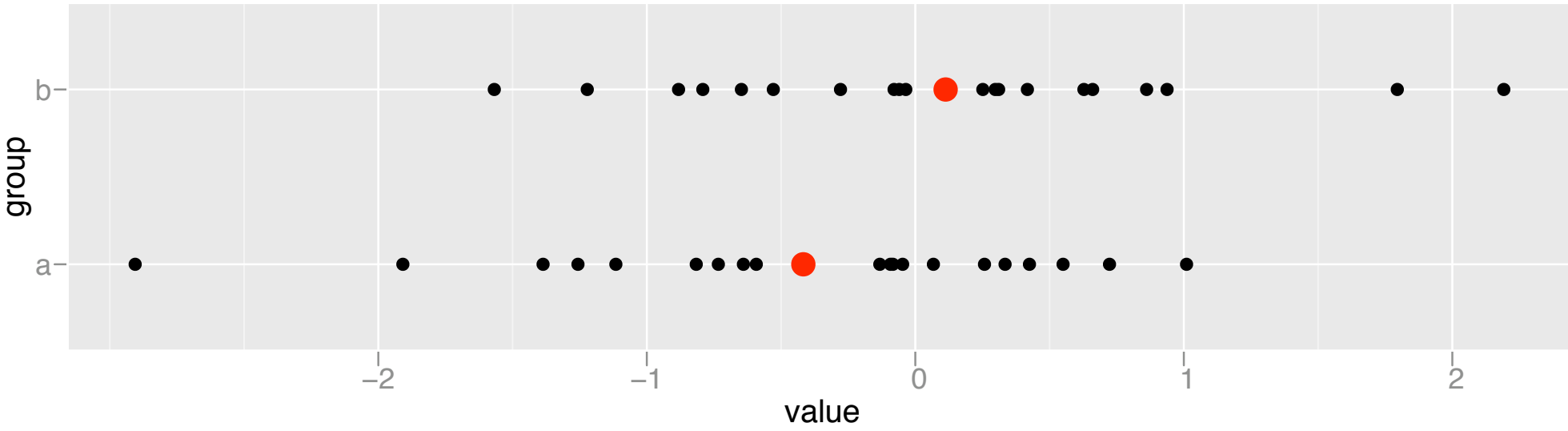
