Scenarios of Visual Inference

- quantifying visual findings -

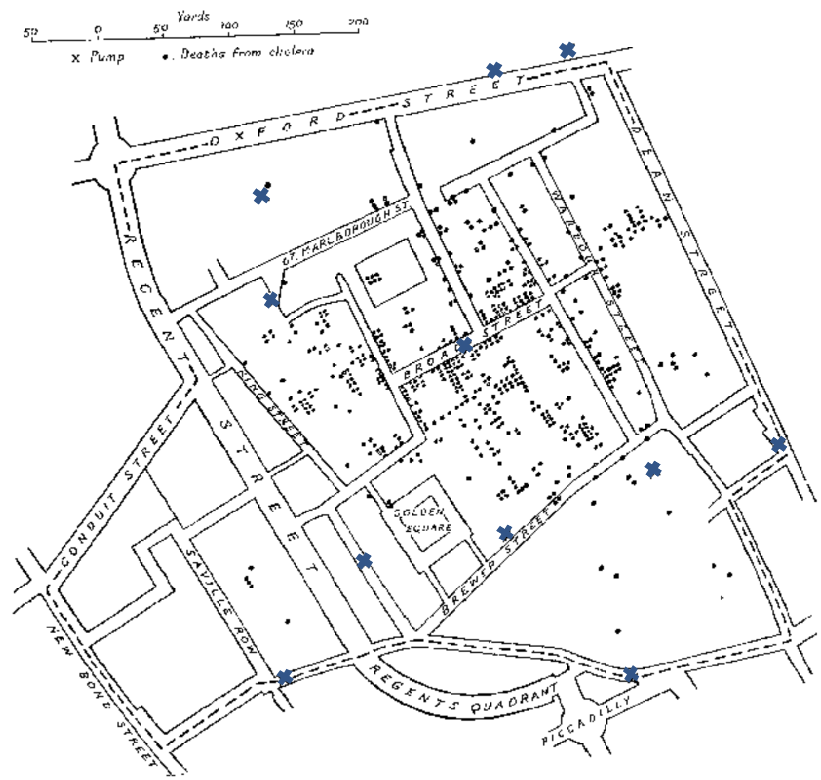
Heike Hofmann
Department of Statistics
Iowa State University

joint work with Susan VanderPlas and Dianne Cook

Outline

- Some examples
- A bit about the Lineup Protocol
- Inference in the lineup protocol

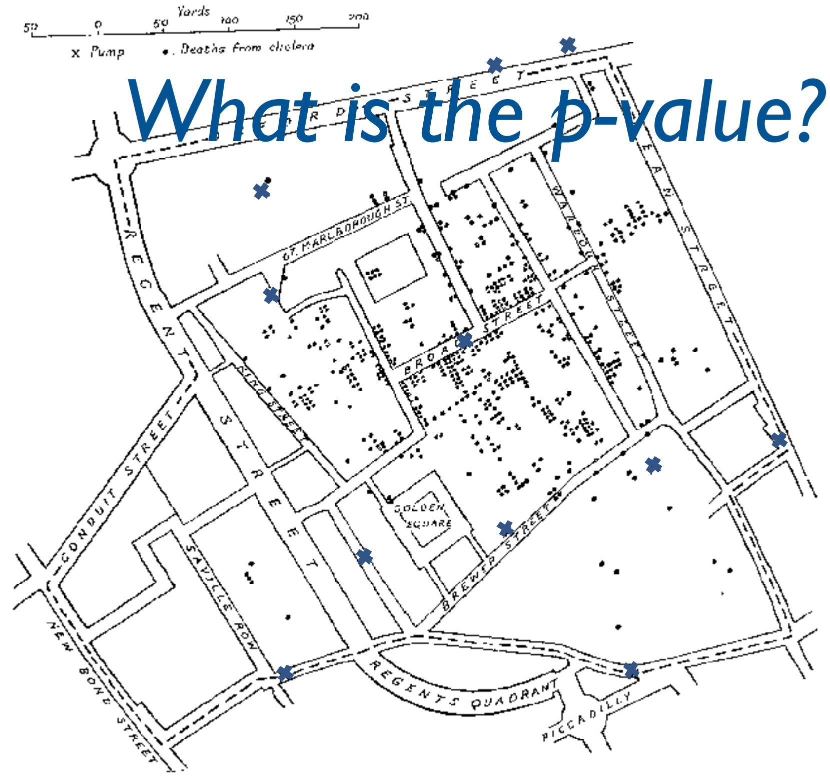
Why Visual Inference?



John Snow 1854

- Graphics are essential tools for data exploration, but …
- … post-hoc inferential results are invalid (data fishing, trawling, snooping …)
- Need: quantitative assessment of significance of graphical finding based directly on graphic

Why Visual Inference?

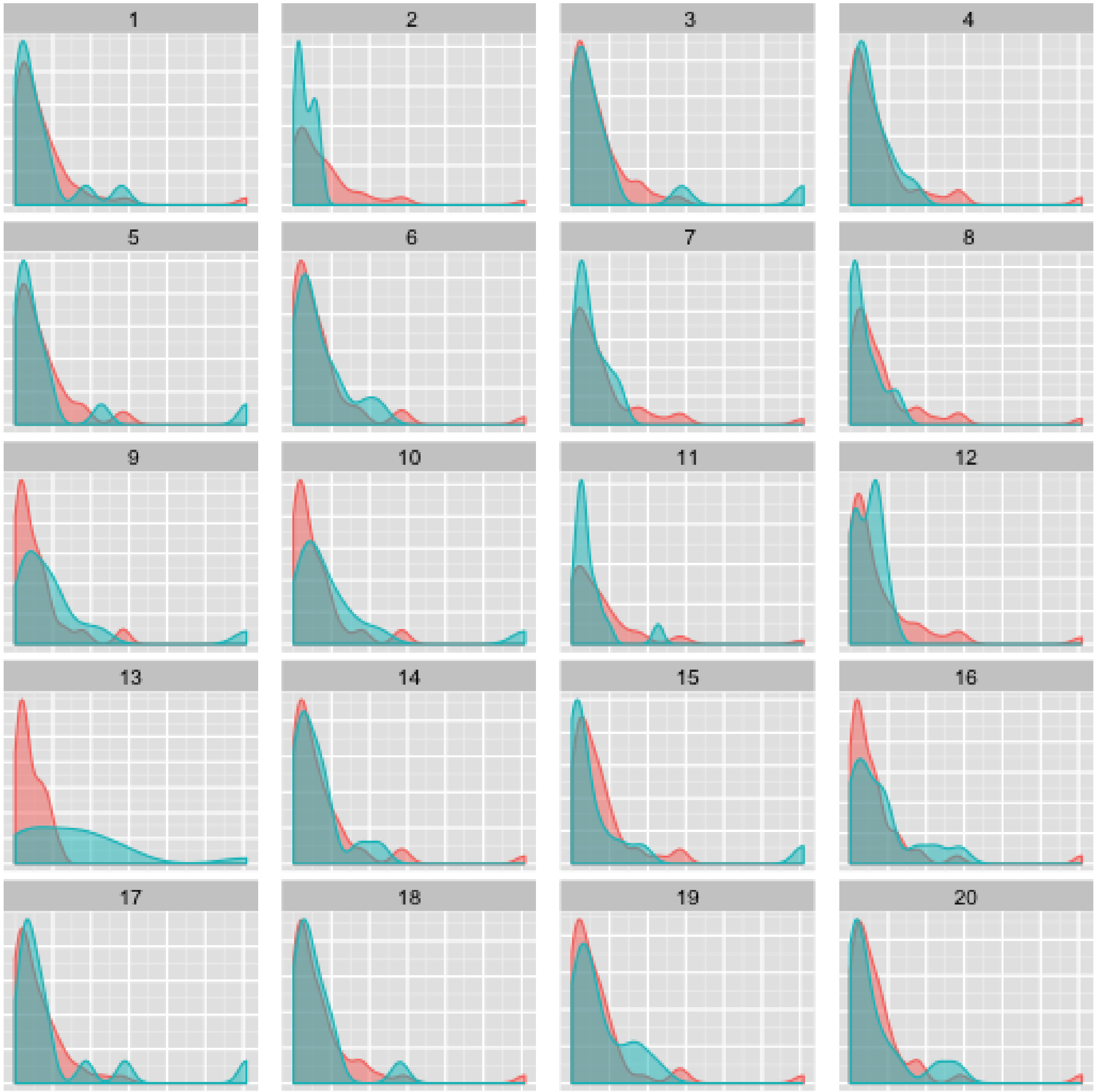


What is the p-value?

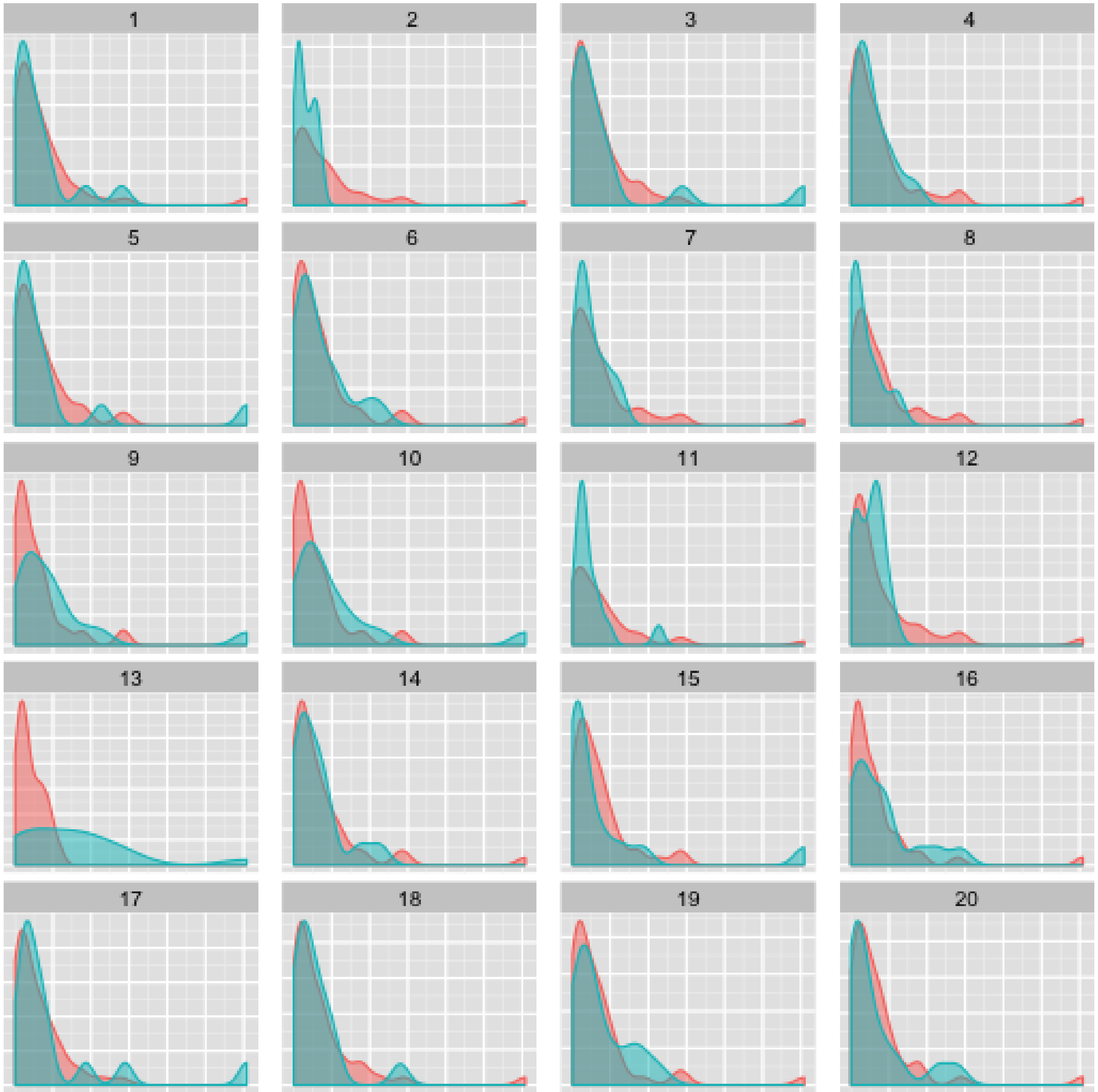
John Snow 1854

- Graphics are essential tools for data exploration, but …
- … post-hoc inferential results are invalid (data fishing, trawling, snooping …)
- Need: quantitative assessment of significance of graphical finding based directly on graphic

Which of these panels looks the most different?

