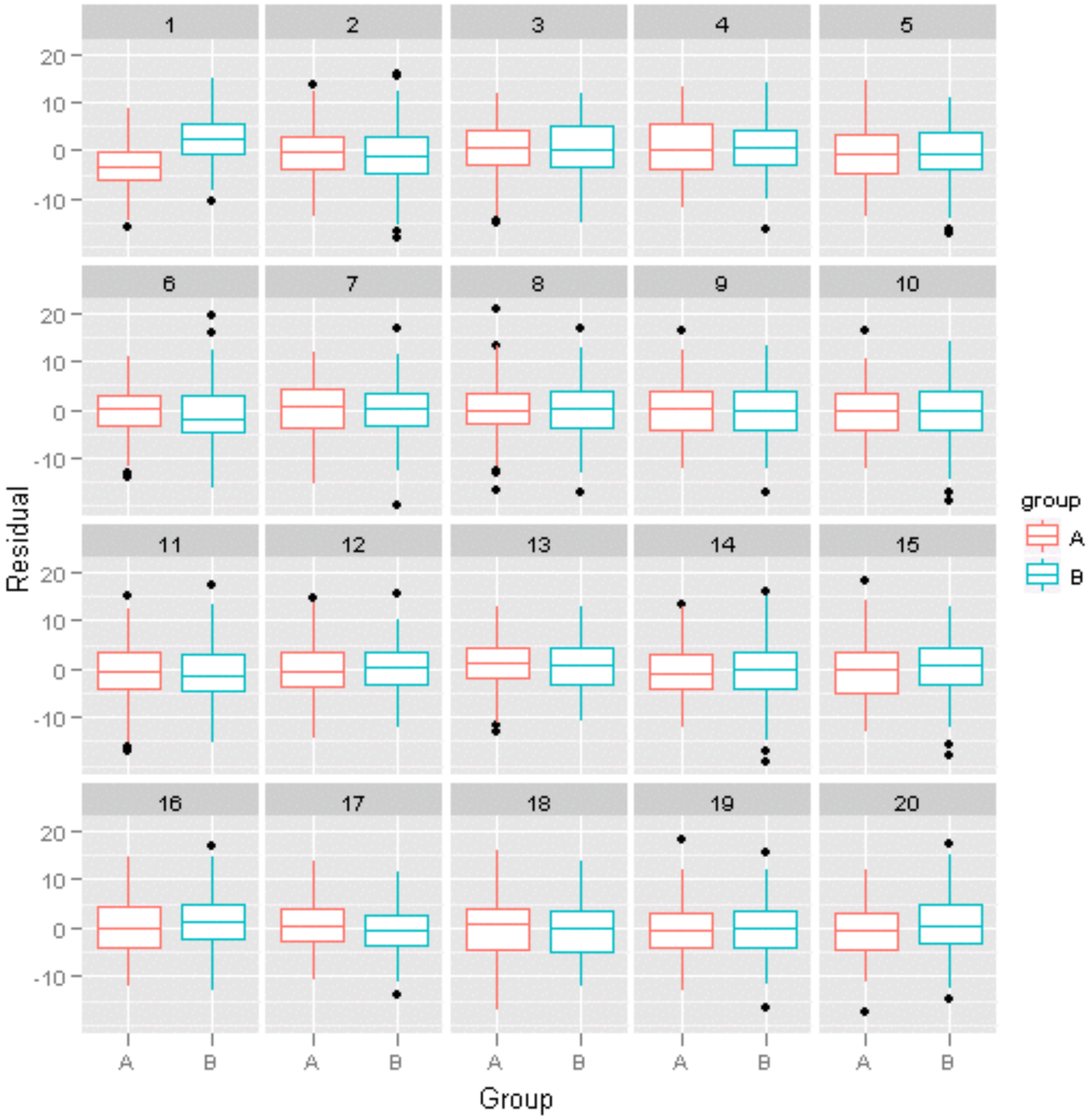
For the next 5 pages, record your choice (between 1, ..., 20) to answer this question:

Which plot shows the biggest vertical difference between group A and B?

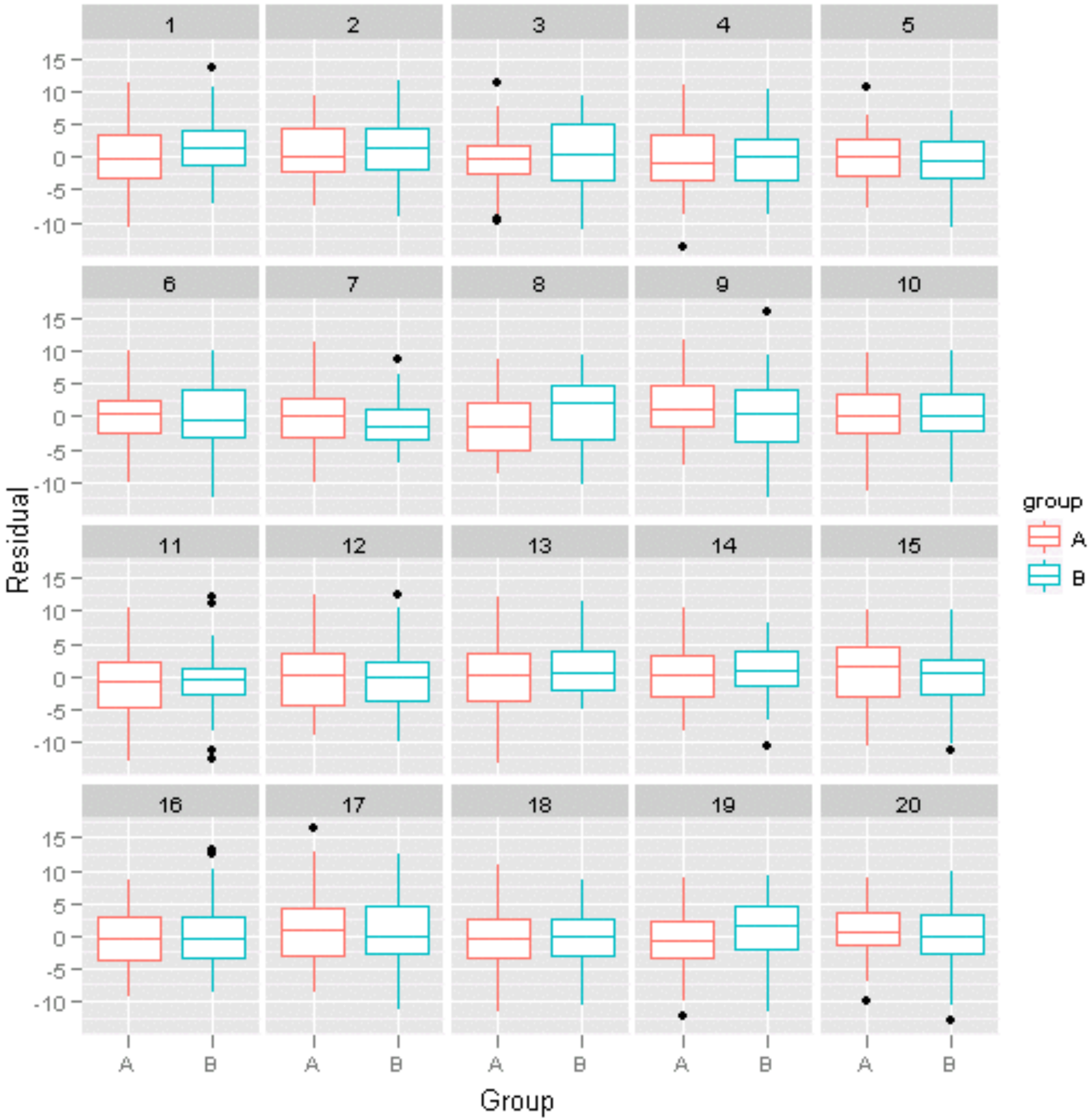
Enter your choice at:

http://bit.ly/NZSAORSNZ

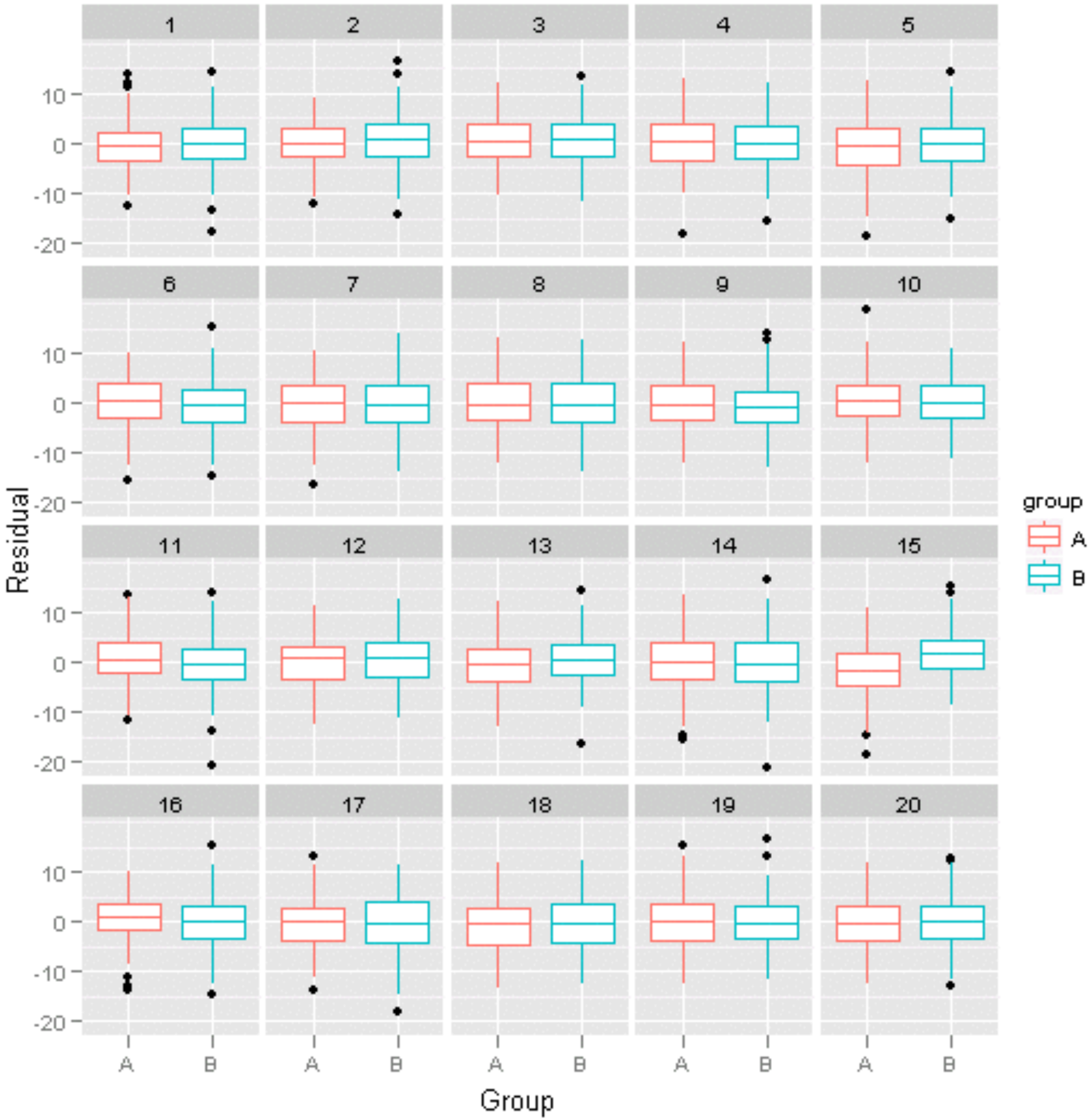
1


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
20
10
0
-10
Residual
A
B
Group
group
A
B

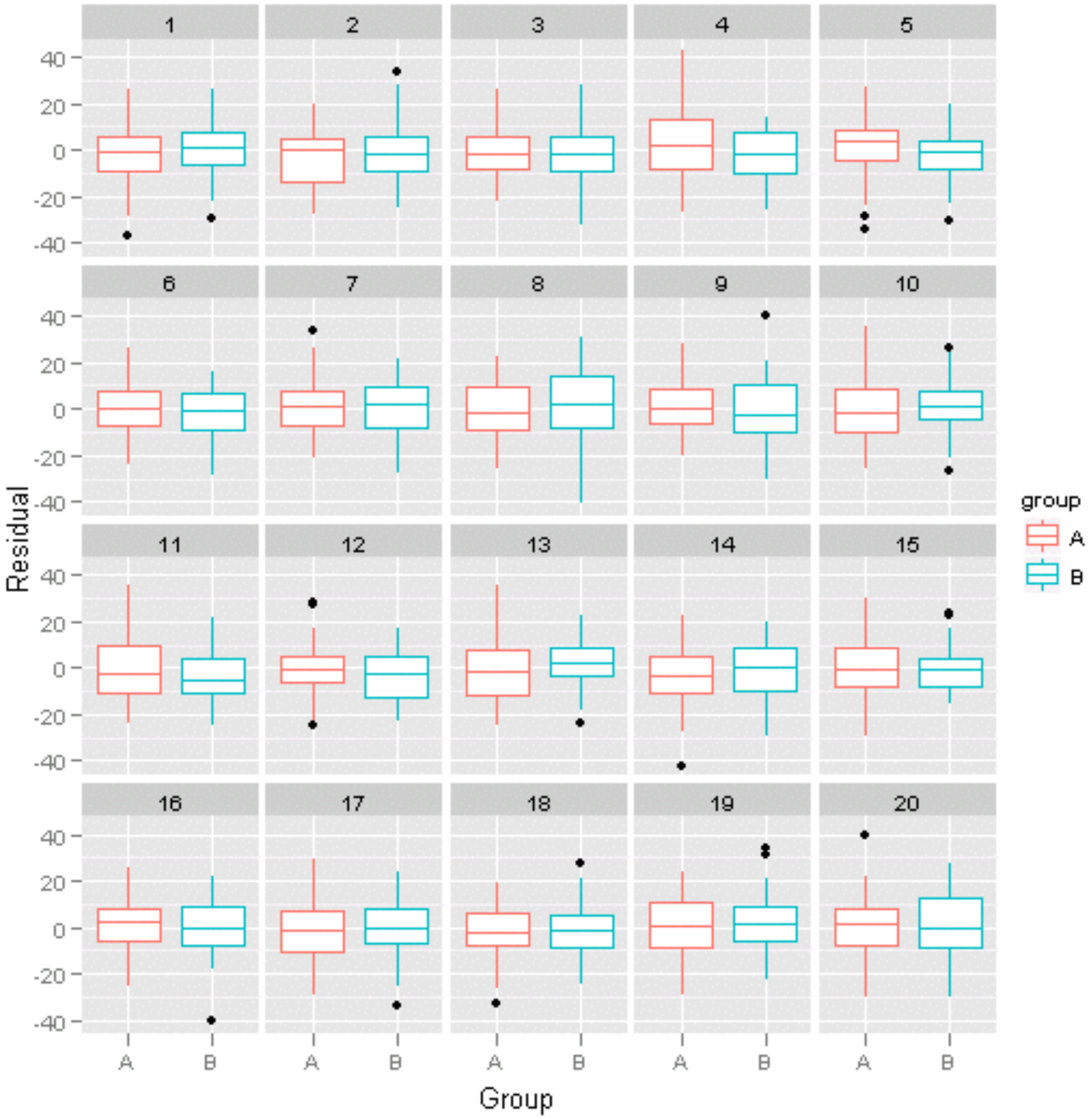
2


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
15
10
5
0
-5
-10
Residual
A
B
Group
group
A
B

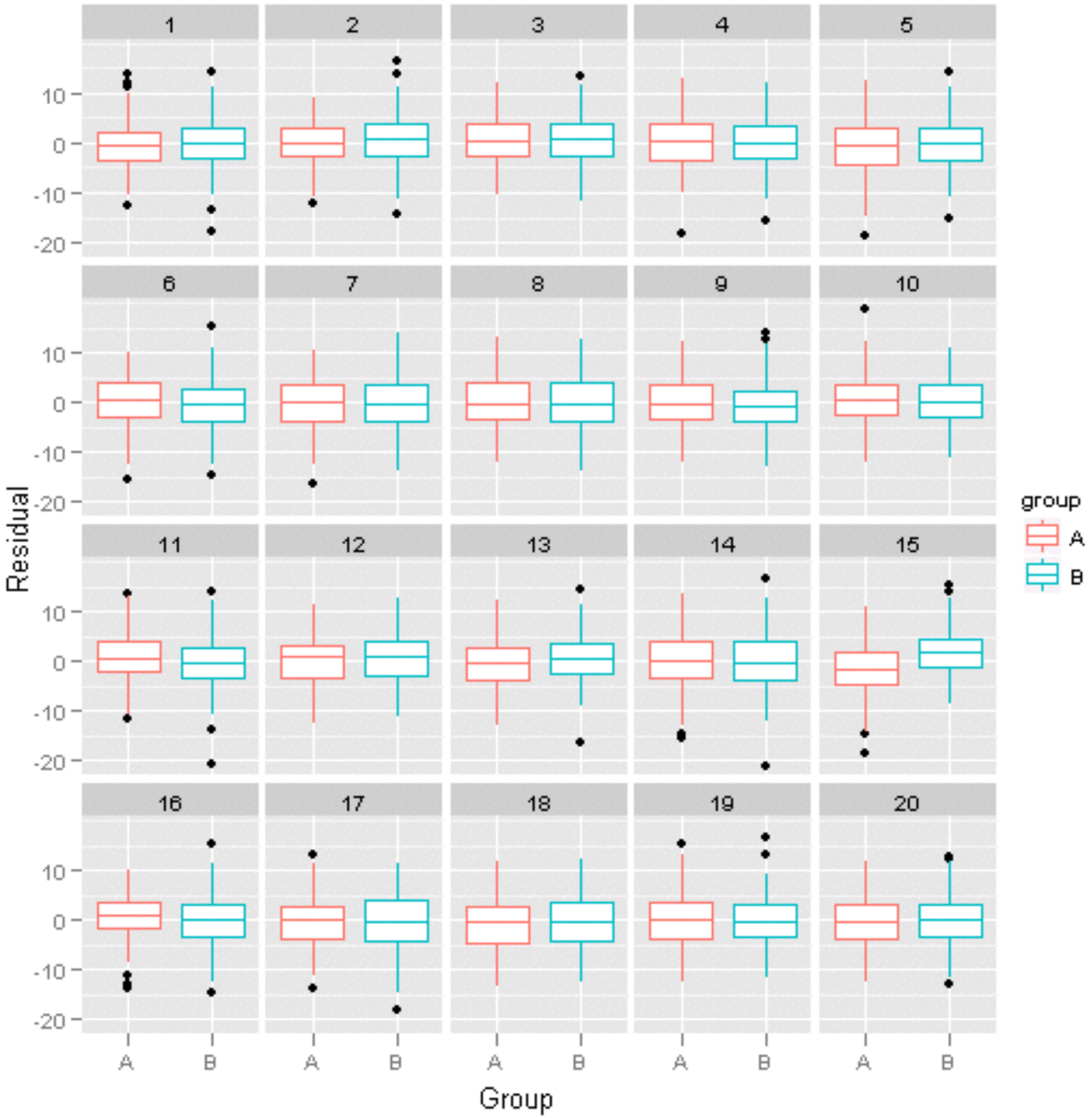
3


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
10
0
-10
-20
Residual
A
B
Group
group
A
B