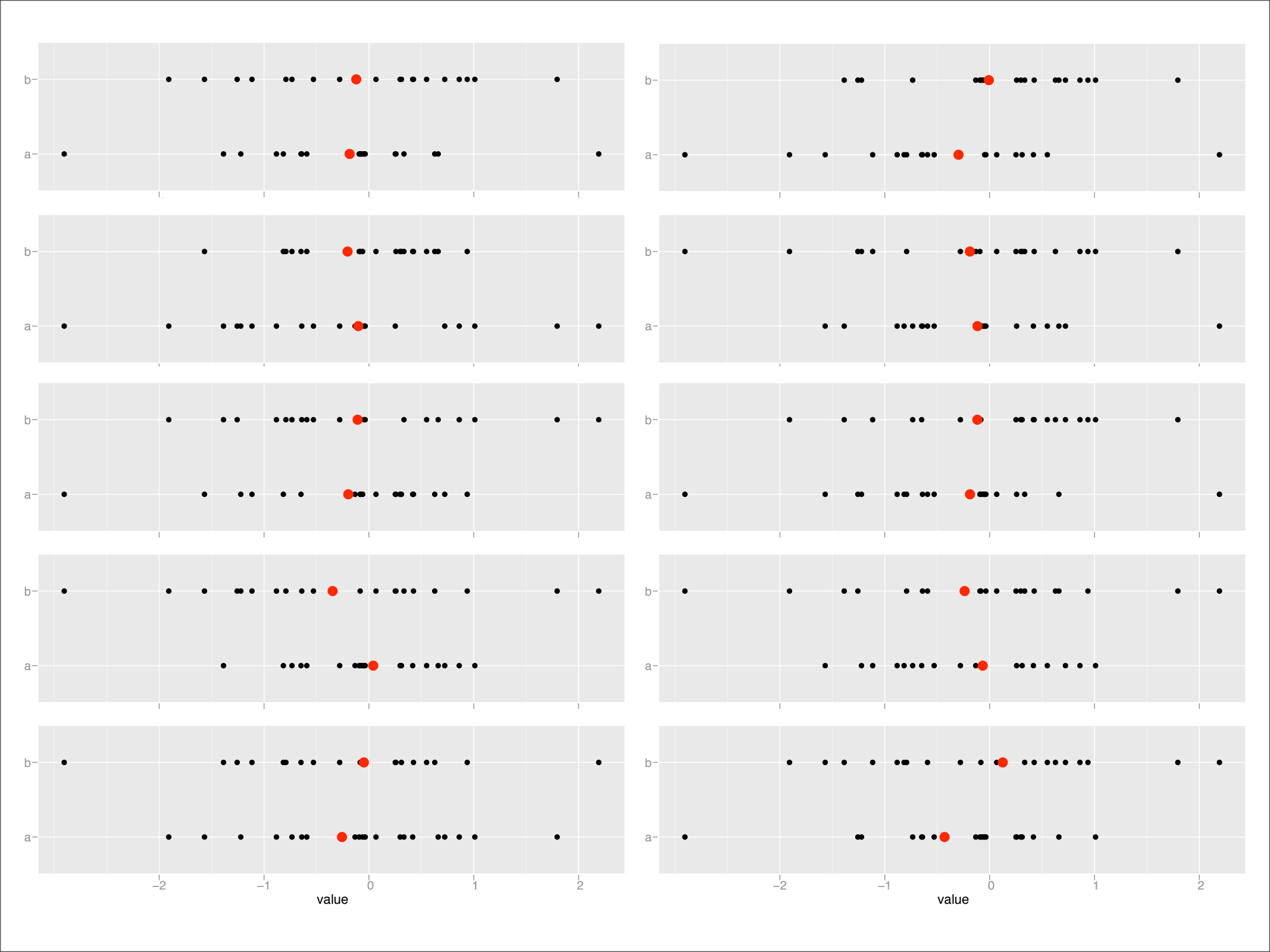
Inference