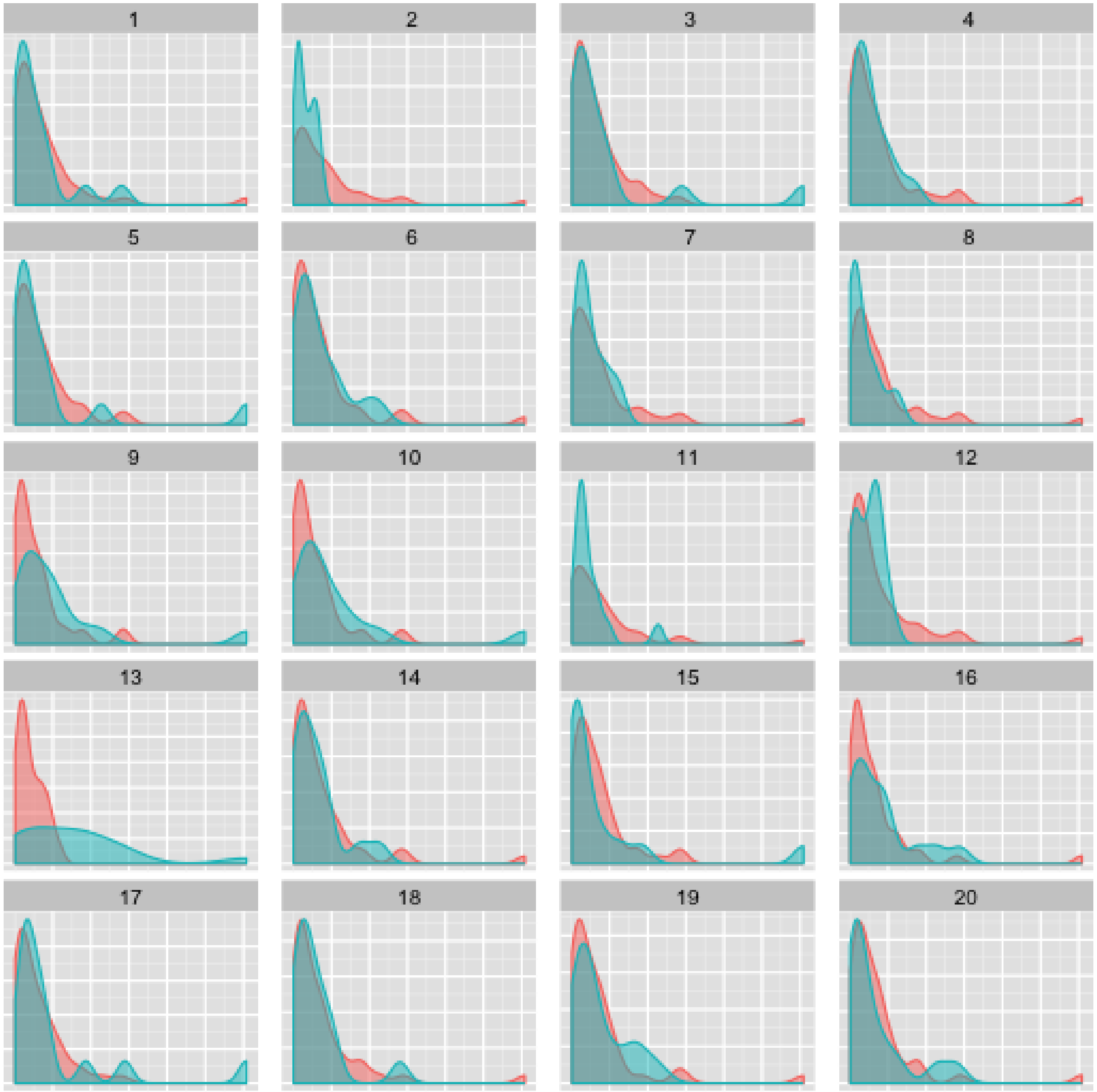


Which of these panels looks the most different?



data is in panel #13

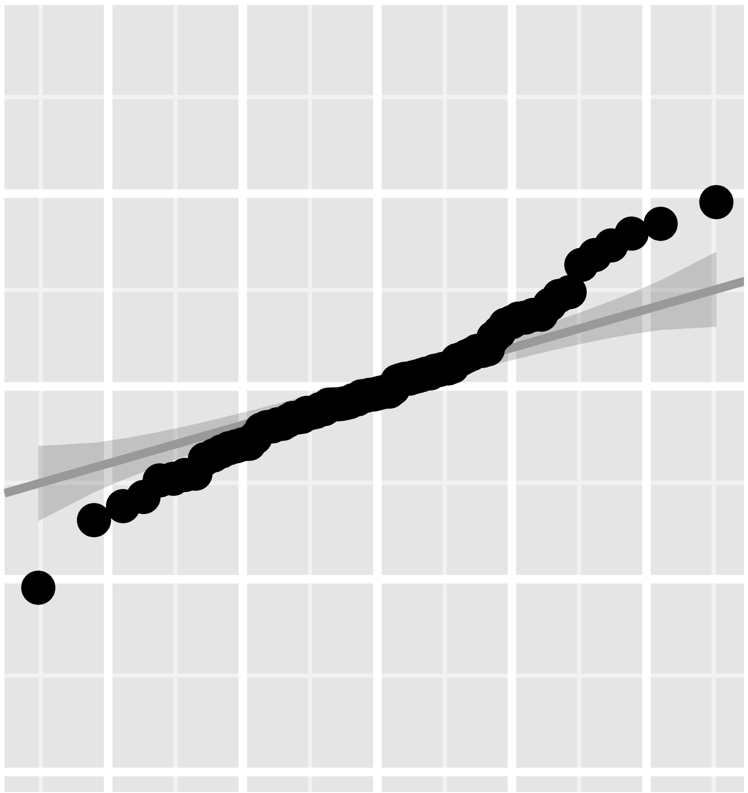
Which of these panels looks the most different?



data is in panel #13

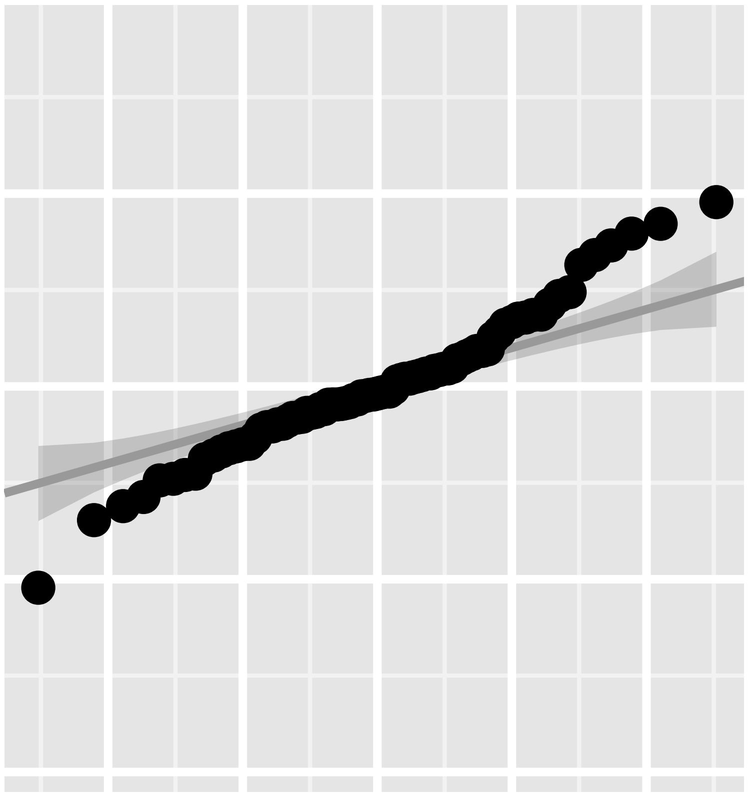
20/23 participants identified #13 as the most different

Is this normal?



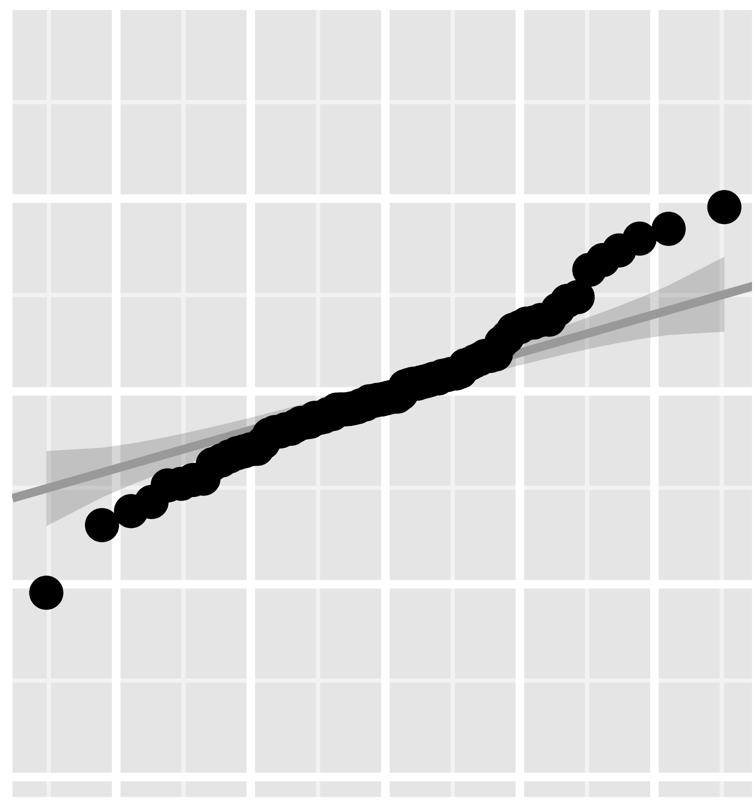
Is this normal?

Normal Q-Q plot



Is this normal?

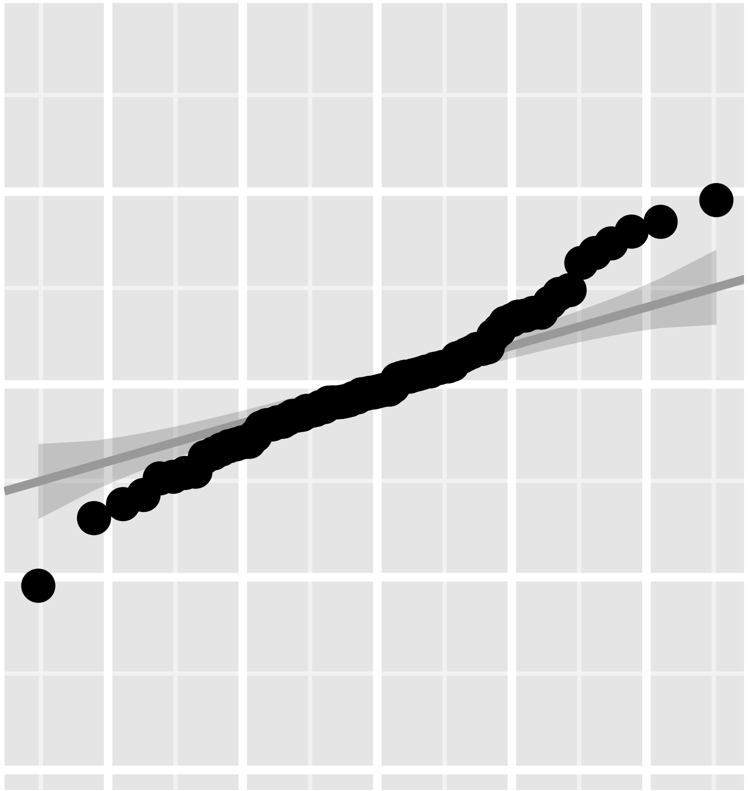
Normal Q-Q plot



Obvious deviations from normality assumption

Is this normal?