4


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
40
20
0
-20
-40
Residual
A
B
Group
group
A
B

5


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
10
0
-10
-20
Residual
A
B
Group
group
A
B

Five examples

	TRUE	n	β	σ	effect
1	1	300	5	5	17.00
2	3	100	0	5	0.00
3	15	300	3	5	10.30
4	13	100	1	12	0.83
5	15	300	3	12	4.30

Significance

- What is the *p*-value?
- For one observer, the probability of randomly selecting the data plot is $^1/_m$, where m is the number of plots in the lineup.
- With multiple observers, the *p*-value is estimated by

Number of independent observers

$$P(X \geq x) = 1 - Binom_{K,1/m}(x-1) = \sum_{i=x}^{K} \binom{K}{i} \left(\frac{1}{m}\right)^i \left(\frac{m-1}{m}\right)^{K-i}$$

Number of observers choosing data plot

Source: Majumder et al (2013) JASA

Let's calculate the p-values for our trial

Power of the Test

- Power is used to compare the performance of tests in statistics.

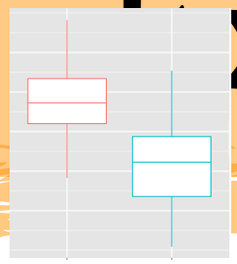
Definition 2.6. The *power* of a visual test, V_θ, is defined as the probability to reject the null hypothesis for a given parameter value θ:

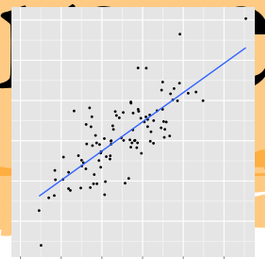
$$\text{Power}_V(\theta) = \Pr(\text{Reject } H_0 \mid \theta).$$

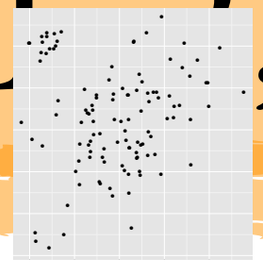
- Estimation requires assuming that the data plot has lowest p-value (controlled simulation studies), and incorporating the number of observers.

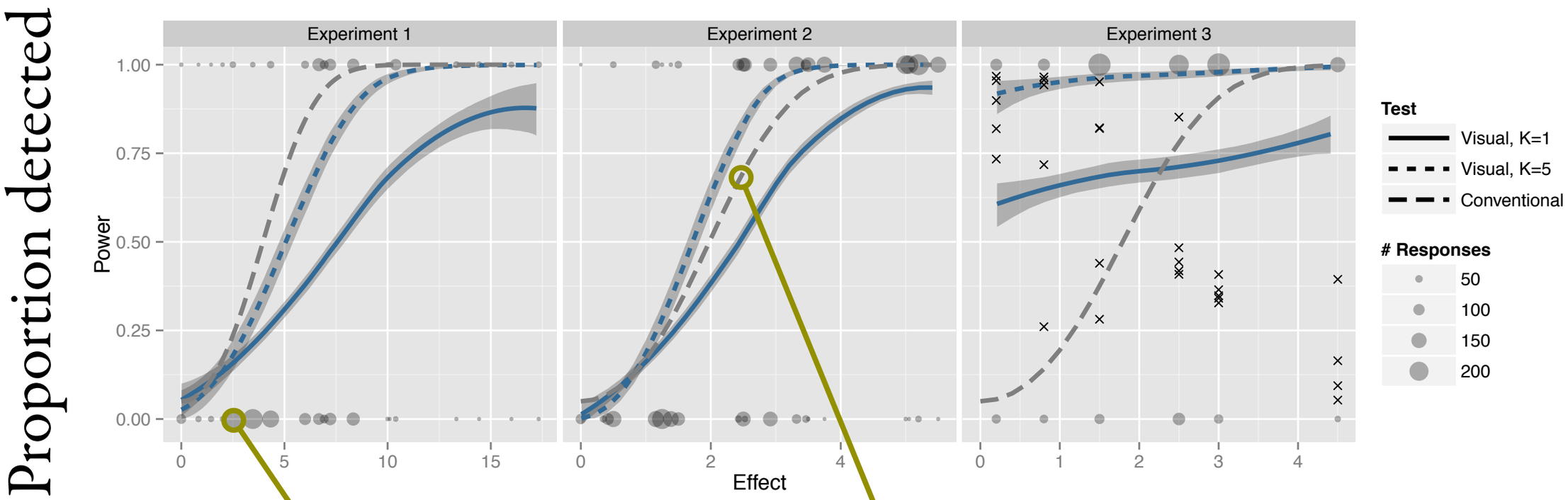
Source: Majumder et al (2013) JASA

Experiments 1, 2, 3









Effect $E = \sqrt{n} \cdot \beta / \sigma$

Subject choices

Conventional test

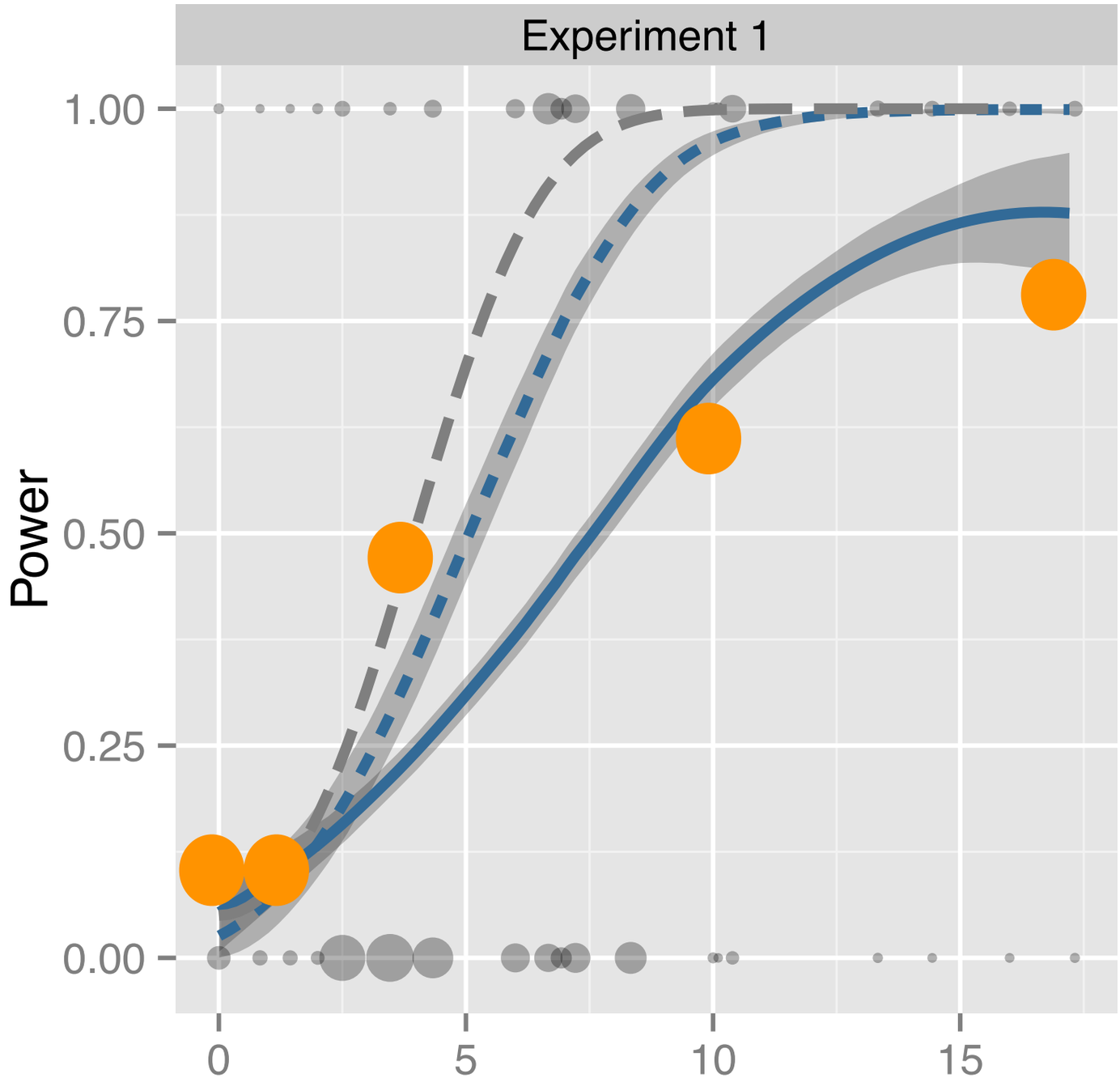
- Power matches conventional test in form, if effect size is "large" people see it.
- Pooling results from (5) observers improves the power, and it is possible to beat conventional test power.
- People beat conventional test when data was contaminated.