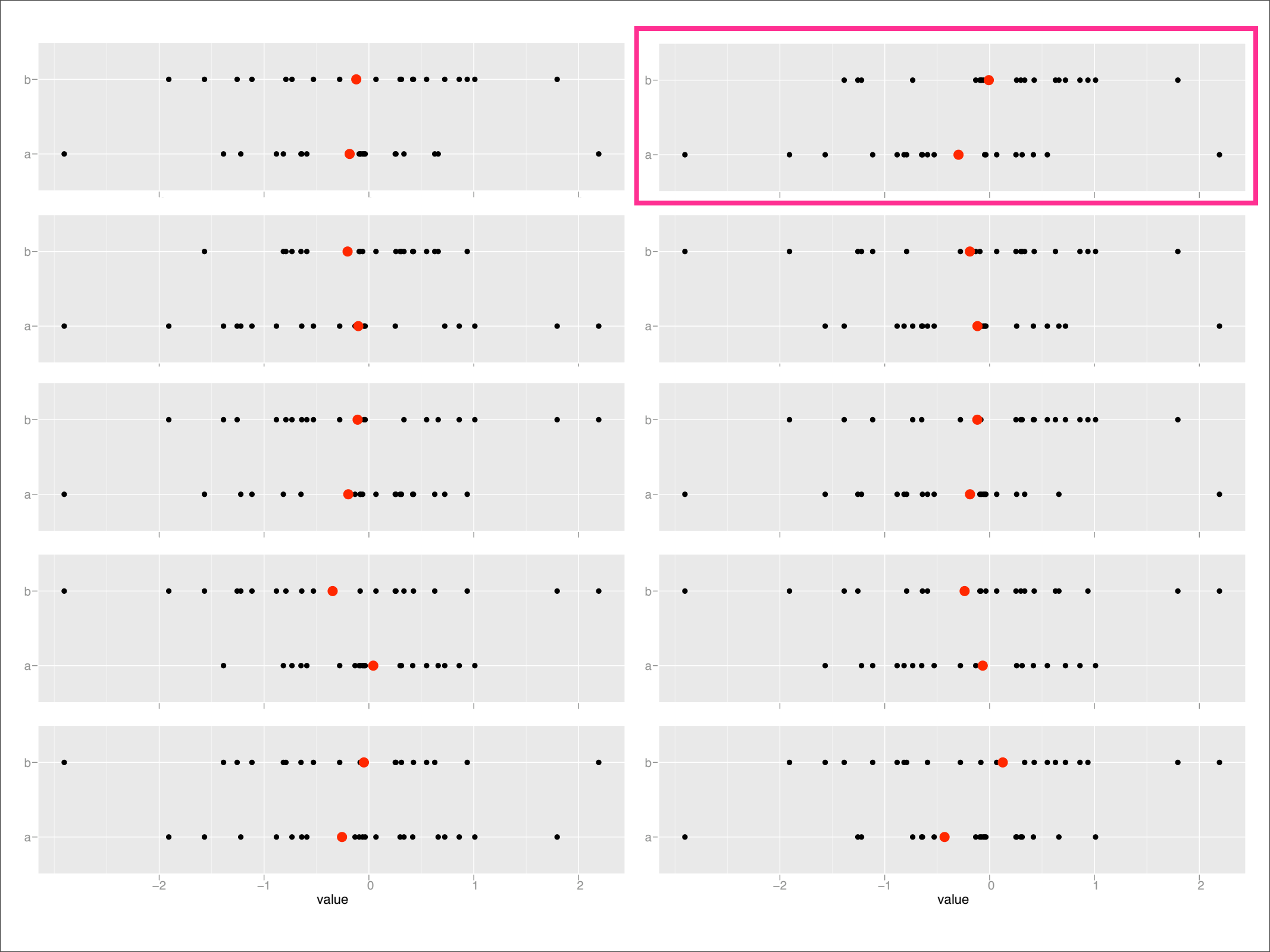
Revealing, informative plots often provoke the question “Is what we see really there?”