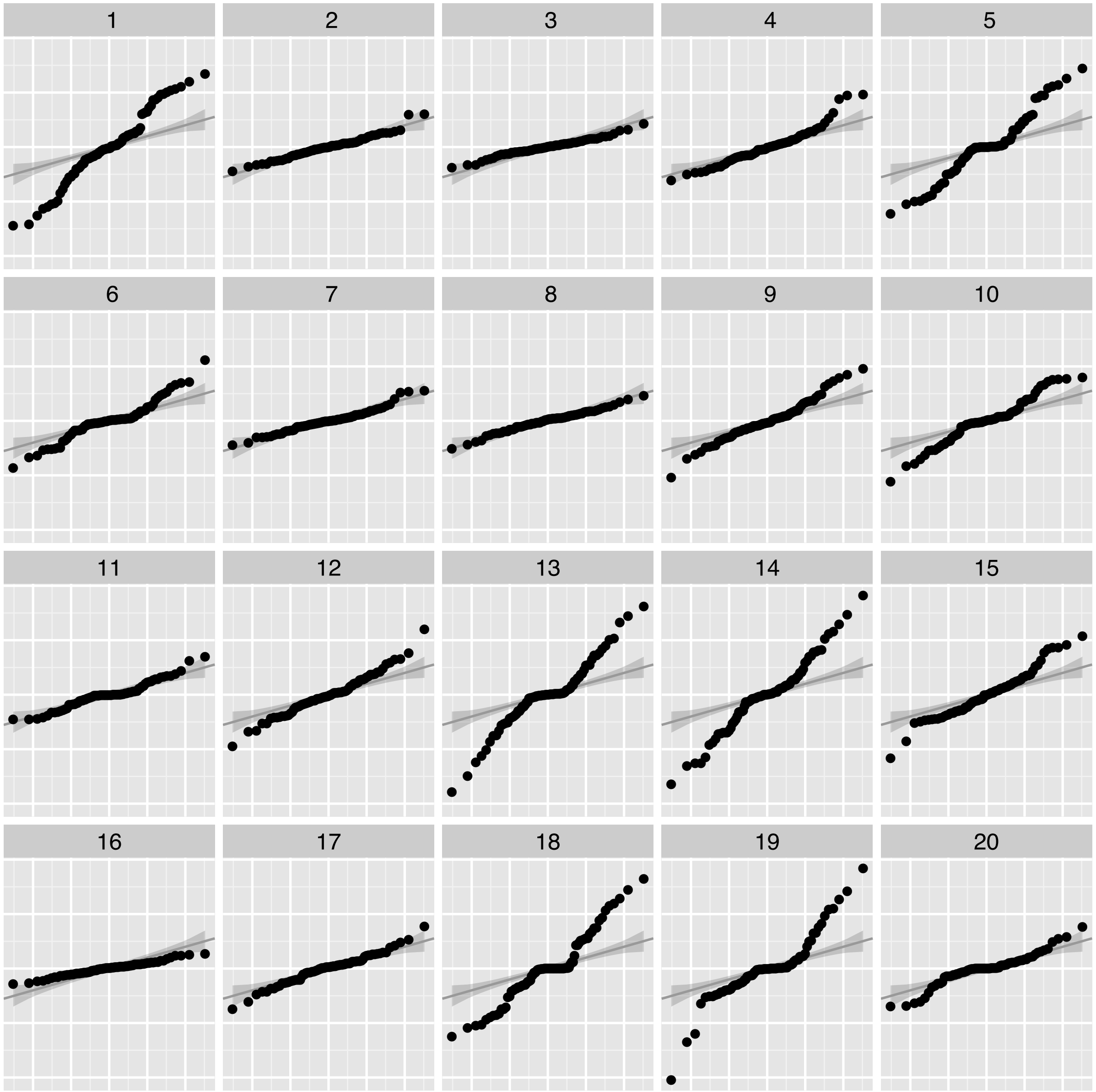
Normal Q-Q plot



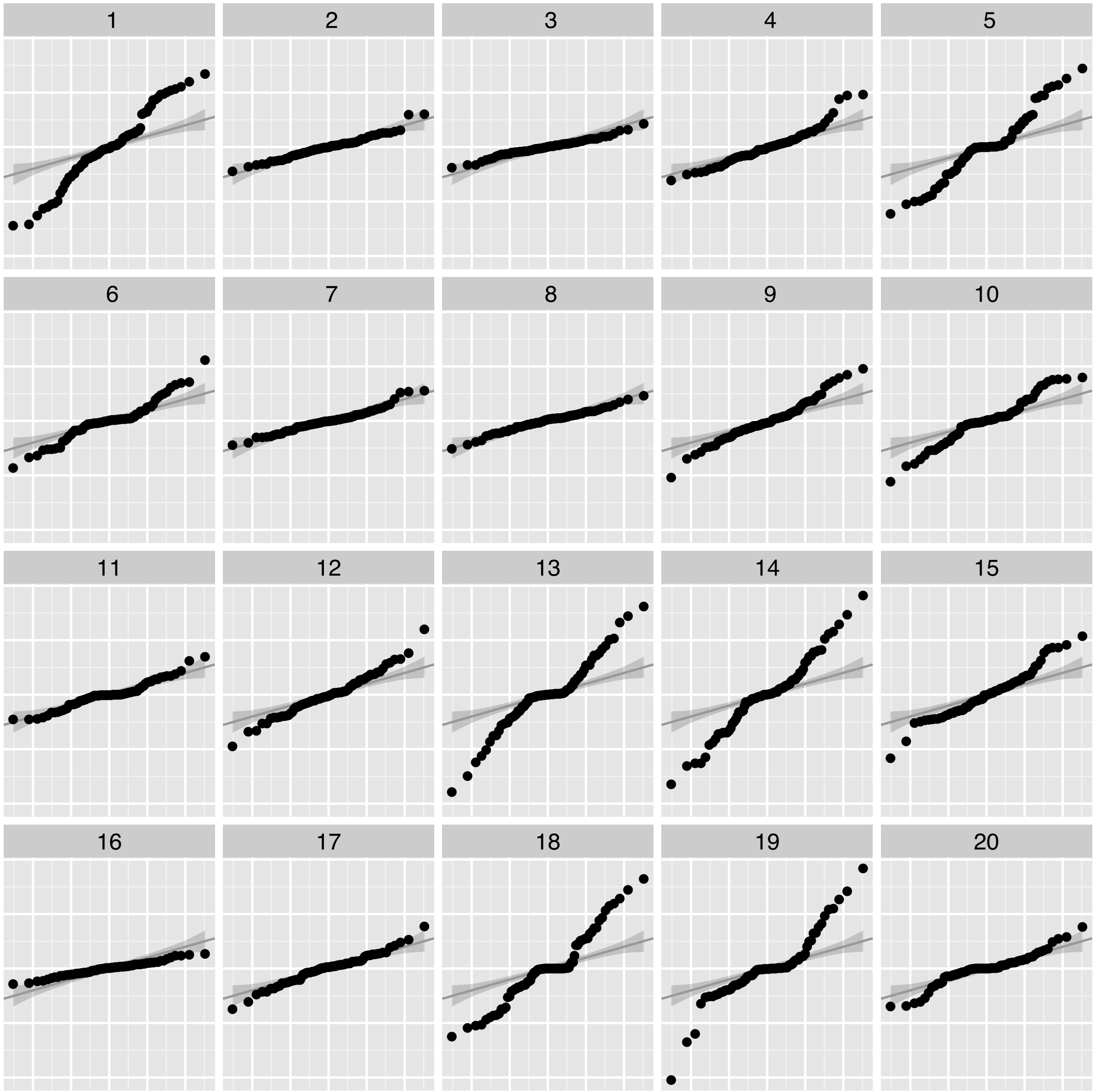
Obvious deviations from normality assumption

but …

Which of these panels looks the most different?

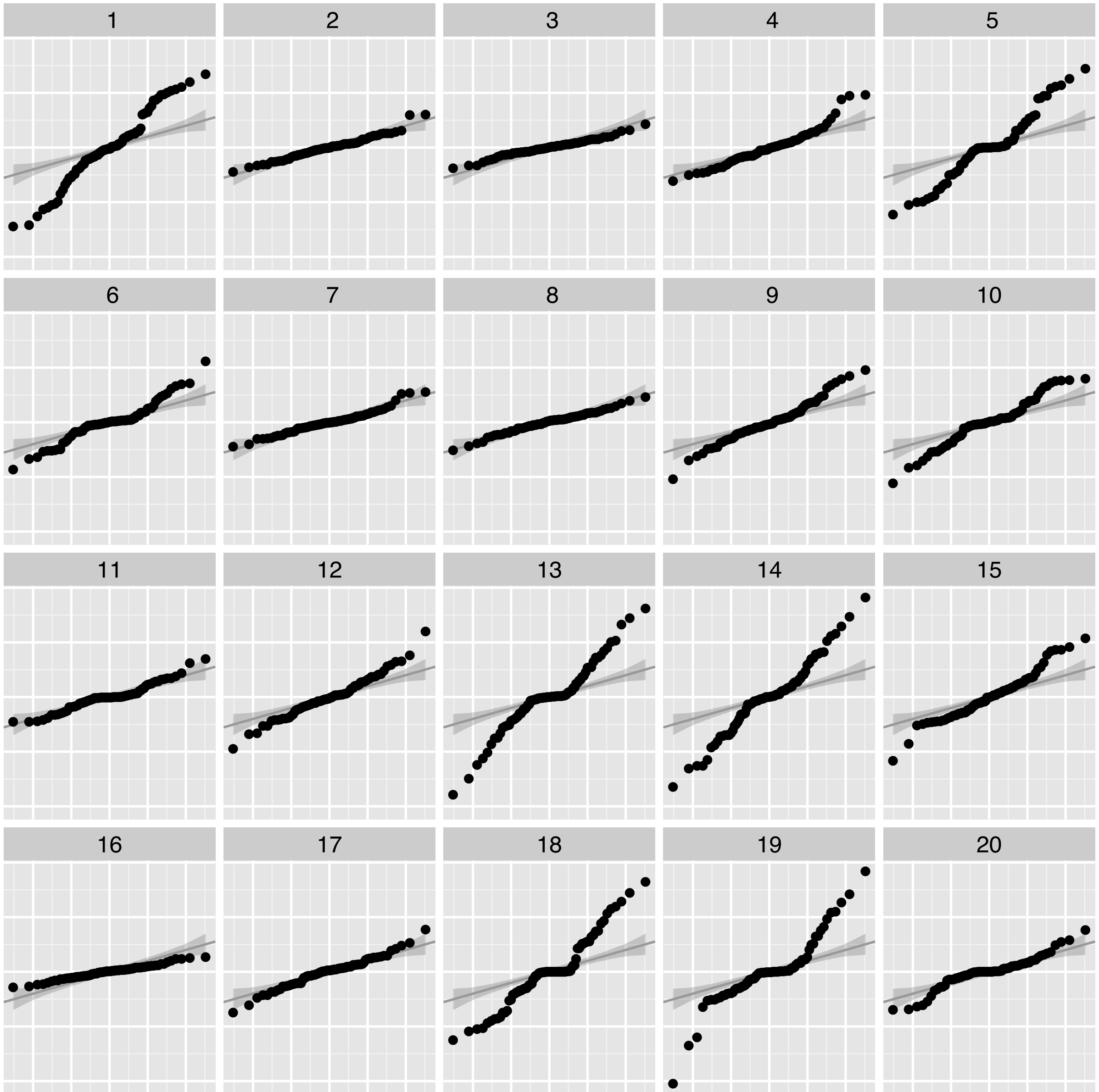


Which of these panels looks the most different?



data is in panel #10

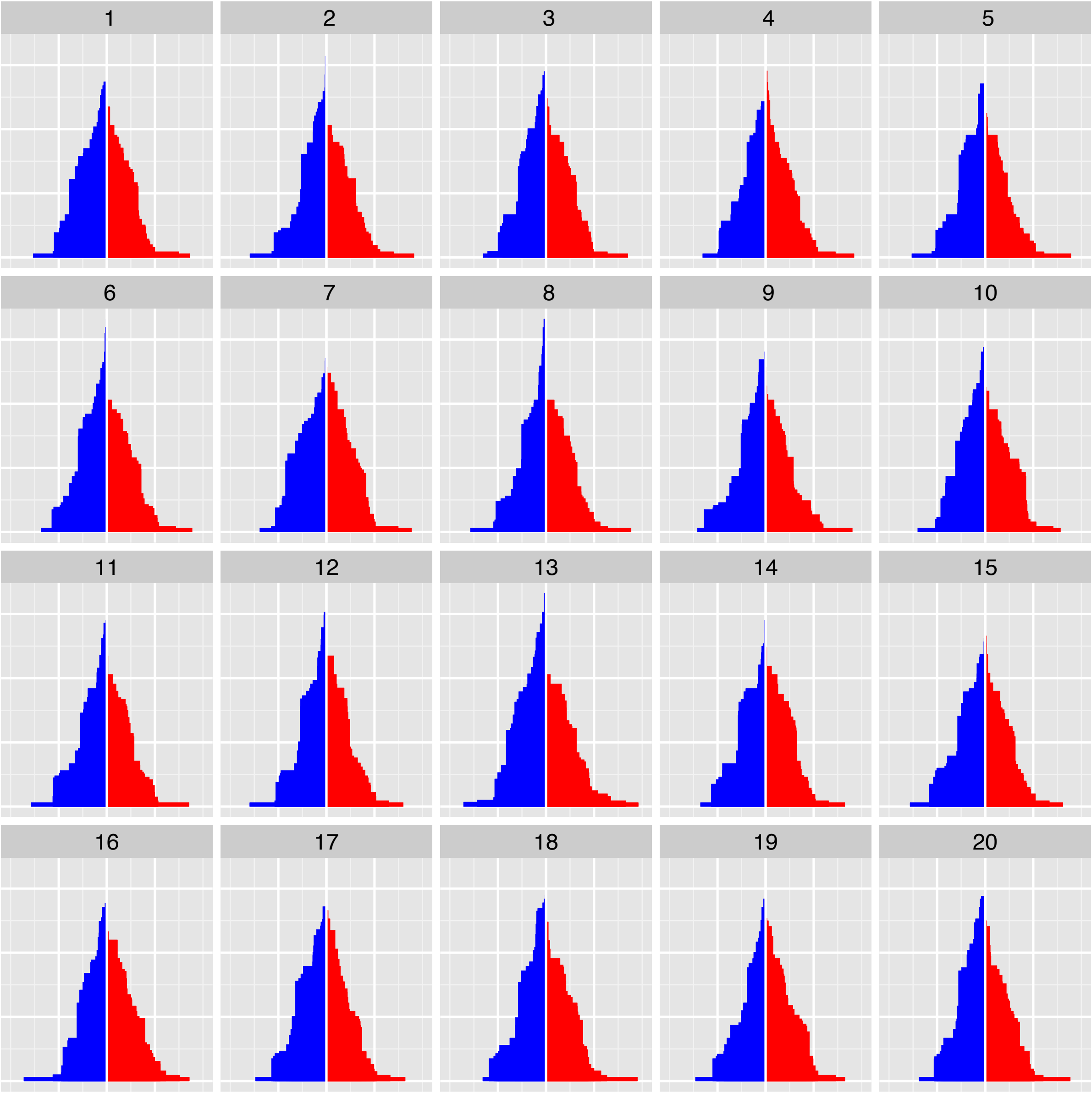
Which of these panels looks the most different?



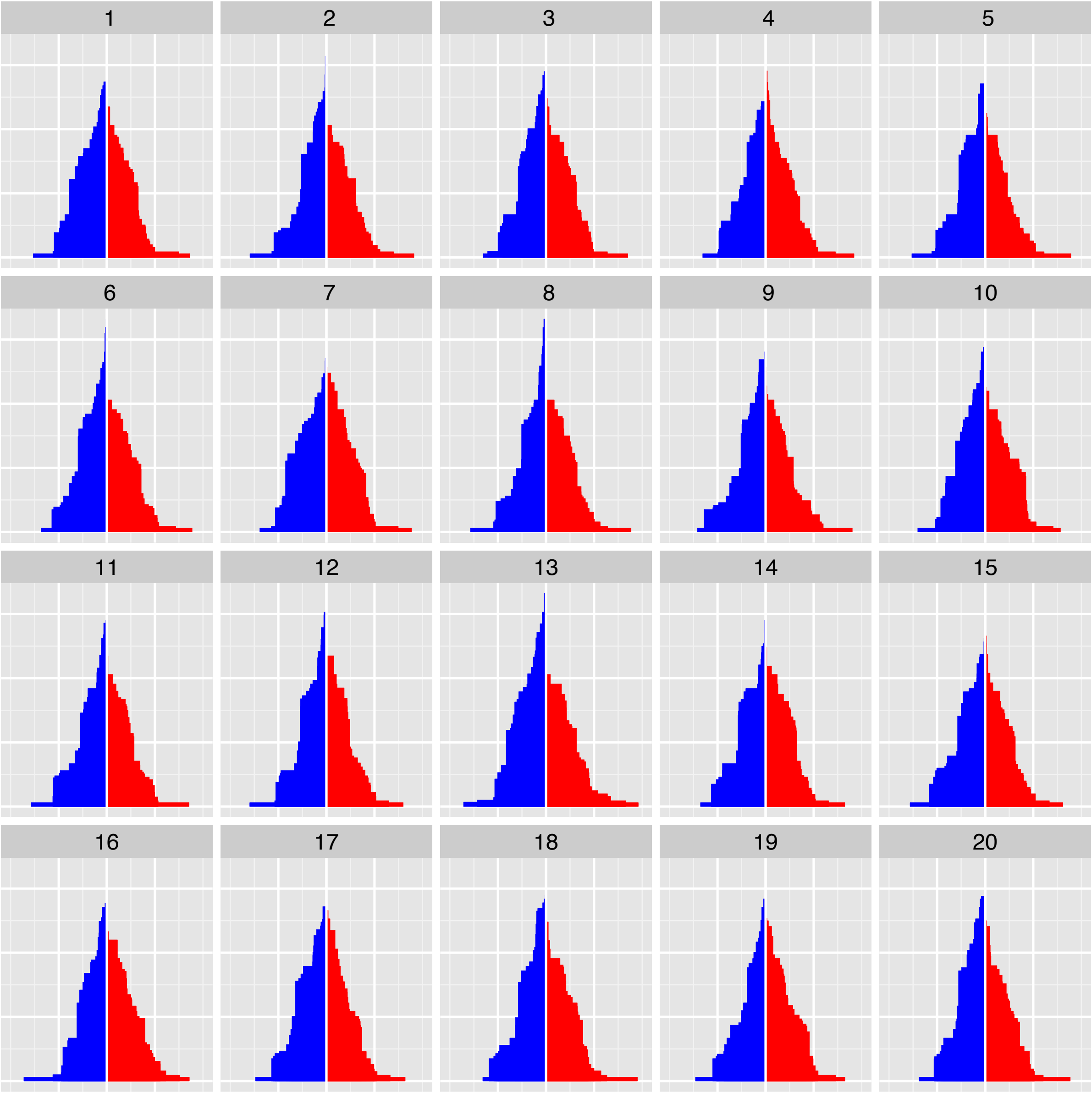
data is in panel #10

0/68 participants identified #10 as the most different

Which of these panels looks the most different?