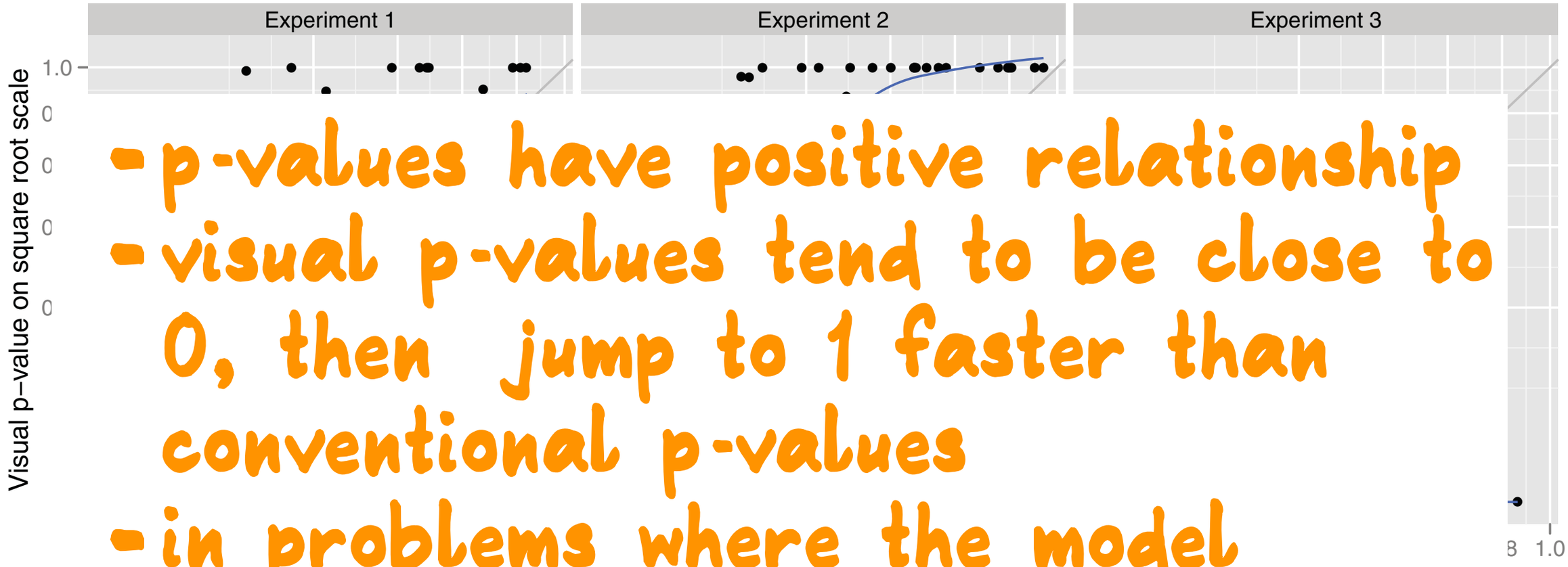
Source: Majumder et al (2013) JASA

Five examples

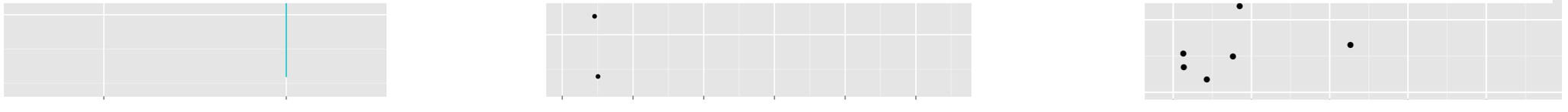
	TRUE	Selections	n	β	σ	effect
1	1	30/43=0.70	300	5	5	17.00
2	3	4/27=0.15	100	0	5	0.00
3	15	83/132=0.63	300	3	5	10.30
4	13	2/19=0.11	100	1	12	0.83
5	15	83/155=0.54	300	3	12	4.30

Where did you fit?

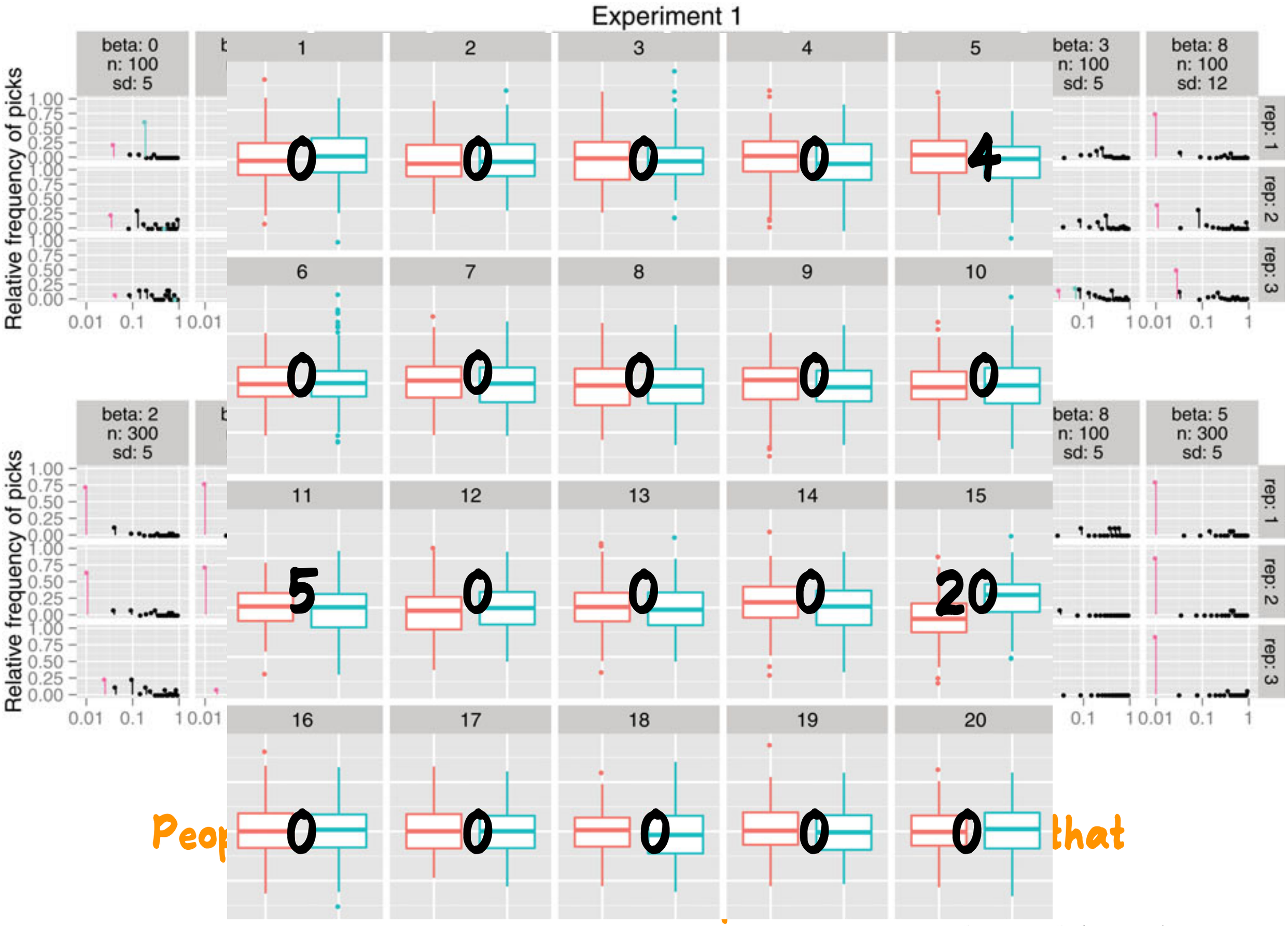




- p-values have positive relationship
- visual p-values tend to be close to 0, then jump to 1 faster than conventional p-values
- in problems where the model assumptions are violated, e.g. with contaminated data, the visual p-value is far superior to the conventional p-value



Source: Majumder et al (2013) JASA



Peop that

Source: Majumder et al (2013) JASA

Quantify patterns?