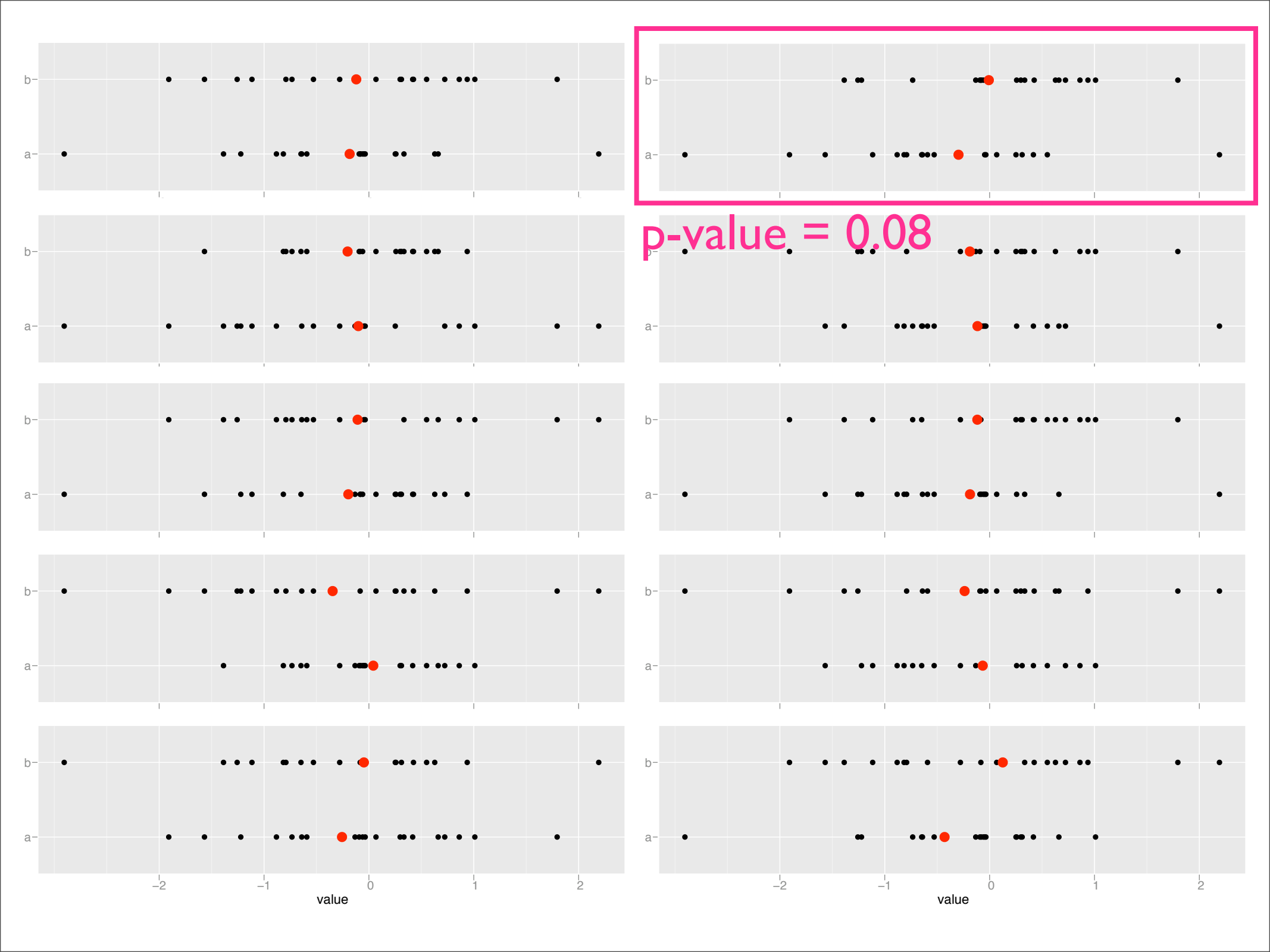

group
b
a
−2
−1
0
1
2
value