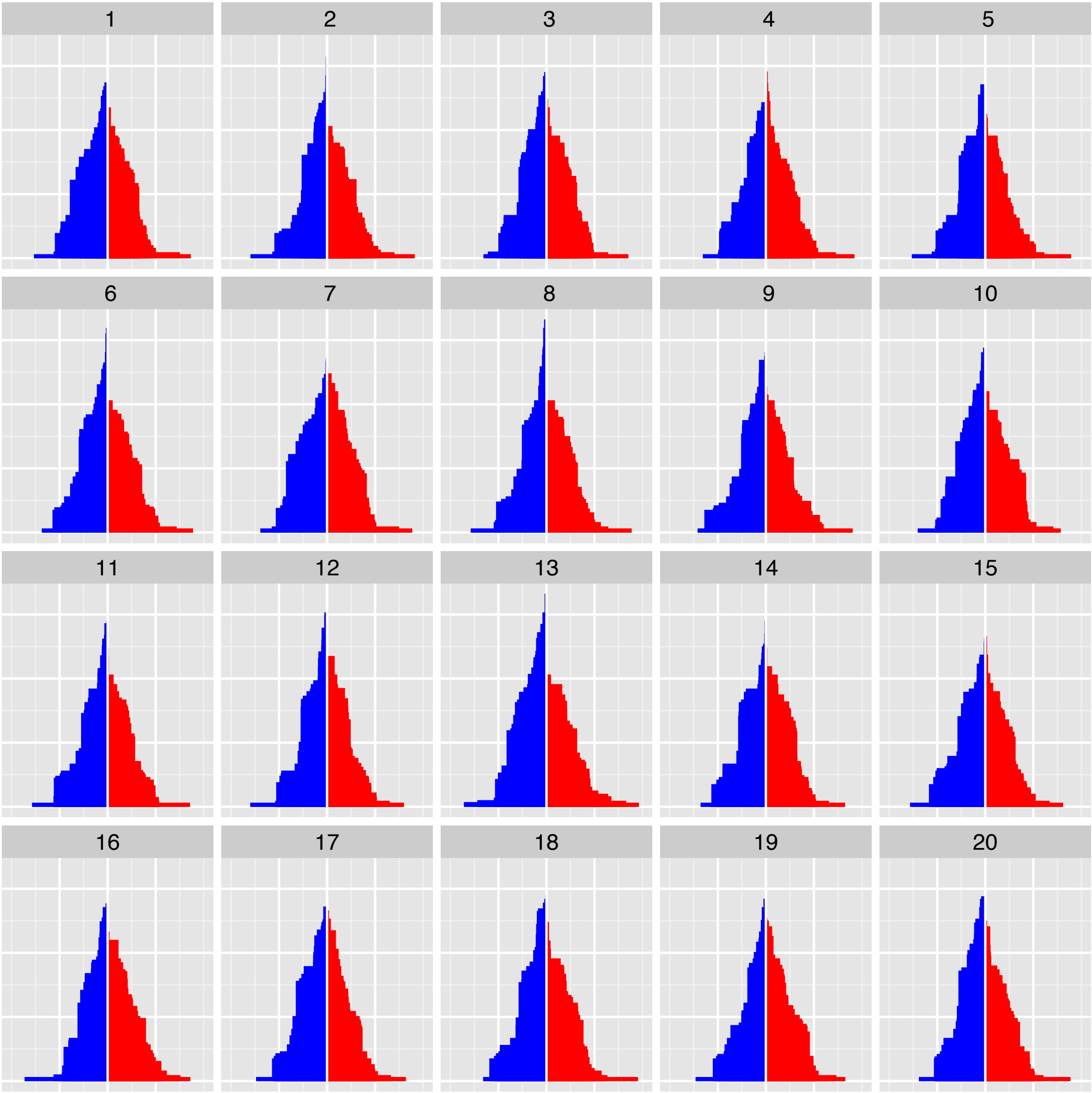


Which of these panels looks the most different?



data is in panel #13

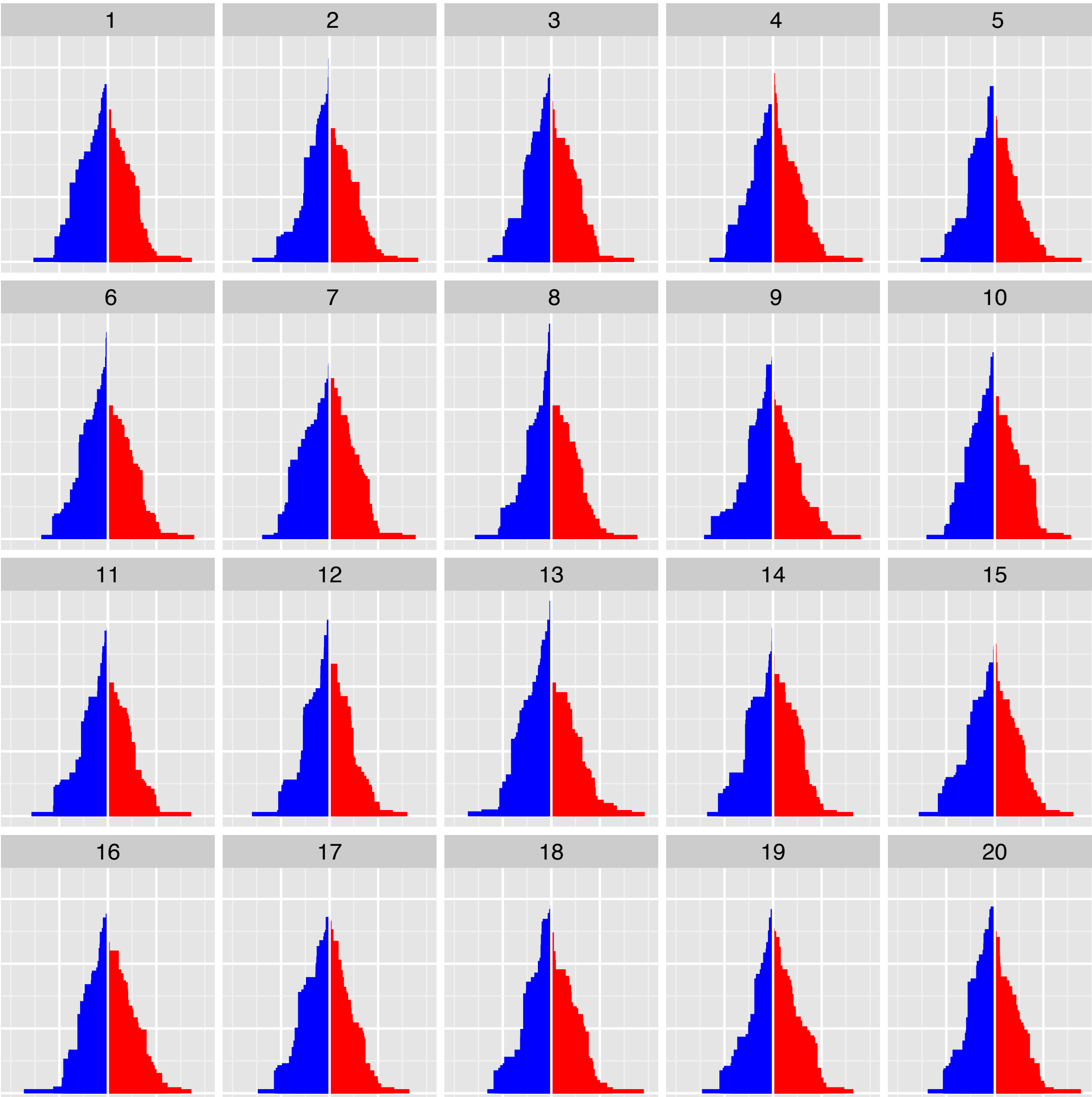
Which of these panels looks the most different?



data is in panel #13

12/72 participants identified #13 as the most different

Which of these panels looks the most different?



What is the p-value of this finding?

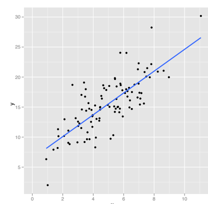
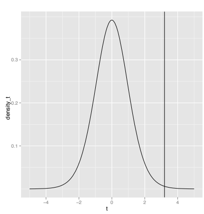
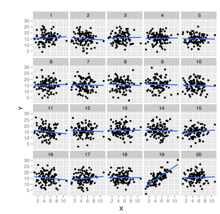
data is in panel #13

12/72 participants identified #13 as the most different

Back up:

- Lineup protocol in general
- Construction of Lineup in this example

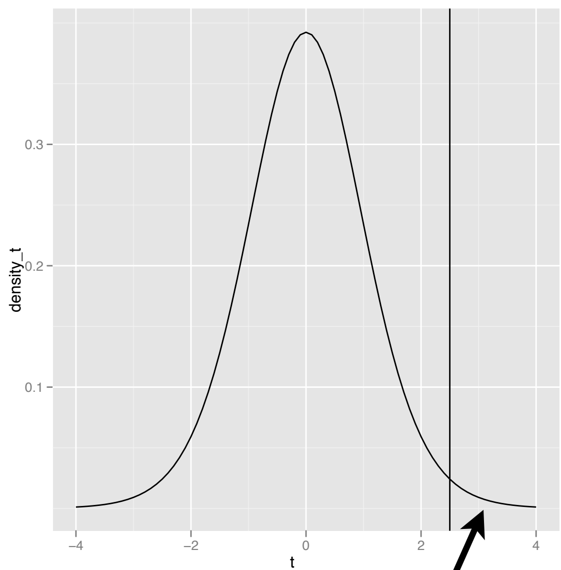
Graphical vs Classical

	Mathematical Inference	Visual Inference
Hypothesis	$H_0: \beta = 0$ vs $H_1: \beta \neq 0$	$H_0: \beta = 0$ vs $H_1: \beta \neq 0$
	$\downarrow$	$\downarrow$
Test statistic	$T(y) = \frac{\hat{\beta}}{se(\hat{\beta})}$	$T(y) =$ 
	$\downarrow$	$\downarrow$
Null Distribution	$f_{T(y)}(t);$ 	$f_{T(y)}(t);$ 
	$\downarrow$	$\downarrow$
Reject H_0 if	observed T is extreme	observed T is identifiable

Test

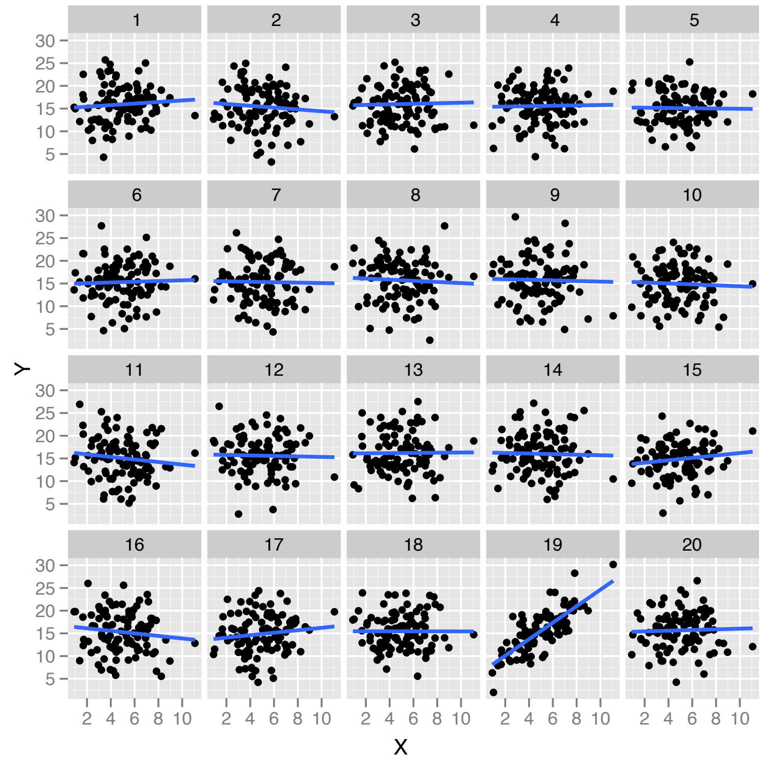
Compare test statistic to values generated consistently with the null distribution