- Can we use this process to design metrics that capture structure, so that we can automate pattern recognition?
- Measure difference between data plot and null plots
- Match these to what people choose

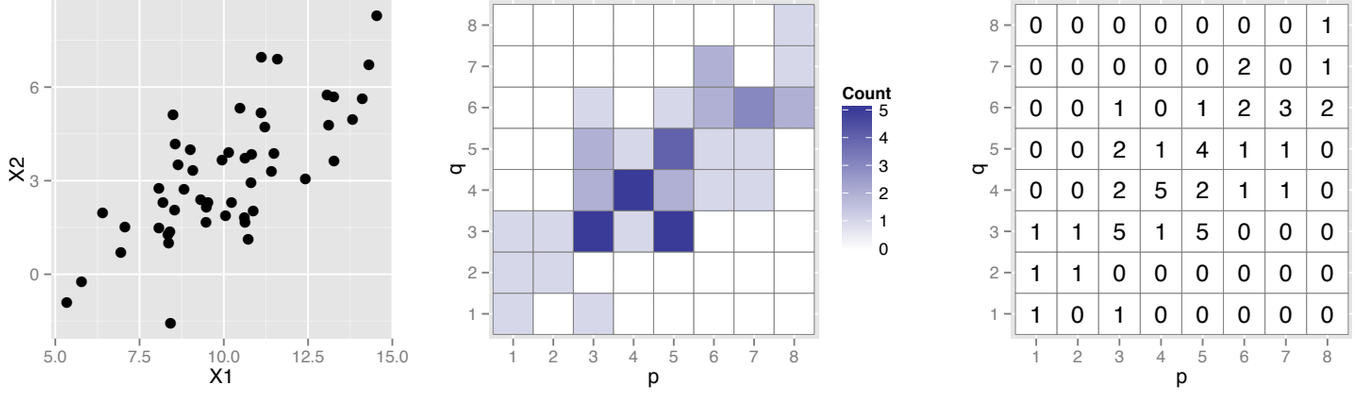
Source: Roy Chowdhury et al (2016) JCGS, submitted

Metrics

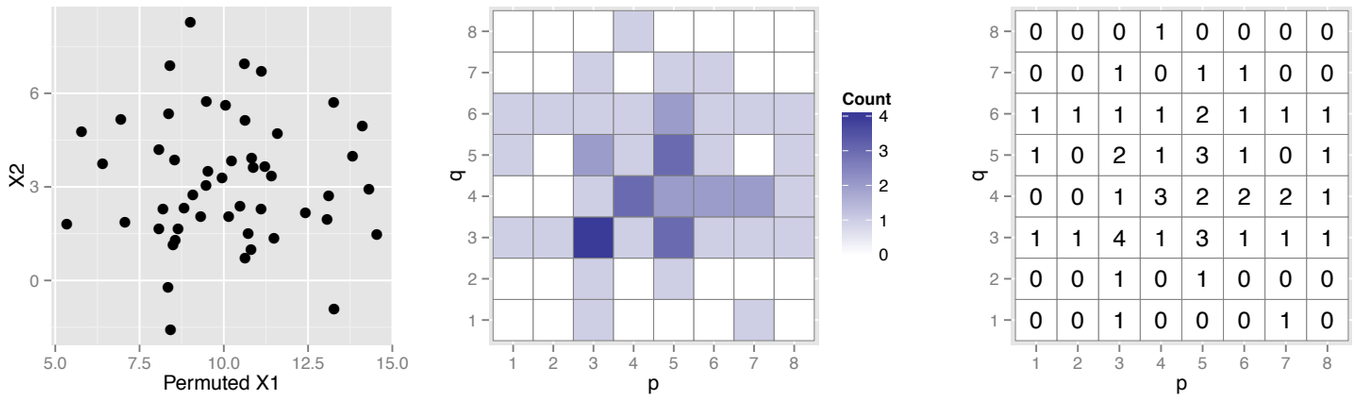
- Binned: bin plot space, count observations
- Boxplot: difference the Q1, M, Q3 between groups
- Regression line: slope and intercept

- Generate many more null plots. Compute average difference between (1) data plot and these nulls, and (2) null plots from lineup and these nulls, called δ-difference.

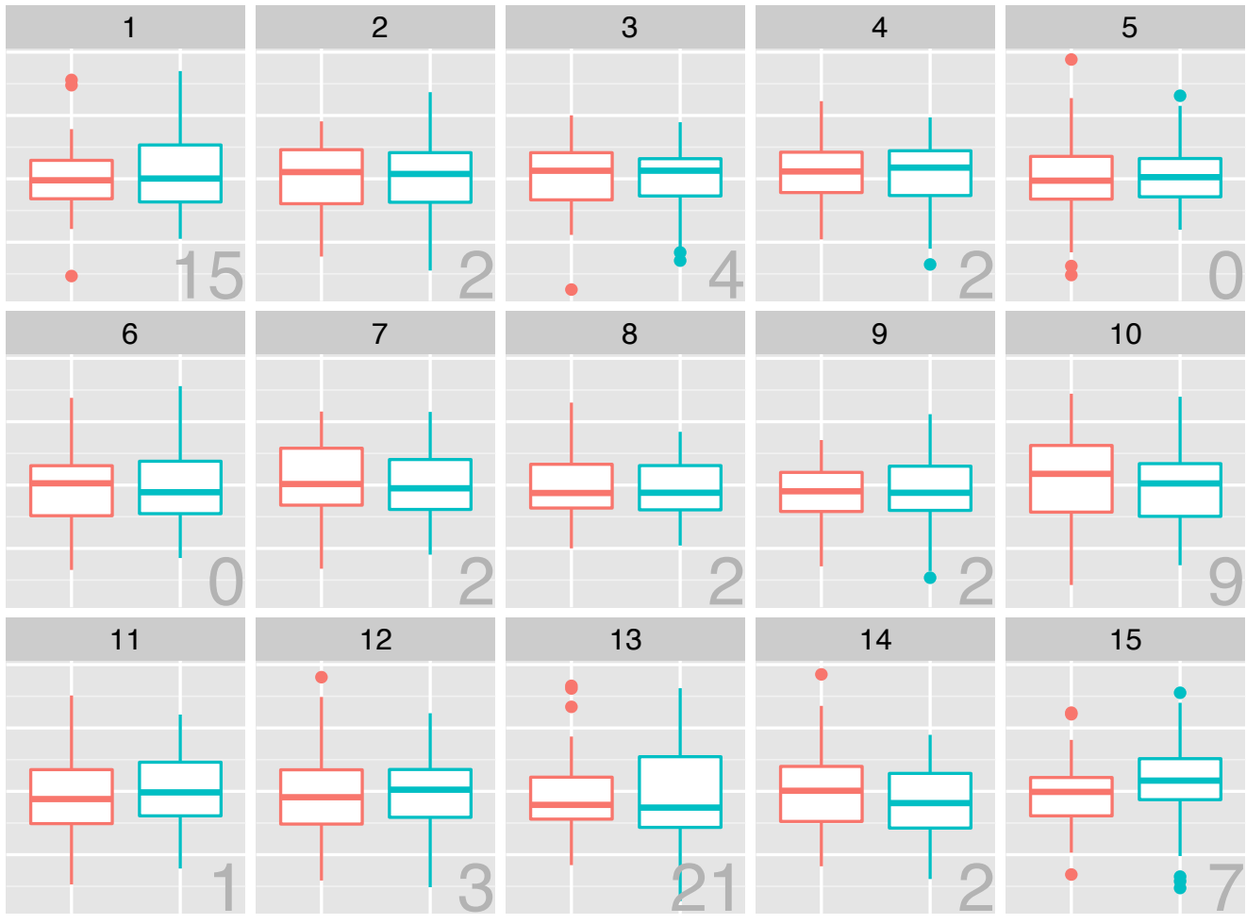
Binned distance



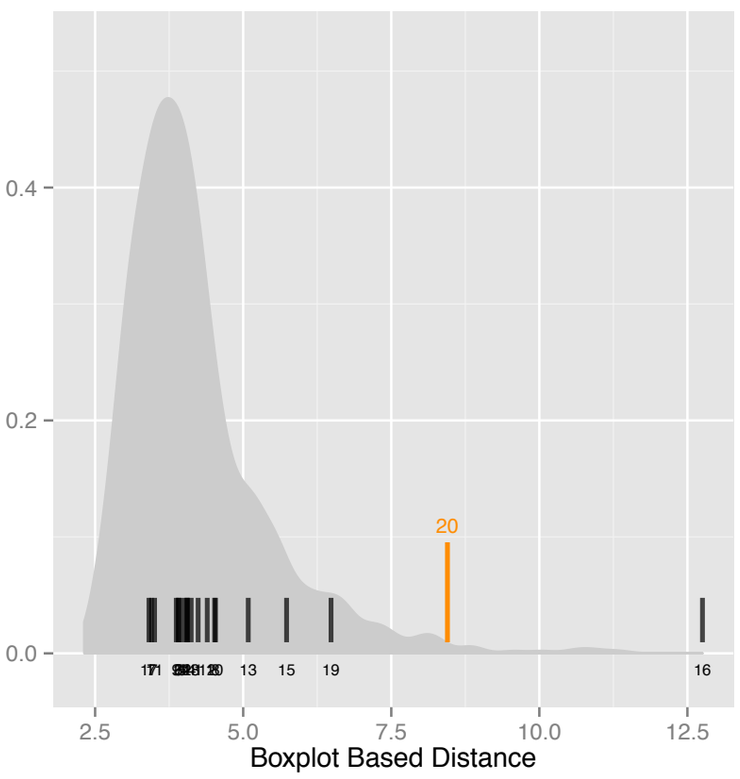
(b) Dataset Y with permuted X_1 and original X_2