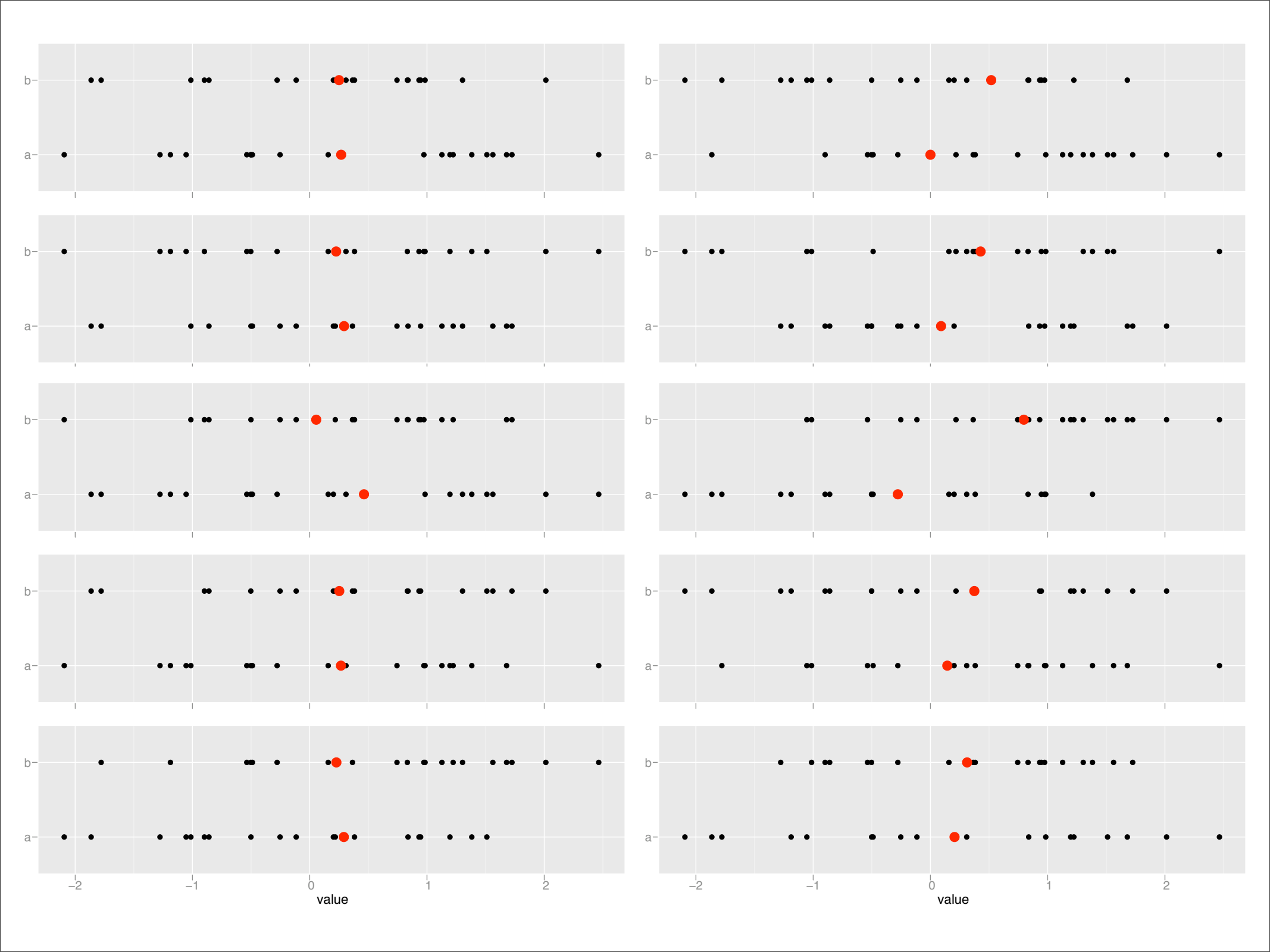
What is statistical significance?