Classical



reject null, if test statistic is here

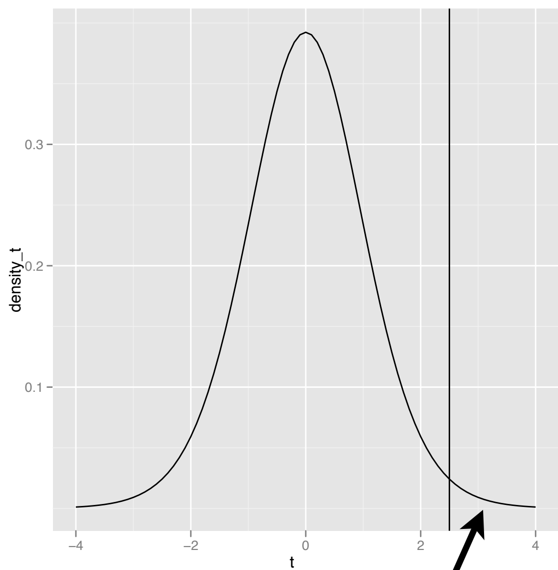
Visual



Test

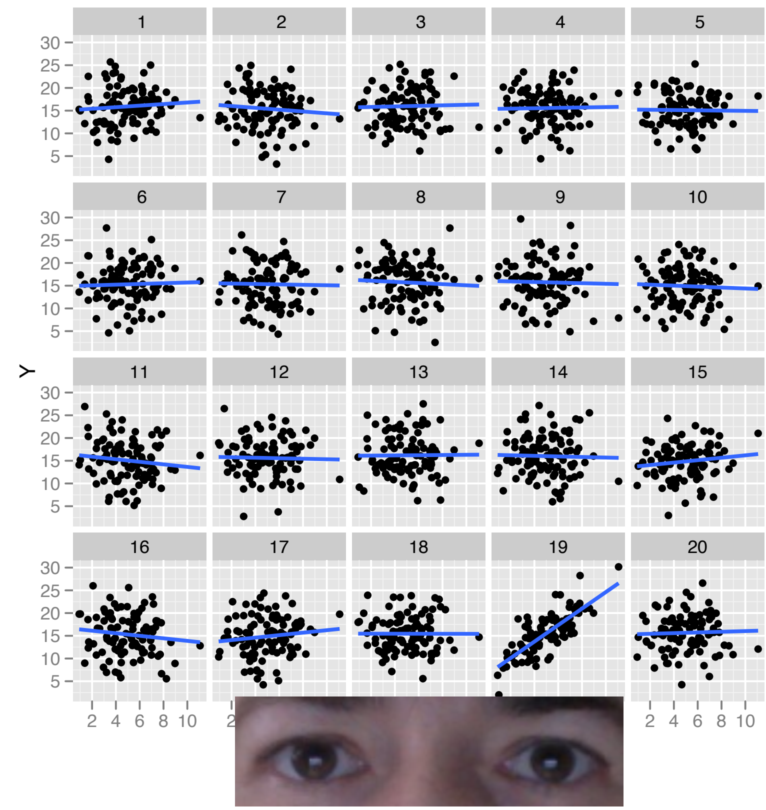
Compare test statistic to values generated consistently with the null distribution

Classical



reject null, if test statistic is here

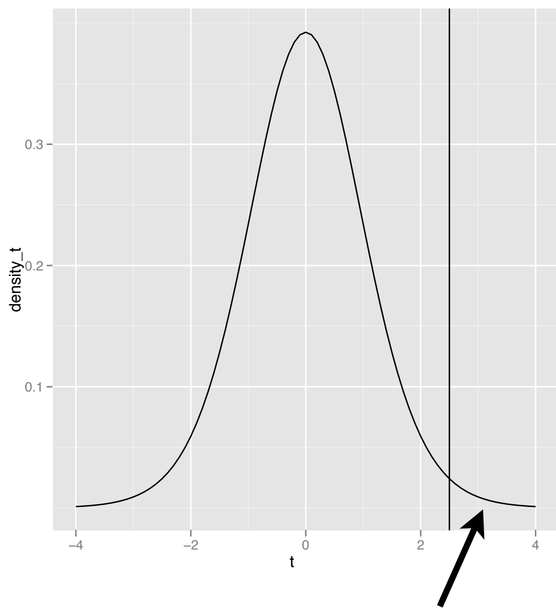
Visual



Test

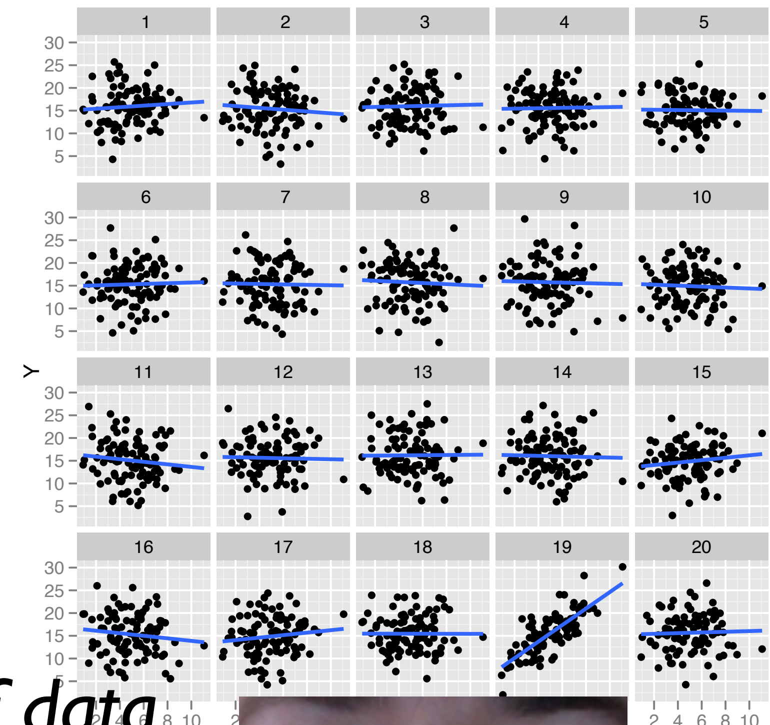
Compare test statistic to values generated consistently with the null distribution

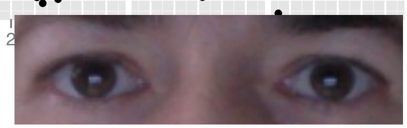
Classical



reject null, if test statistic is here

Visual



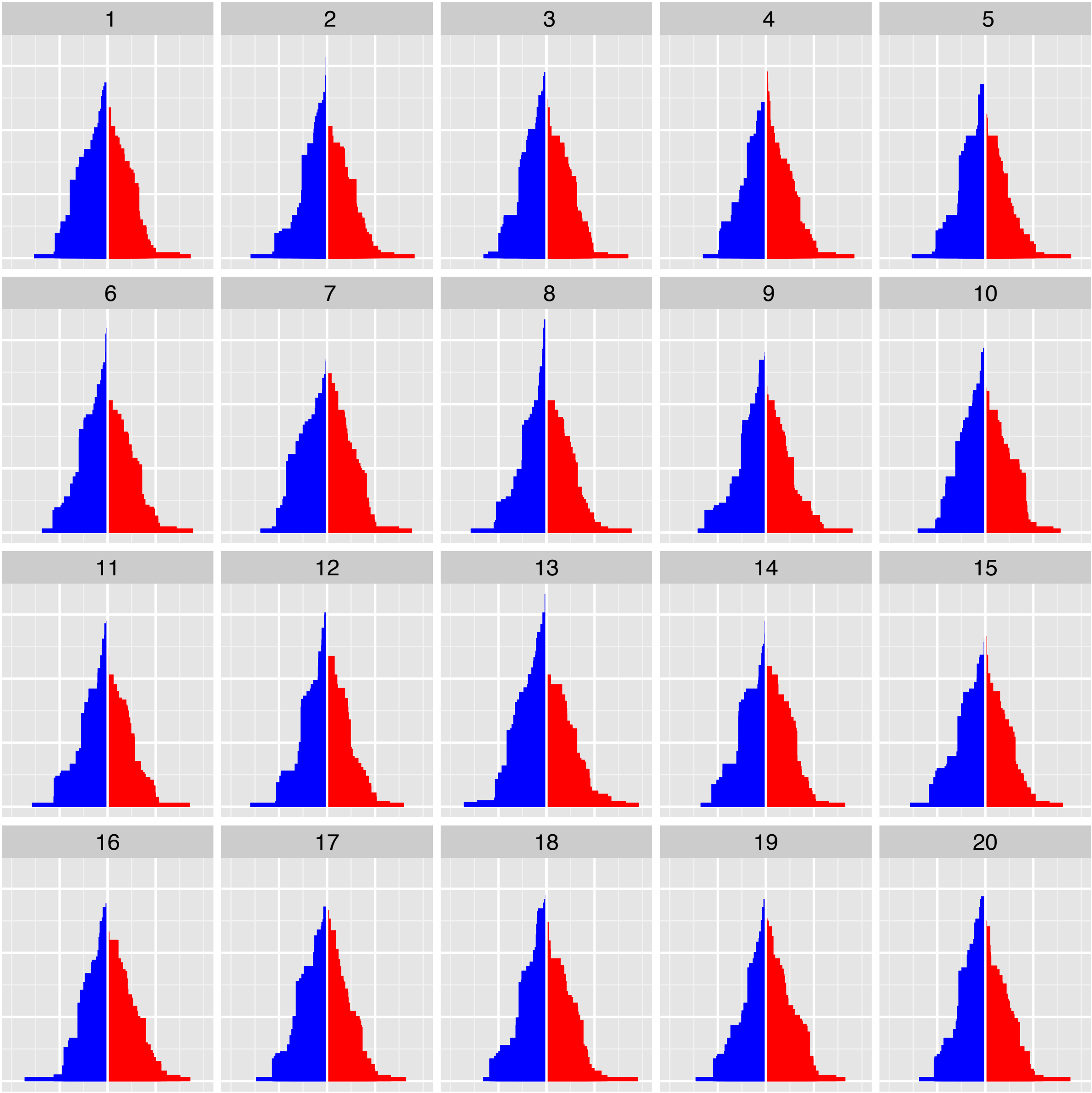


reject null, if data plot is 'identifiable'

Visual p-values

- Assume K independent observers evaluate a lineup
- Let X denote the number of data identifications
- quantify visual p-value: $\Pr(X \geq x \mid H_0 \text{ true})$

Which of these panels looks the most different?



What is the p-value of this finding?