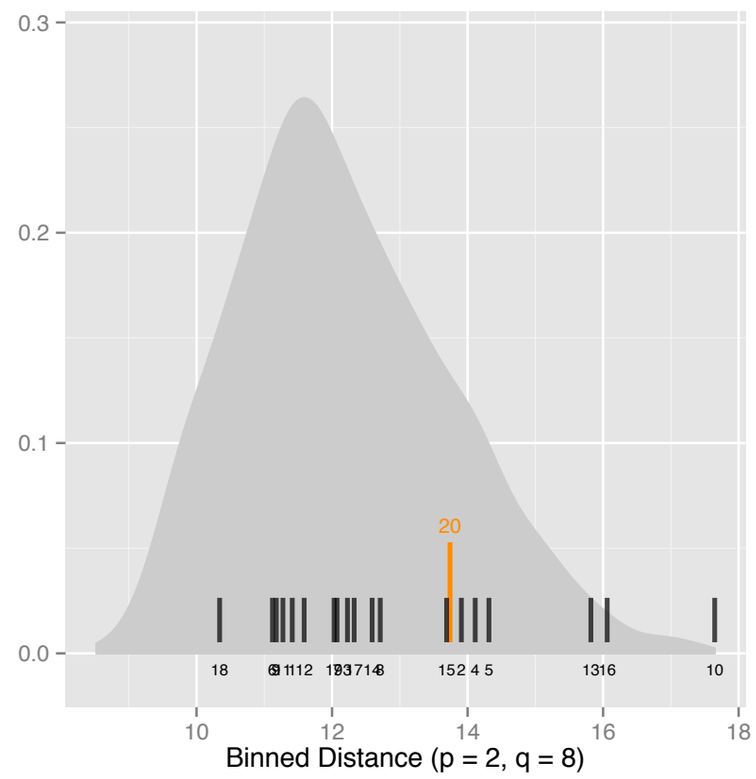


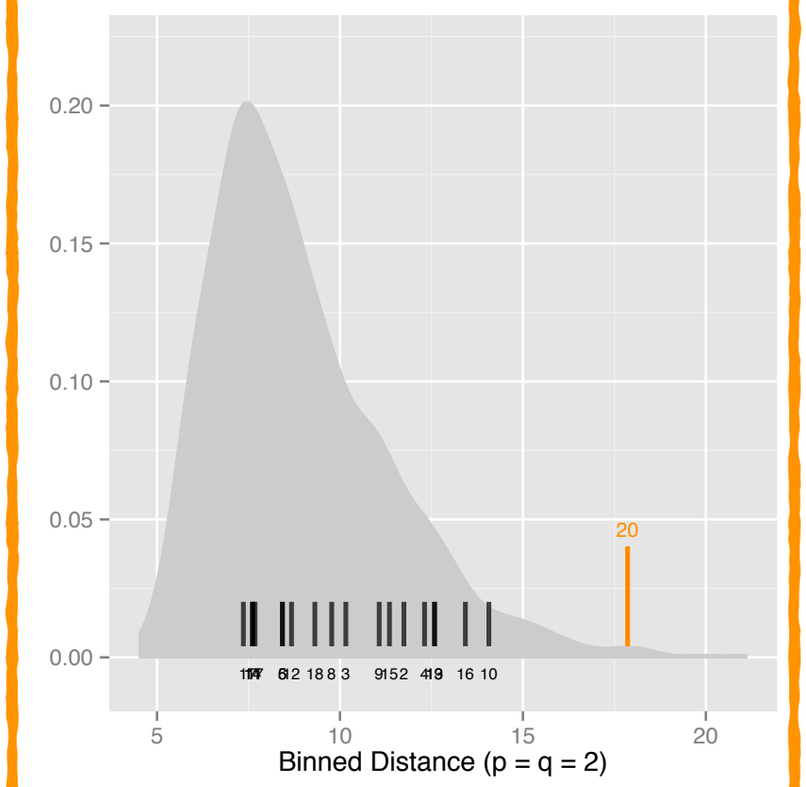
$$
\begin{aligned}
d^2_{BN}(X,Y) &:= ||C_X(X_1,X_2) - C_Y(X_1,X_2)||^2 \\
&= \sum_{i=1}^{p}\sum_{j=1}^{q}(C_X(X_{1i},X_{2j}) - C_Y(X_{1i},X_{2j}))^2.
\end{aligned}
$$



This metric best captures what people select



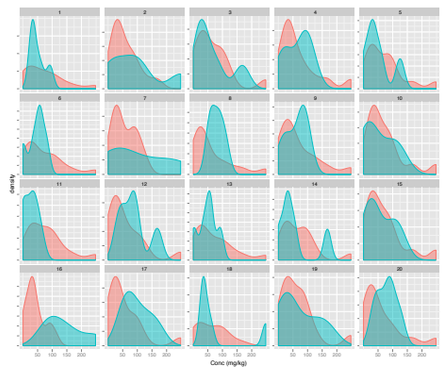
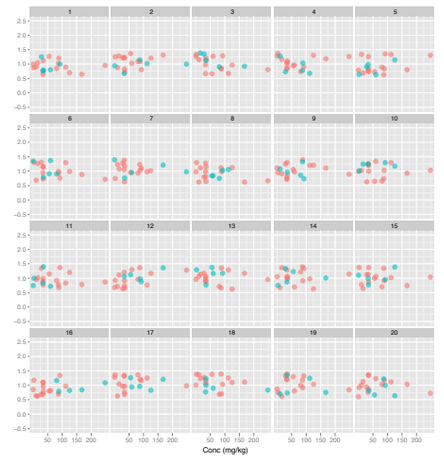
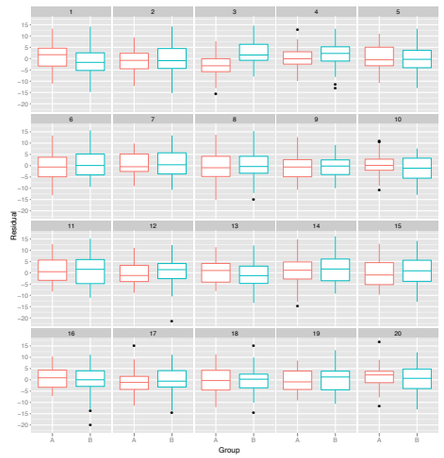
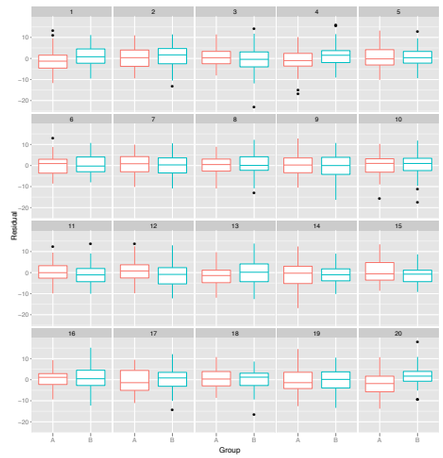
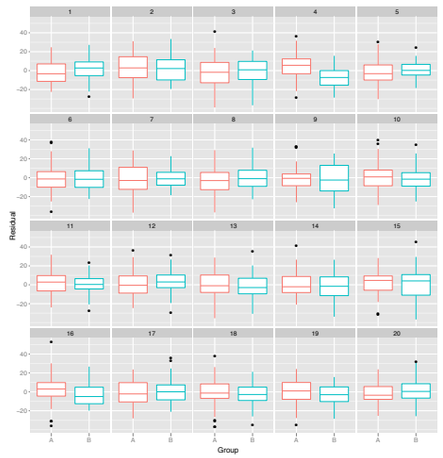
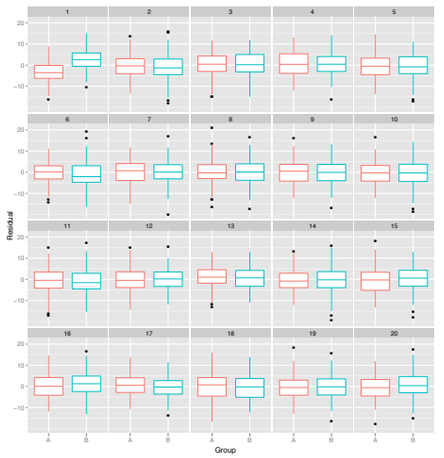
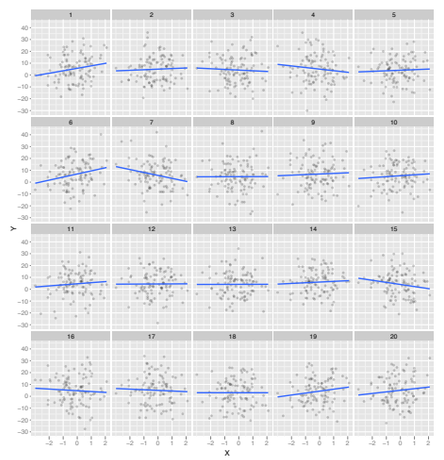
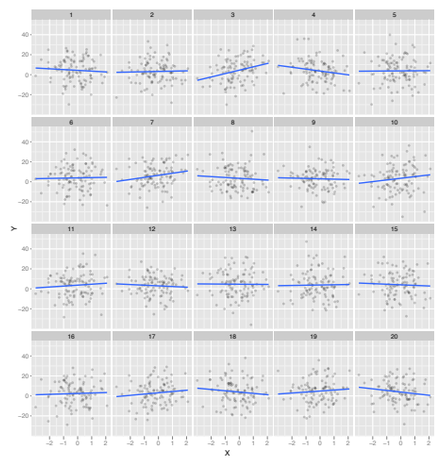
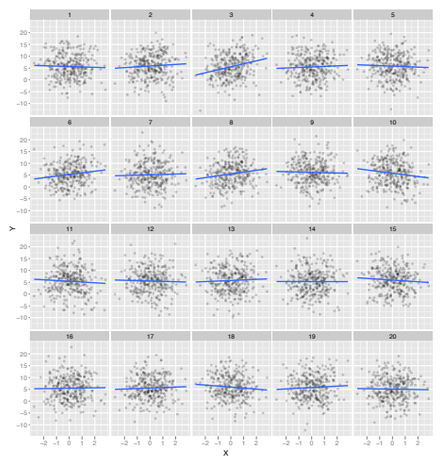
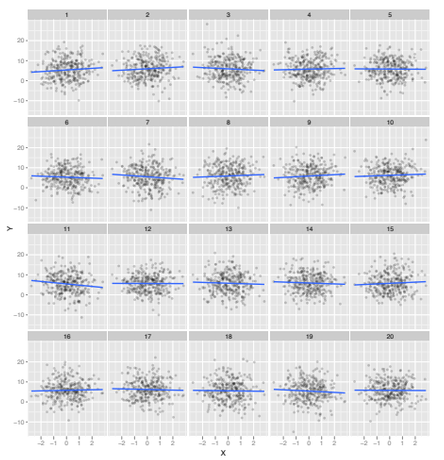