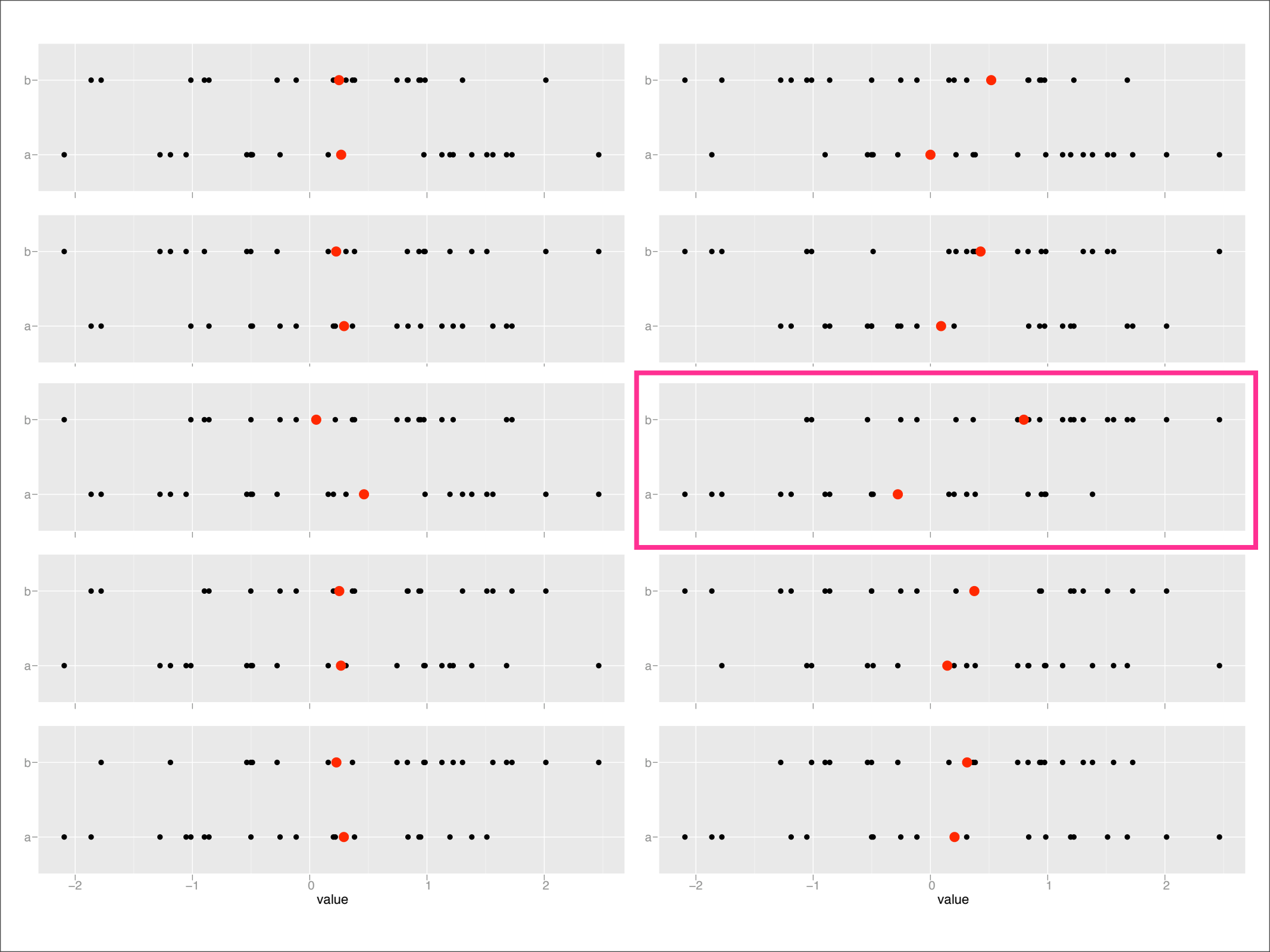
- Closely related to inference
- How likely are these results to be true for the population we're interested in?
- Or, what's the probability we see a strong effect from chance alone?