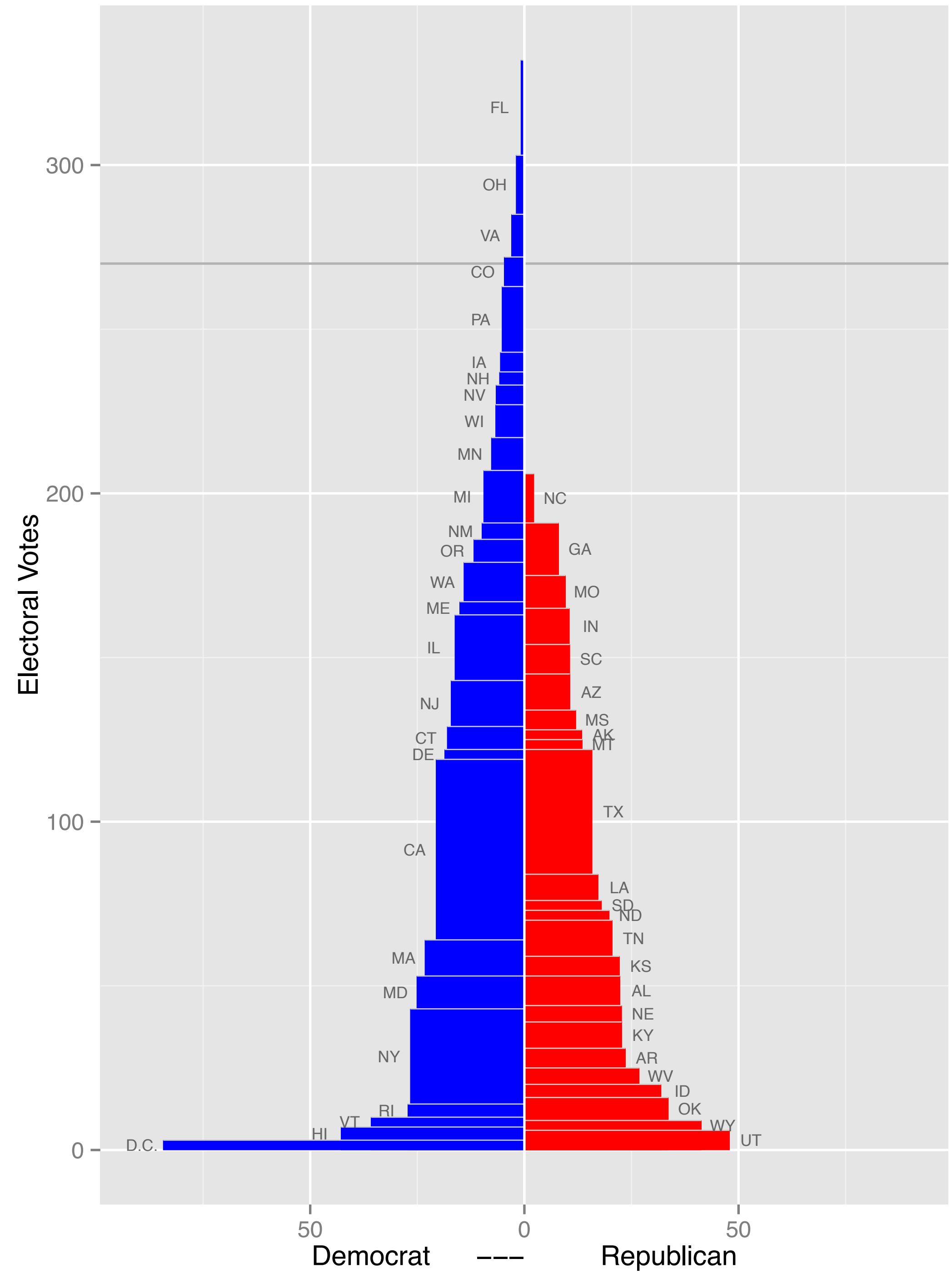
data is in panel #13

12/72 participants identified #13 as the most different

The Electoral Building

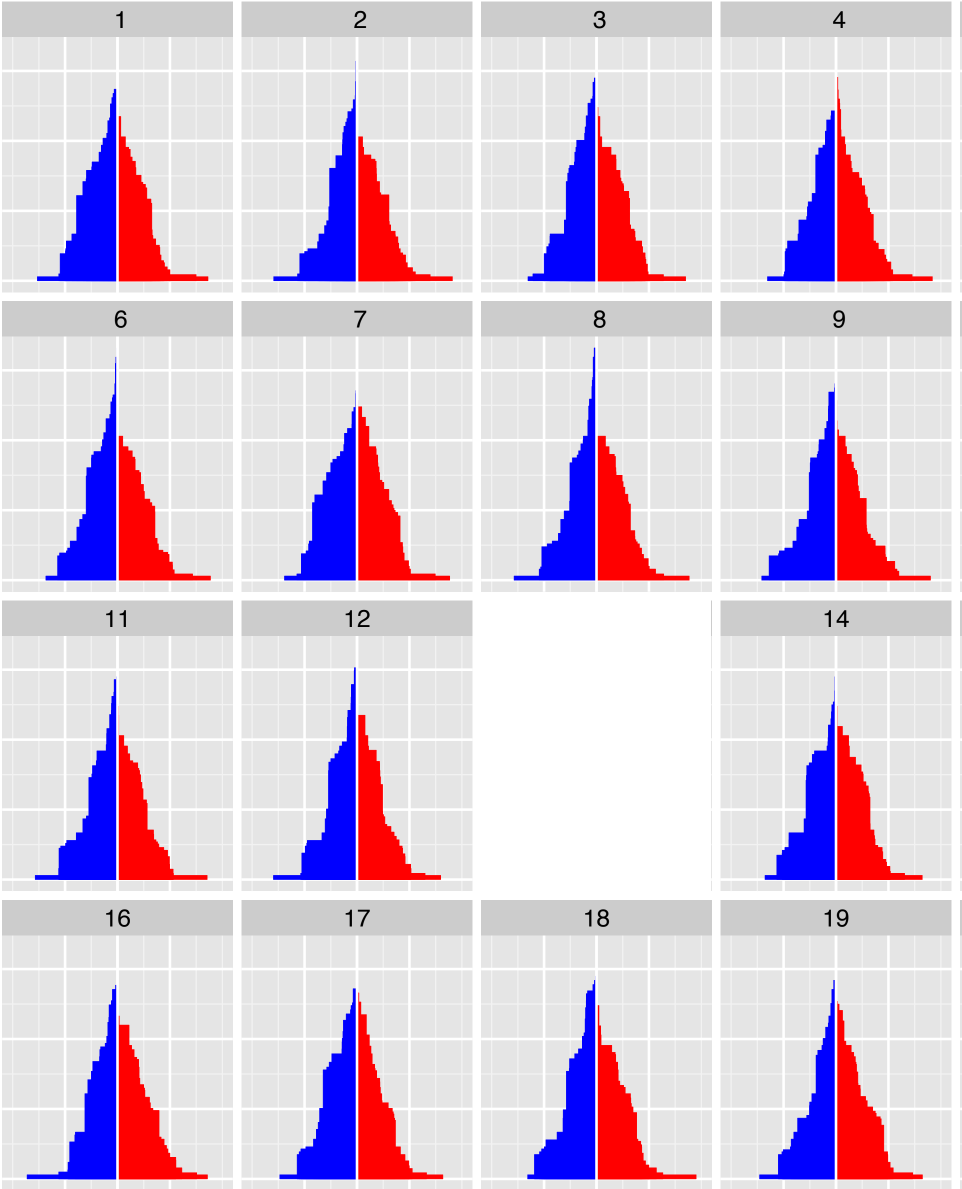
- result from the 2012 US election
- each state a rectangle:
  width: margin of majority party over minority
  height: #electoral votes

the test statistic



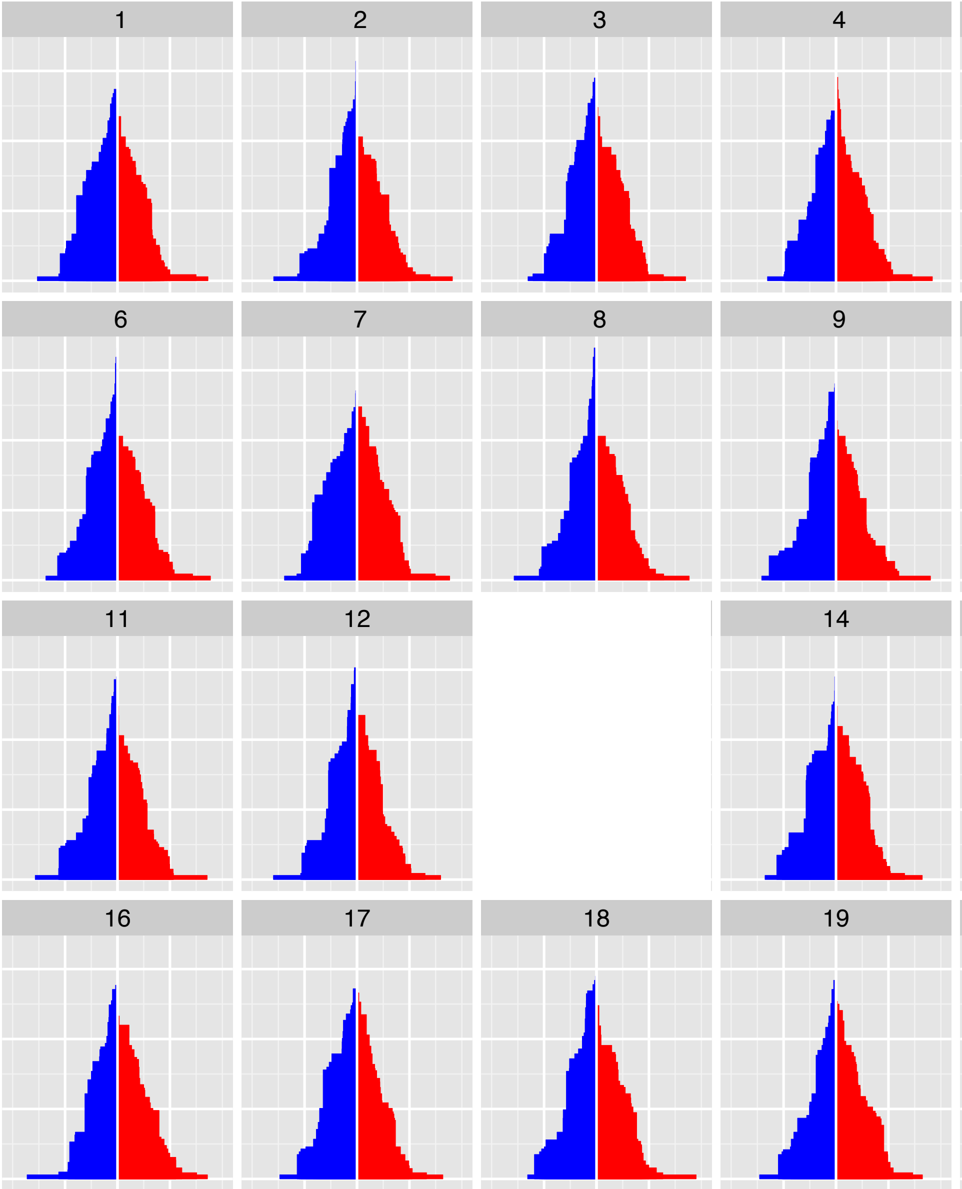
Null plots

- Null hypothesis: election outcome is consistent with polling results
- Each null plot consists of sample from a pollster's predictions



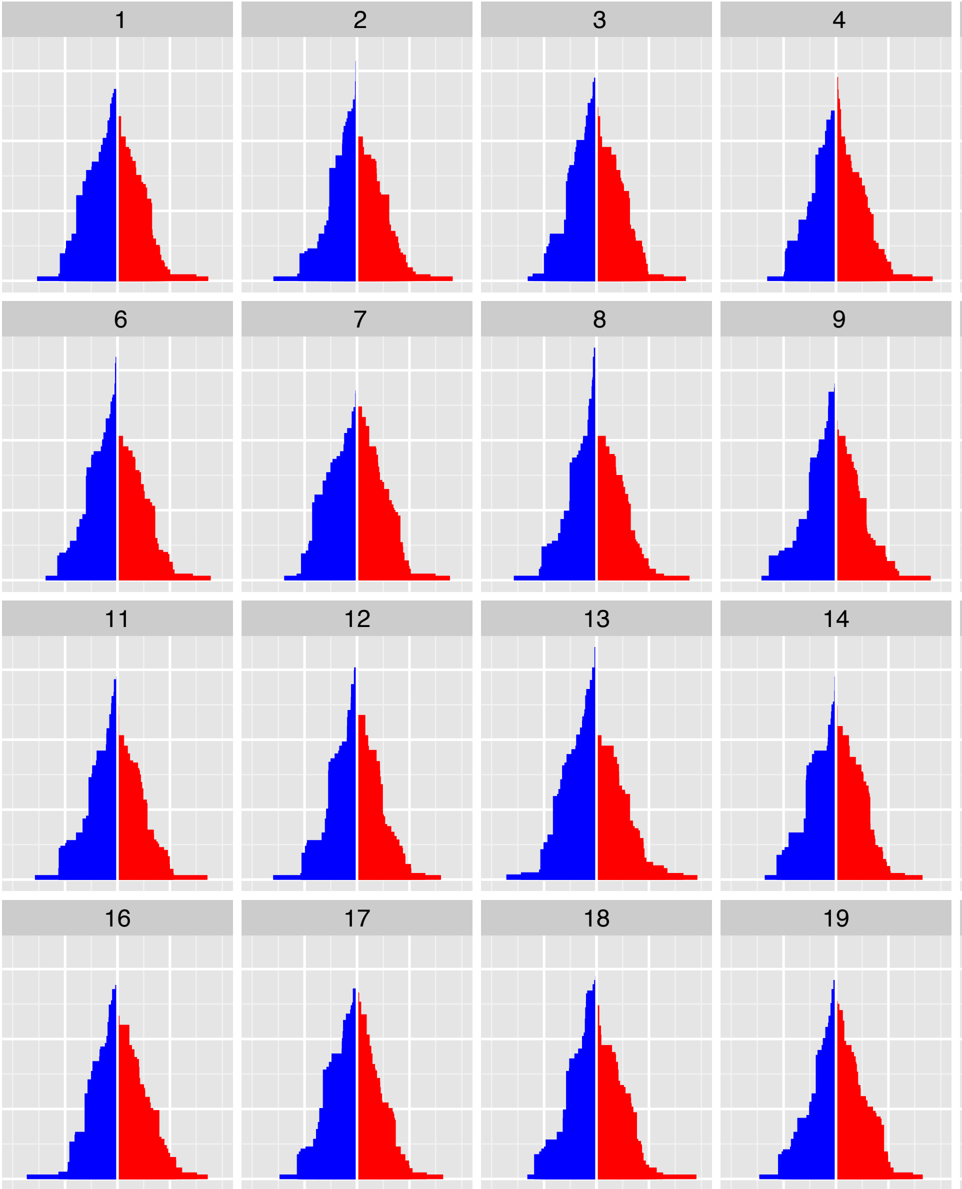
Lineup

- Data is randomly placed among the null plots
- If the data is indistinguishable from the null, the election results are consistent with the poll



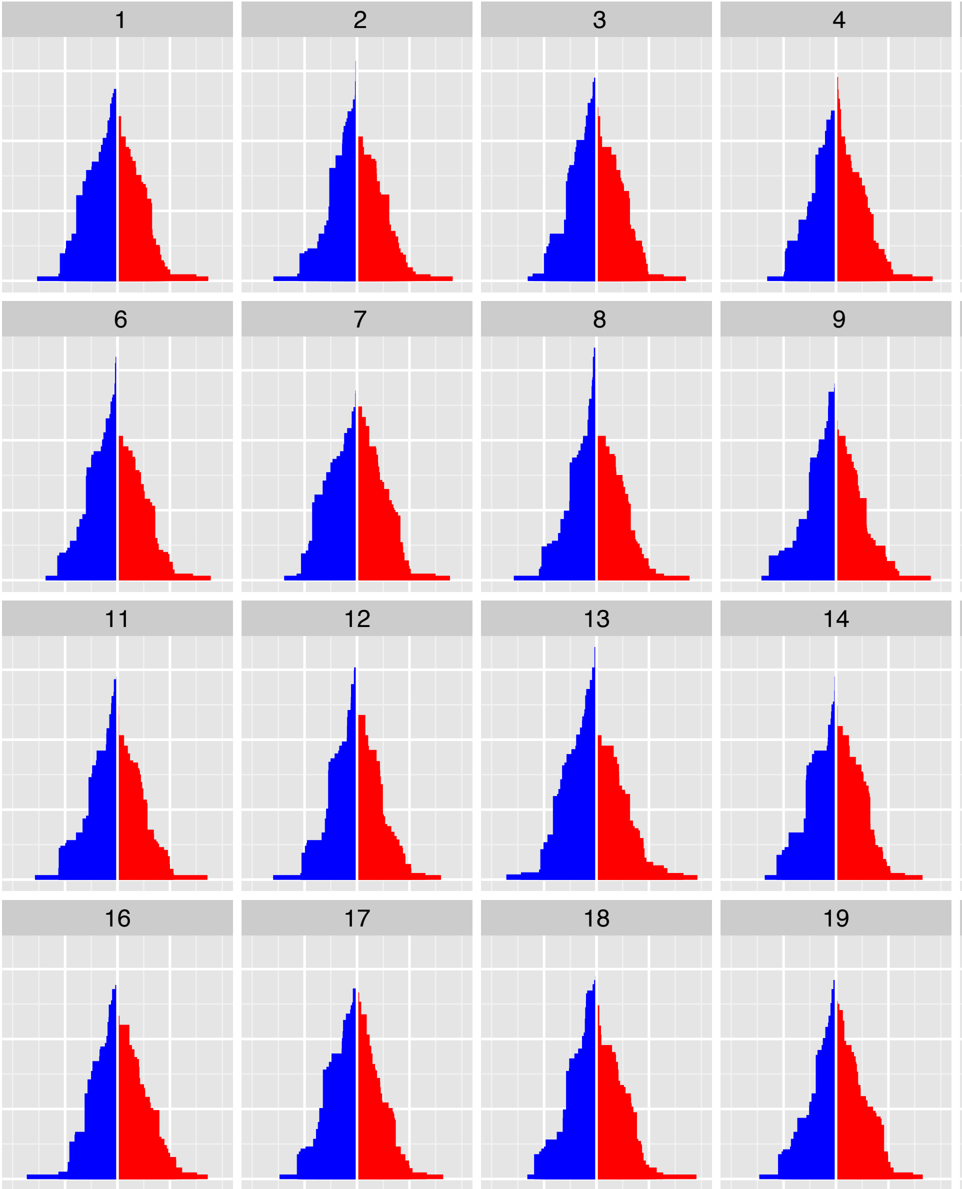
Lineup

- Data is randomly placed among the null plots
- If the data is indistinguishable from the null, the election results are consistent with the poll



Lineup

- Data is randomly placed among the null plots
- If the data is indistinguishable from the null, the election results are consistent with the poll



visual p-value: P(#data plot picks ≥ 12)

Data from lineup evaluation

- For lineup of size m we observe
  $X = (X_1, \ldots, X_m) \sim Mult_{p1, p2, \ldots, pm}$
- with $0 \leq p_i \leq 1$ and $\sum_i p_i = 1$
- w.lo.g. data plot in panel m,
  ie $X_m \sim Binom(K, p_m)$
  K independent evaluations
- What is distribution of X_m under null?