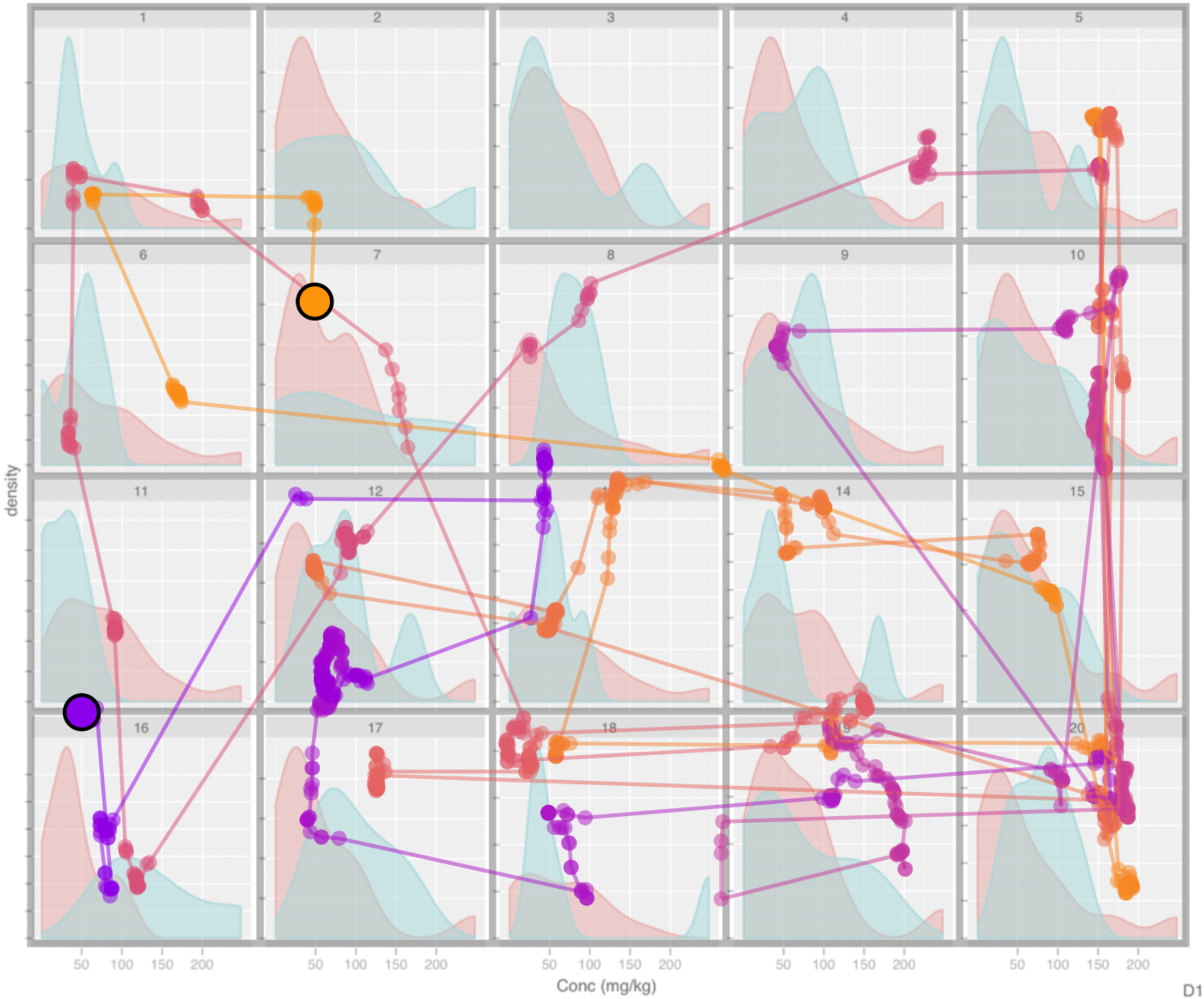


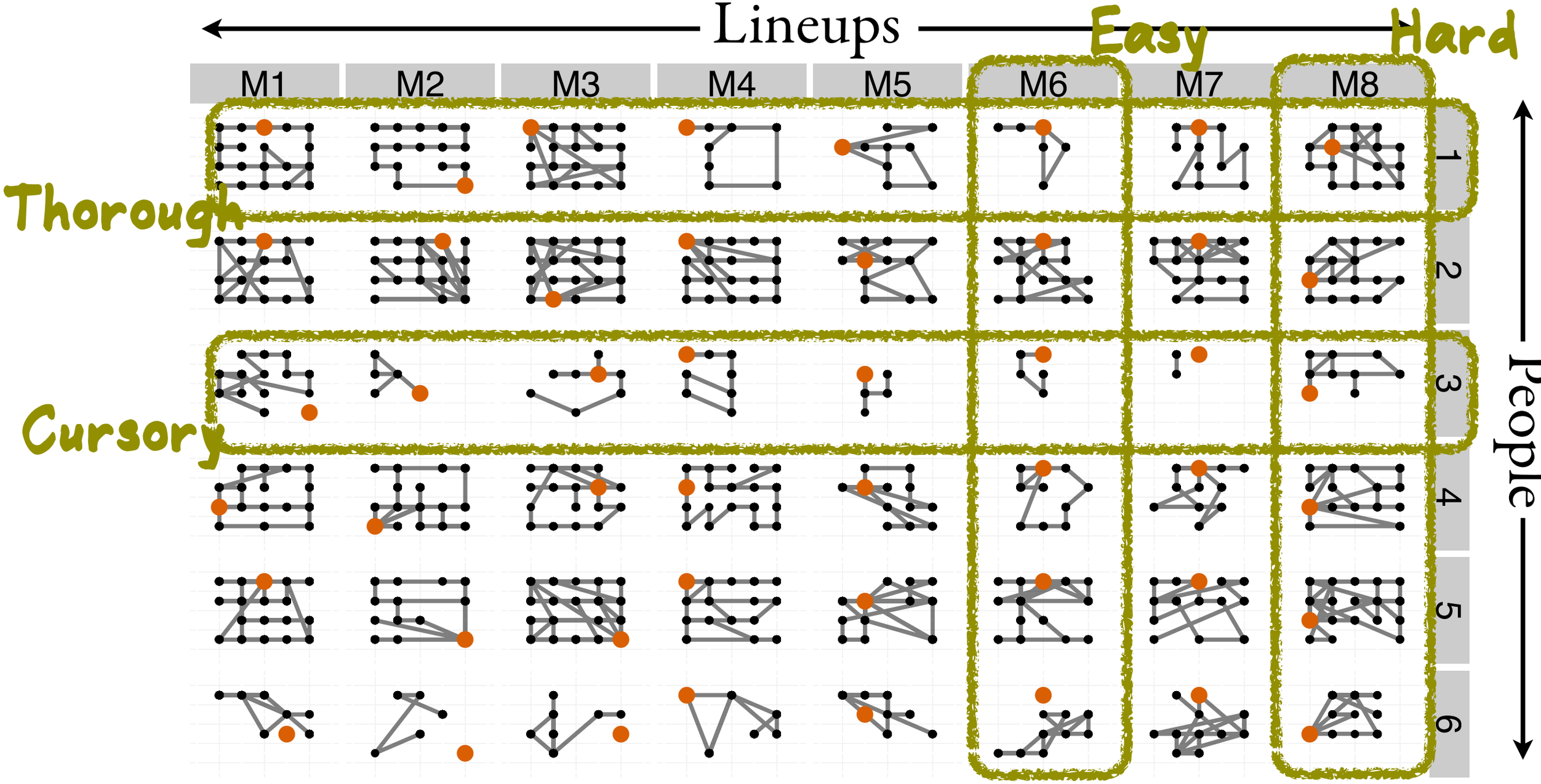
Follow-up eye tracking

- 12 lineups shown to 24 subjects
- Examined where there eyes focused in the page of plots while answering the same question as the Turk study