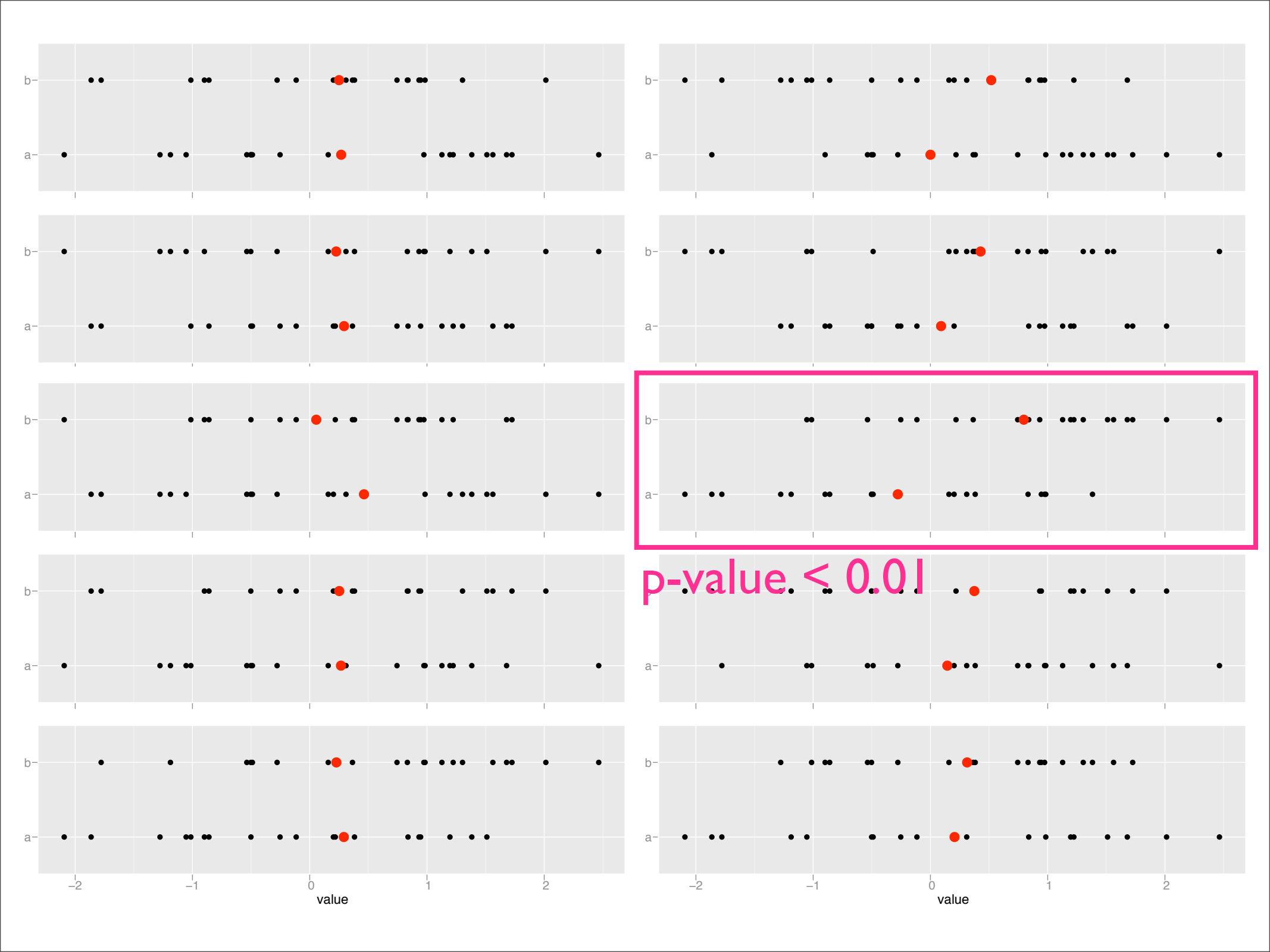

b
a
−2
−1
0
1
2
value


b
a
−2
−1
0
1
2
value


p-value = 0.08
b
a
−2
−1
0
1
2
value


b
a
−2
−1
0
1
2
value

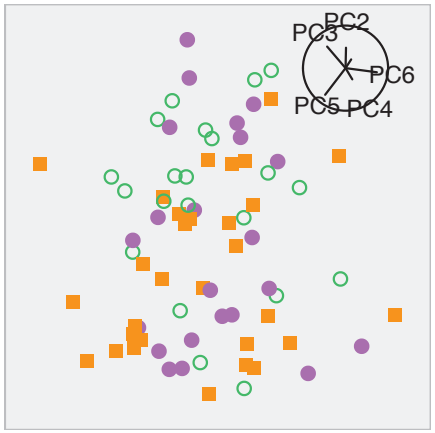

b
a
−2
−1
0
1
2
value

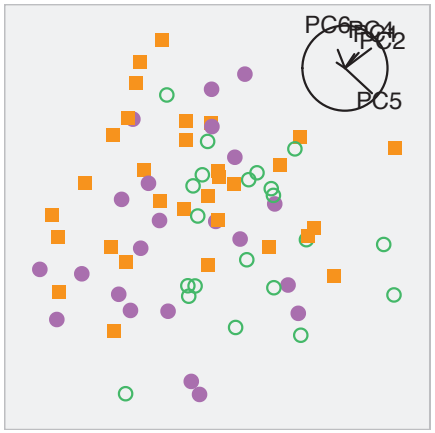

b
a
b
a
b
a
b
a
b
a
b
a
b
a
b
a
b
a
b
a
p-value < 0.01
−2
−1
0
1
2
value
−2
−1
0
1
2
value