Data from lineup evaluation

- For lineup of size m we observe
  $X = (X_1, \ldots, X_m) \sim \text{Mult}_{p1, p2, \ldots, pm}$
- with $0 \leq p_i \leq 1$ and $\sum_i p_i = 1$
- w.lo.g. data plot in panel m,
  ie $X_m \sim \text{Binom}(K, p_m)$
  K independent evaluations
- What is distribution of X_m under null?

  if all plots were indistinguishable, we could assume $p_m = 1/m$

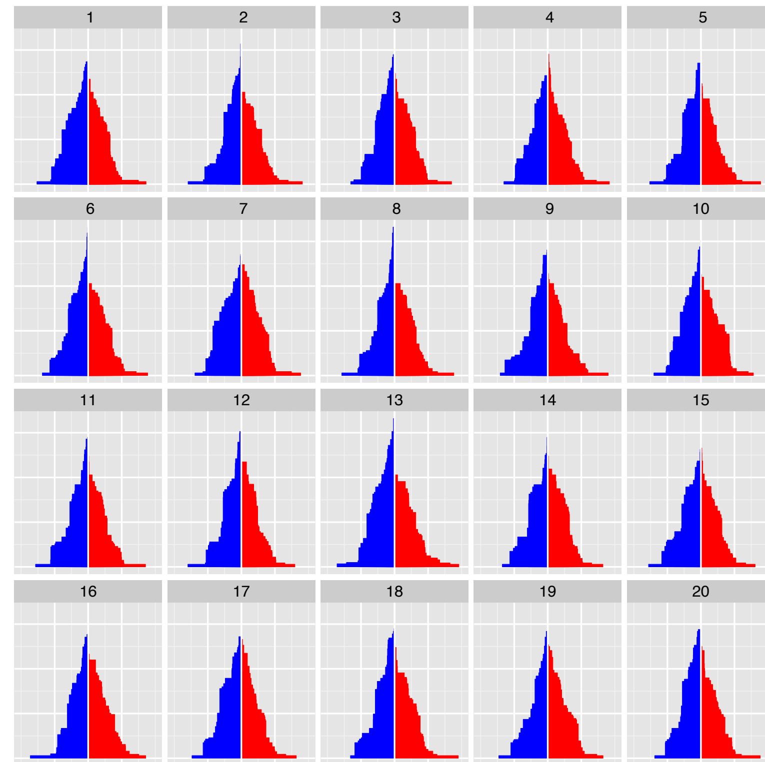
Data from lineup evaluation

- For lineup of size m we observe
  $X = (X_1, \ldots, X_m) \sim \text{Mult}_{p1, p2, \ldots, pm}$
- with $0 \leq p_i \leq 1$ and $\sum_i p_i = 1$
- w.lo.g. data plot in panel m,
  ie $X_m \sim \text{Binom}(K, p_m)$
  K independent evaluations
- What is distribution of X_m under null?

  if all plots were indistinguishable, we could assume $p_m = 1/m$

Evaluating lineup evaluations

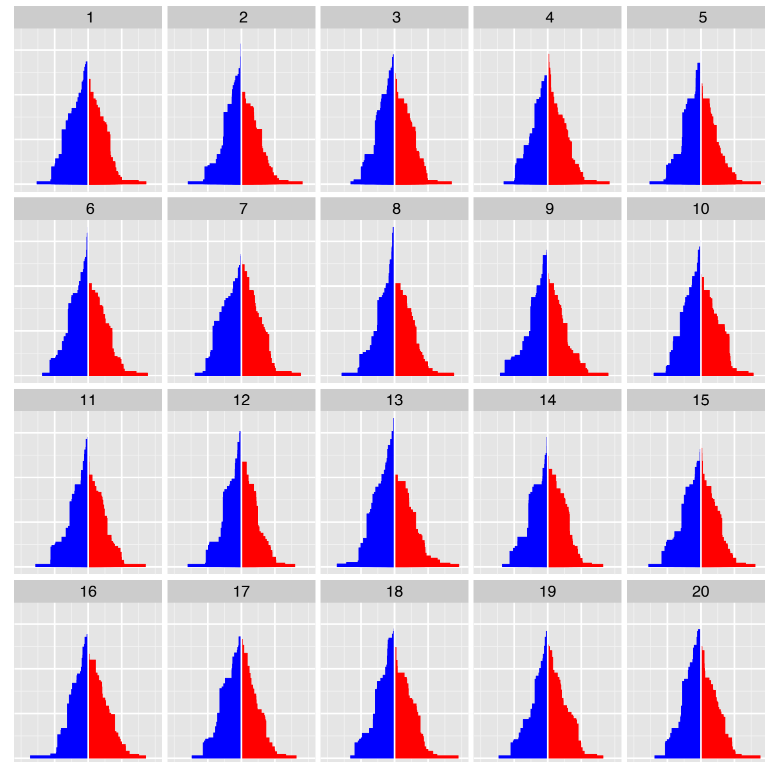
- Assuming X ~ Binom(72, 1/20)
- p-value for 12 data picks is $P(X \geq 12) = 0.00023$



Evaluating lineup evaluations

- Assuming X ~ Binom(72, 1/20)
- p-value for 12 data picks is $P(X \geq 12) = 0.00023$

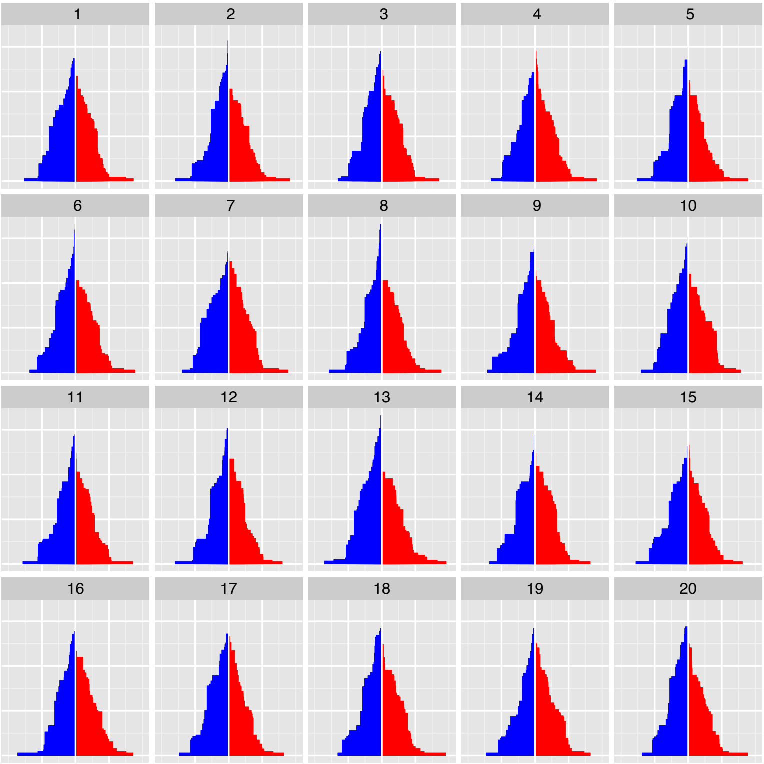
fails the sniff test!



Evaluating lineup evaluations

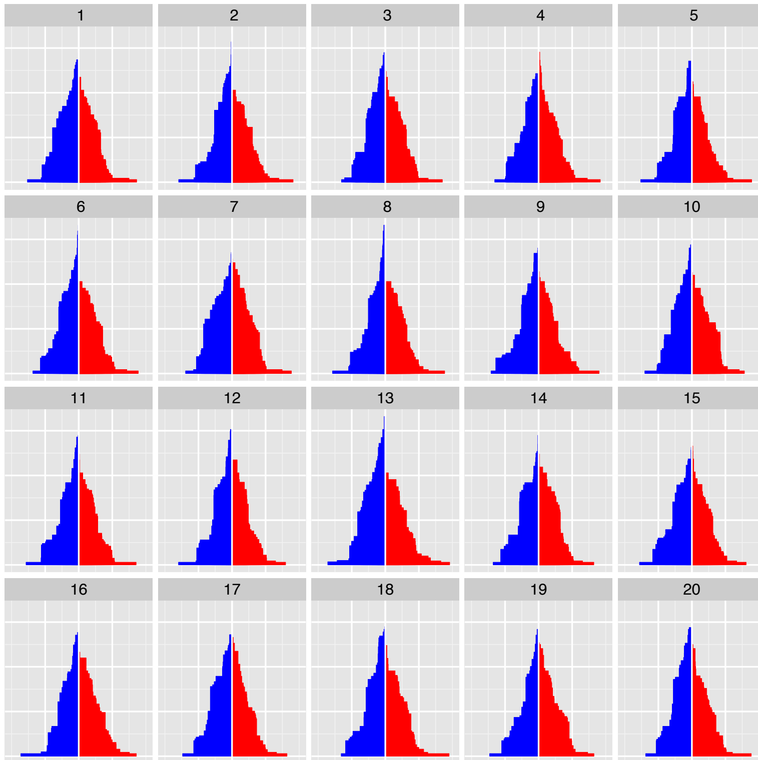
- Assuming X ~ Binom(72, 1/20)
- p-value for 12 data picks is $P(X \geq 12) = 0.00023$

fails the sniff test!



Problem: if *all plots were indistinguishable*, we could assume $p_m = 1/m$ (and all $p_i = 1/m$)

Evaluating lineup evaluations



- Assuming X ~ Binom(72, 1/20)
- p-value for 12 data picks is $P(X \geq 12) = 0.00023$

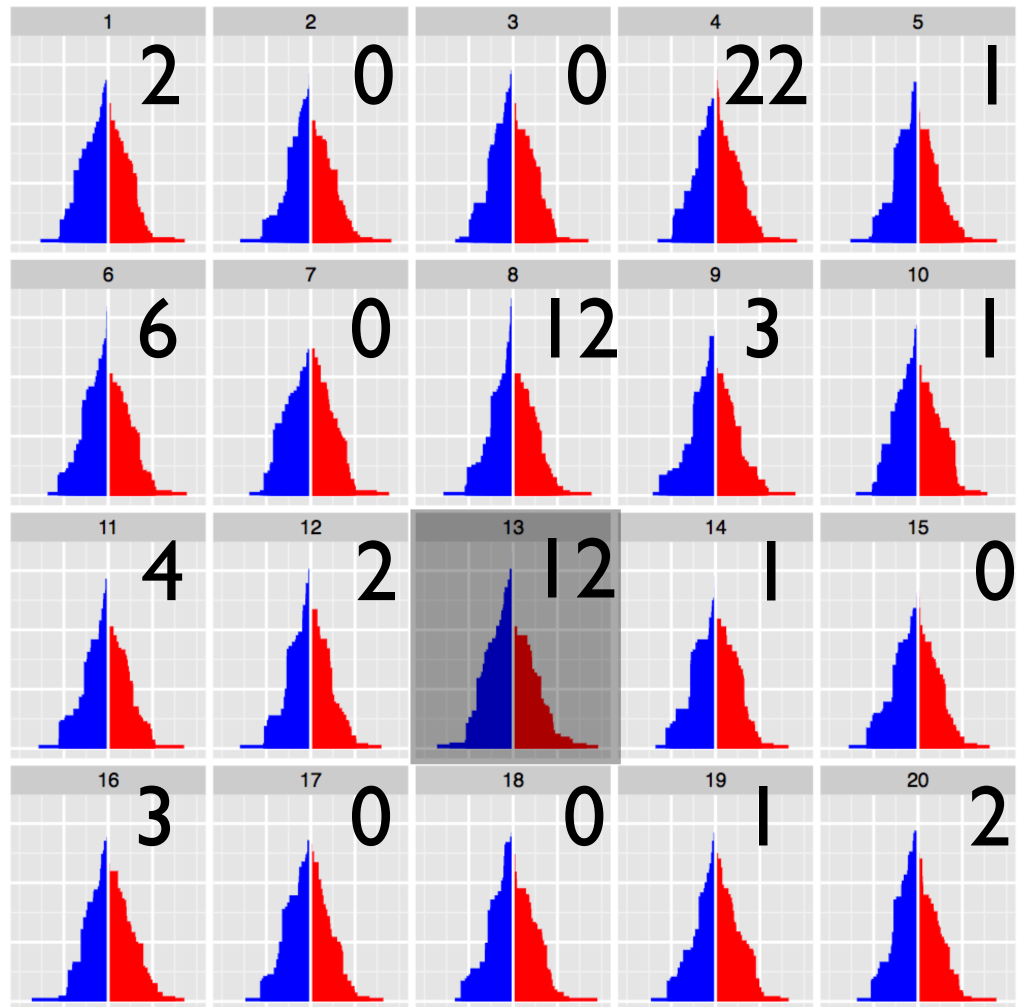
fails the sniff test!