Source: Zhao et al (2013) IJITAS

D1	D2	D3	D4
			
M1	**M2**	**M3**	**M4**
			
M5	**M6**	**M7**	**M8**
			


1
2
3
4
5
6
7
8
9
10
11
12
14
15
16
17
18
20
density
50
100
150
200
Conc (mg/kg)
D1

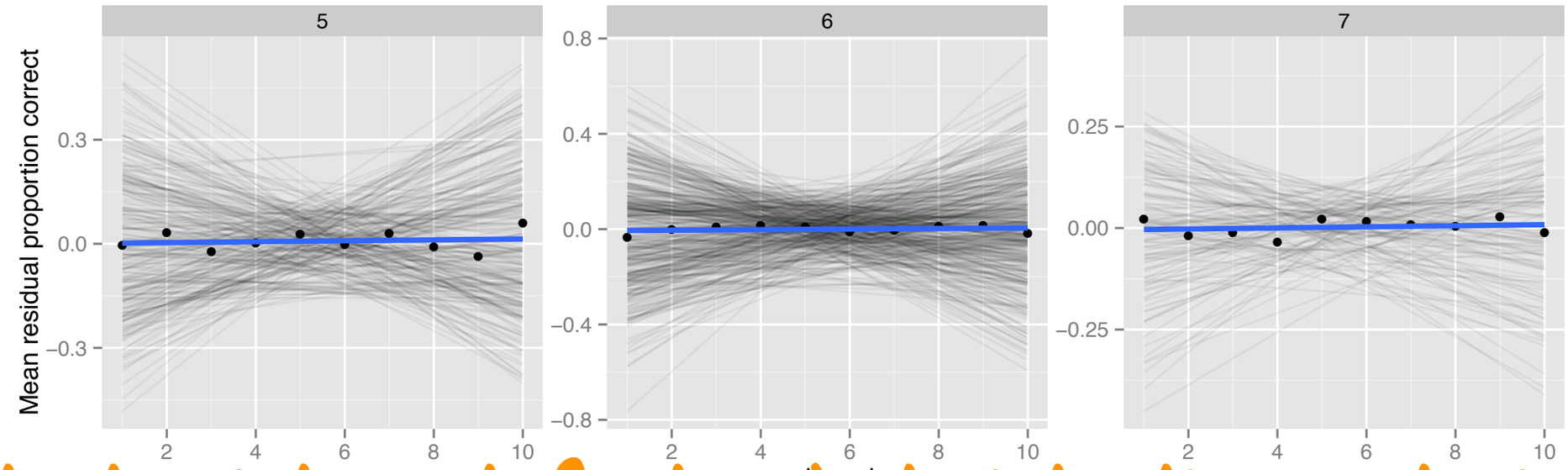


Source: Zhao et al (2013) IJITAS

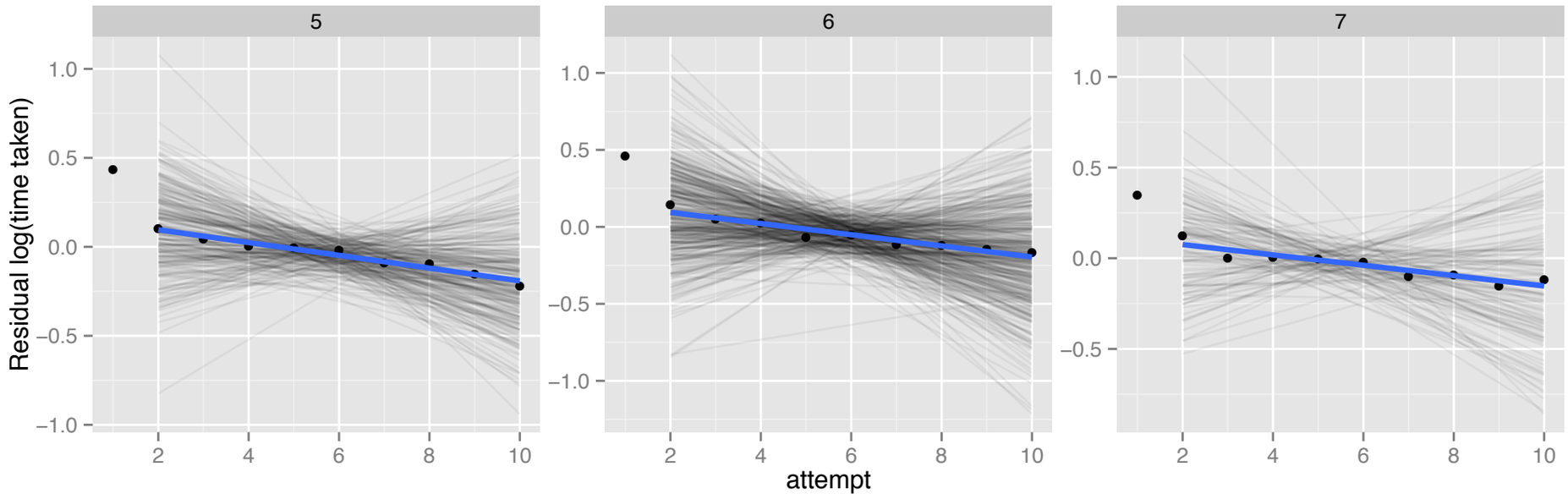
Eye-tracking

- People tended to methodically look around the lineup to find the most unusual plot
- For more difficult lineups all looked harder, and for easier lineups the looking was fast
- Some people were much faster than others in looking
- There are two stages, initial scan, and a final comparison between a few plots

Learning trends



Subjects tend to get faster but detection rate is same



Source: Majumder et al (2016) arXiv:1408.1974

Turkers

Factor	Levels	Participants Total	Participants %	Average Time	Number of Responses
Gender	Male	1348	57.63	48.51	13493
	Female	991	42.37	43.75	10564
Education	High school or less	193	8.24	37.21	2241
	Some under graduate courses	418	17.85	42.84	4070
	Under graduate degree	584	24.93	44.29	5775
	Some graduate courses	245	10.46	43.43	2460
	Graduate degree	902	38.51	52.18	9511
Age	18-25	740	31.61	42.97	7311
	26-30	547	23.36	46.27	5585
	31-35	376	16.06	44.27	3923
	36-40	257	10.98	55.03	2714
	41-45	141	6.02	43.90	1519
	46-50	95	4.05	49.29	1003
	51-55	83	3.54	48.67	867
	56-60	64	2.73	59.73	678
	above 60	38	1.62	48.67	457
Country	Uhited States	1087	46.83	39.64	10769
	India	980	42.22	52.63	10227
	Rest of the world	254	10.94	46.86	2819

R Package

- Nullabor package on CRAN
- When you plot your data, plot it first in a lineup, so you can be the unbiased observer

```
> lineup(null_permute("Obama.Romney"),
tracking.polls[,c(9,11)])
> decrypt("fg0t DARA up iYzuRuYp Q")
[1] "True data in position 5"
```

- Several null generating procedures included

Some foundations...

- Scott et al (1954): Generated synthetic plates to compare with real astronomical plates, acknowledged in Brillinger's (2005) Neyman lecture.
- Daniel (1976) had 40 pages of null plots for industrial applications.
- Diaconis (1983) describes 'magical thinking'.
- Buja et al (1988) describe 'Informal Statistical Inference' in association with the software Dataviewer.
- Gelman (2004) simulate data from statistical models.
- Davies (2008) suggest viewing null data sets.

Our experiments

- Experiments 1, 2, 3 compare lineup protocol with relevant classical test, result published in JASA
- Experiments 4, 5 examine plot design: cartesian vs polar, side-by-side boxplots vs dot plots
- Experiment 6 compares variations on boxplots: notched, violin, vase, beeswarm, ...
- Experiment 7 assesses large p, small n effects
- Experiment 9 tests for presence of any structure in an RNA-seq experiment

Our experiments

- Experiments 11, 13, 14, 15 assess diagnostic plots for hierarchical linear models.
- Experiment 12 tested significant expression in RNA-Seq data from a published study.
- Experiment 16-17 tested perception of dual structures in plots (clusters, trend) to determine dominant feature
- Experiment 18 compared line plots with scatterplots for perceiving association between two time series.

Summary

- People's eyes operate like classical test statistics.
- The detection is on effect size, yielding results closer to practical significance. We do not expect people to see statistical significance %^)
- The more people the more powerful the procedures.

Can we believe what we see?
Can we believe what we DON't see?
How can we discover what we don't know?

Next steps

- Better estimation of p-values
- Estimation and modeling of power
- Adjustments to these calculations to allow for people making multiple choices
- A web app to allow you to upload your data/plot, null-generating mechanism, to create lineups online, and have a team of expert readers evaluate

Acknowledgements

Plots produced using R package **`ggplot2`** by Hadley Wickham

Lineups made with R package **`nullabor`**

Projection pursuit (experiment 7) done using R package **`tourr`** by Wickham, Cook, with PDA index from Lee

National Science Foundation grant DMS 1007697

Heike

Andreas

Debby

Hadley

Eun-kyung

Eric

Mahbub

Tengfei

Susan

Adam

Niladri

Nathaniel