Problem: if *all plots were indistinguishable*, we could assume $p_m = 1/m$ (and all $p_i = 1/m$)

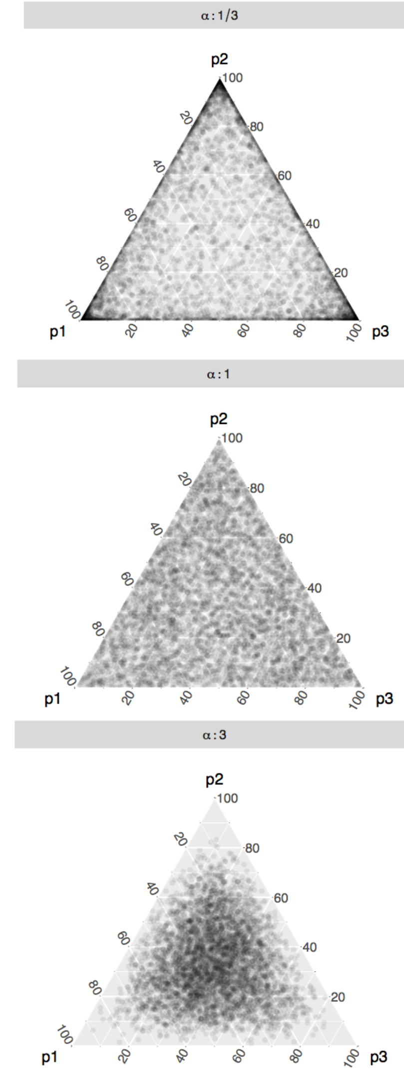
Generally: p_m depends on $p_1, \ldots, p_{m-1}$, varies with lineup

Null Distribution of p

- Two other plots were selected at least as often as the data plot
- Distribution of null plot picks far from uniform



Null Distribution of p



- p_i is probability to pick panel i
- Assume that under the null, all panels have the same distribution:
  $p = (p_1, \ldots, p_m) \sim \text{Dirichlet}(\alpha), \alpha > 0$
  a flat Dirichlet distribution
- Estimate rate α from observed $(p_1, \ldots, p_{m-1})'$ where $(p_1, \ldots, p_{m-1})'$ is rescaled without data plot

Distribution of $(p_1, \ldots, p_{m-1})'$

- flat Dirichlet(α) for $(p_1, \ldots, p_{m-1})'$ seems reasonable
- no obvious preference for location

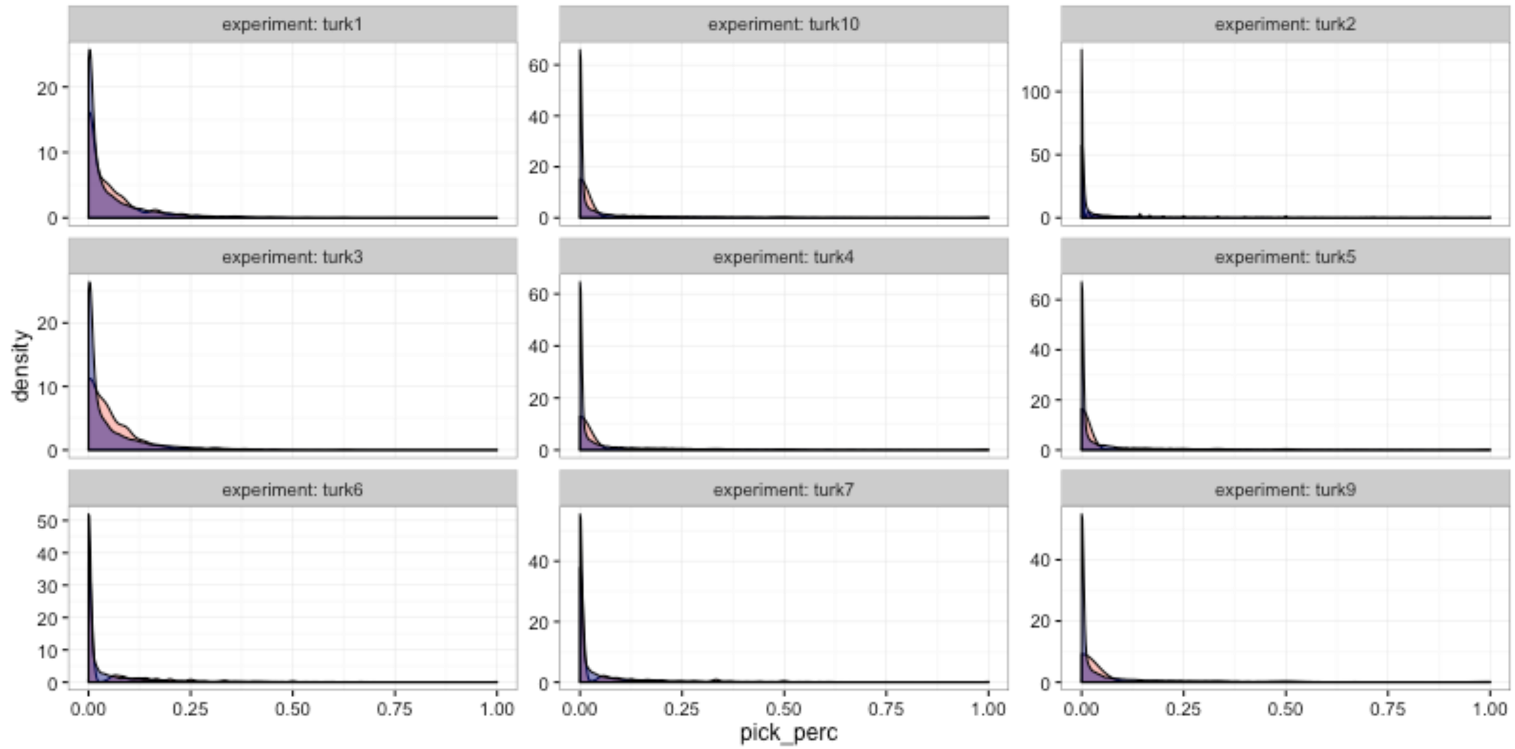
Distribution of $(p_1, \ldots, p_{m-1})'$

- flat Dirichlet(α) for $(p_1, \ldots, p_{m-1})'$ seems reasonable
- no obvious preference for location

Dirichlet distributions estimated
for each of nine different
experiments

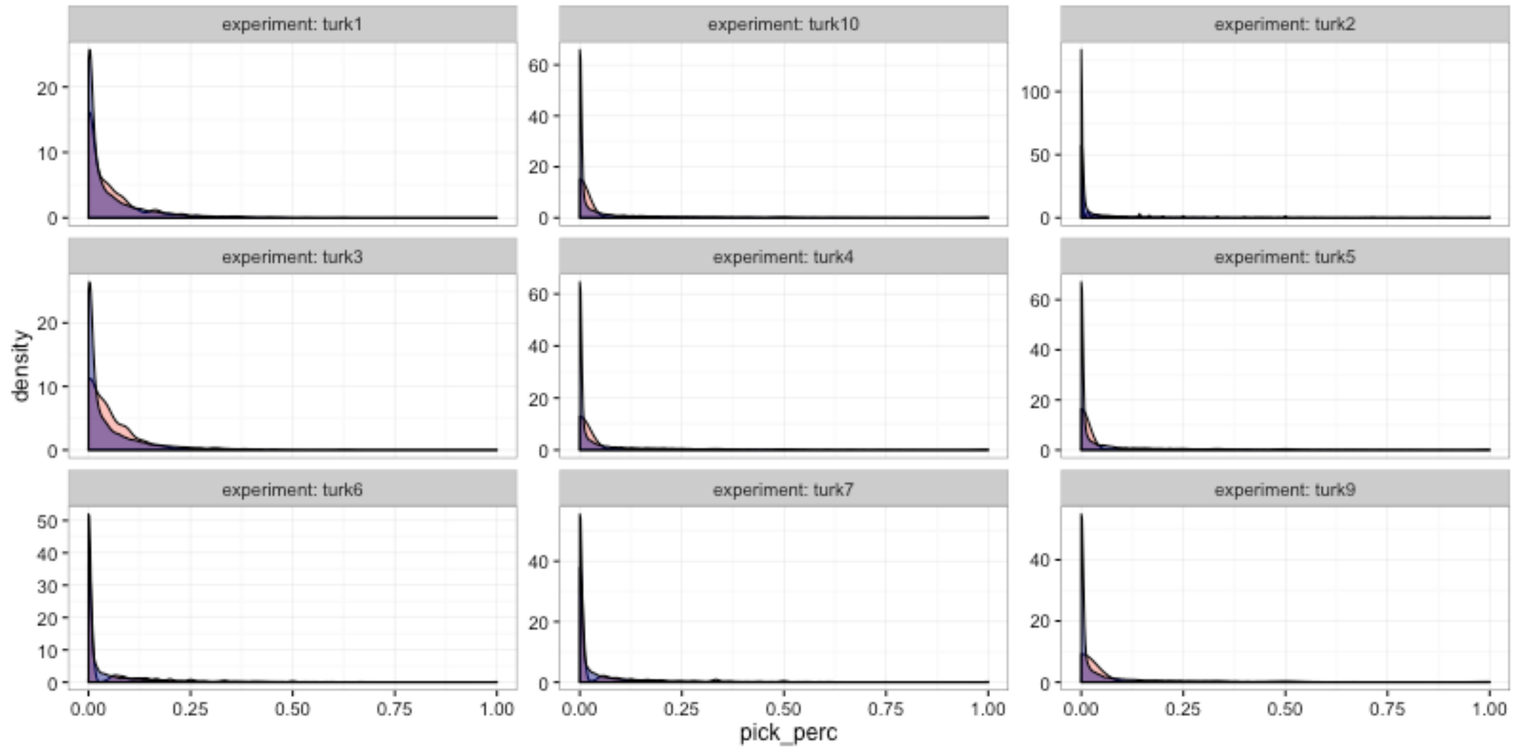
Distribution of $(p_1, \ldots, p_{m-1})'$

- flat Dirichlet(α) for $(p_1, \ldots, p_{m-1})'$ seems reasonable
- no obvious preference for location



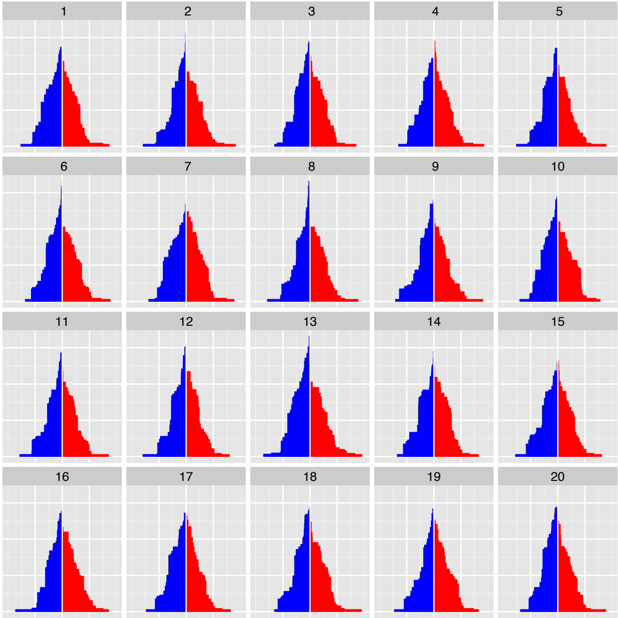
Distribution of $(p_1, \ldots, p_{m-1})'$

- flat Dirichlet(α) for $(p_1, \ldots, p_{m-1})'$ seems reasonable
- no obvious preference for location



turk1	turk2	turk3	turk4	turk5	turk6	turk7	turk9	turk10
0.34	0.14	0.33	0.13	0.13	0.15	0.15	0.14	0.12

visual p-value

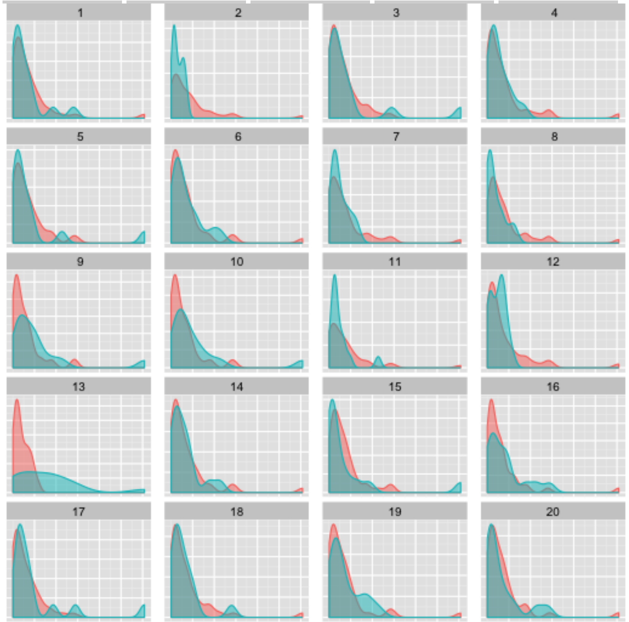


12/72 participants identified #13 as the most different

- p-value based on Binom(72, 1/20)
  $P(X \geq 12) = 0.00023$

- p-value based on Dirichlet approach:
  $P(X \geq 12) = 0.11396$

visual p-value



- p-value based on Binom(23, 1/20)
  $P(X \geq 20) \leq 0.00001$

20/23 participants identified #13 as the most different

- p-value based on Dirichlet approach:
  $P(X \geq 20) = 0.001842$

Dirichlet distributions for null

- seems to work in practice - theoretical densities and observed frequencies of picking null plots match
- α gives a rough estimate of the spread of null distribution/difficulty of a lineup (without regarding : small α = small number of null plots attract picks)
- Weirdly, strong signal in data plot makes estimating α harder: Rorschach for α

Conclusions

- Use lineup scenario to get valid p-values for visual findings
- useful in situations where conventional methods break down
- lineups allow us to ask for 'why' … insight to visual reasoning of participants