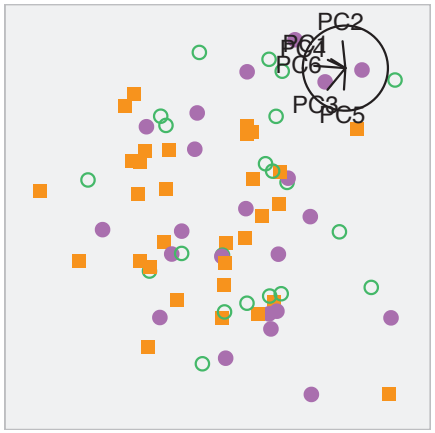
How does it work?

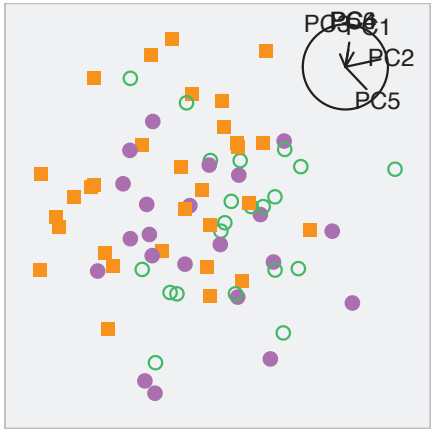
Randomise
Repeat
Reject

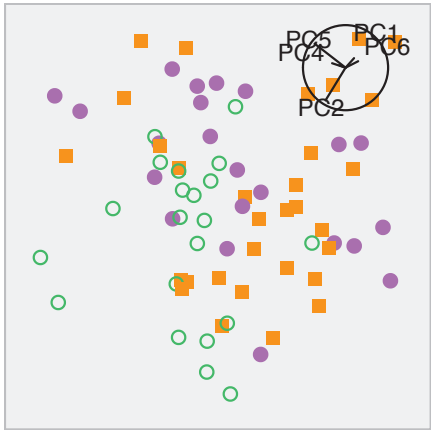
- If the variables are independent, then the structure will remain the same when the dependence between ordered pairs is broken.
- Take one column of data and permute the values, make a plot, repeat many times.
- Is the real data plot distinguishable from the permuted data plots? If so, the structure in the plot of data is real.

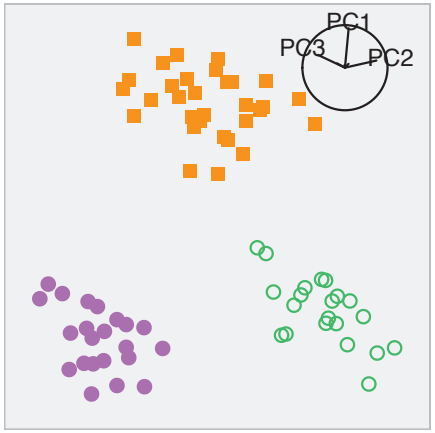


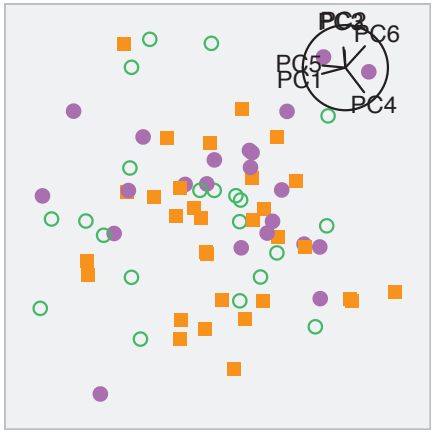






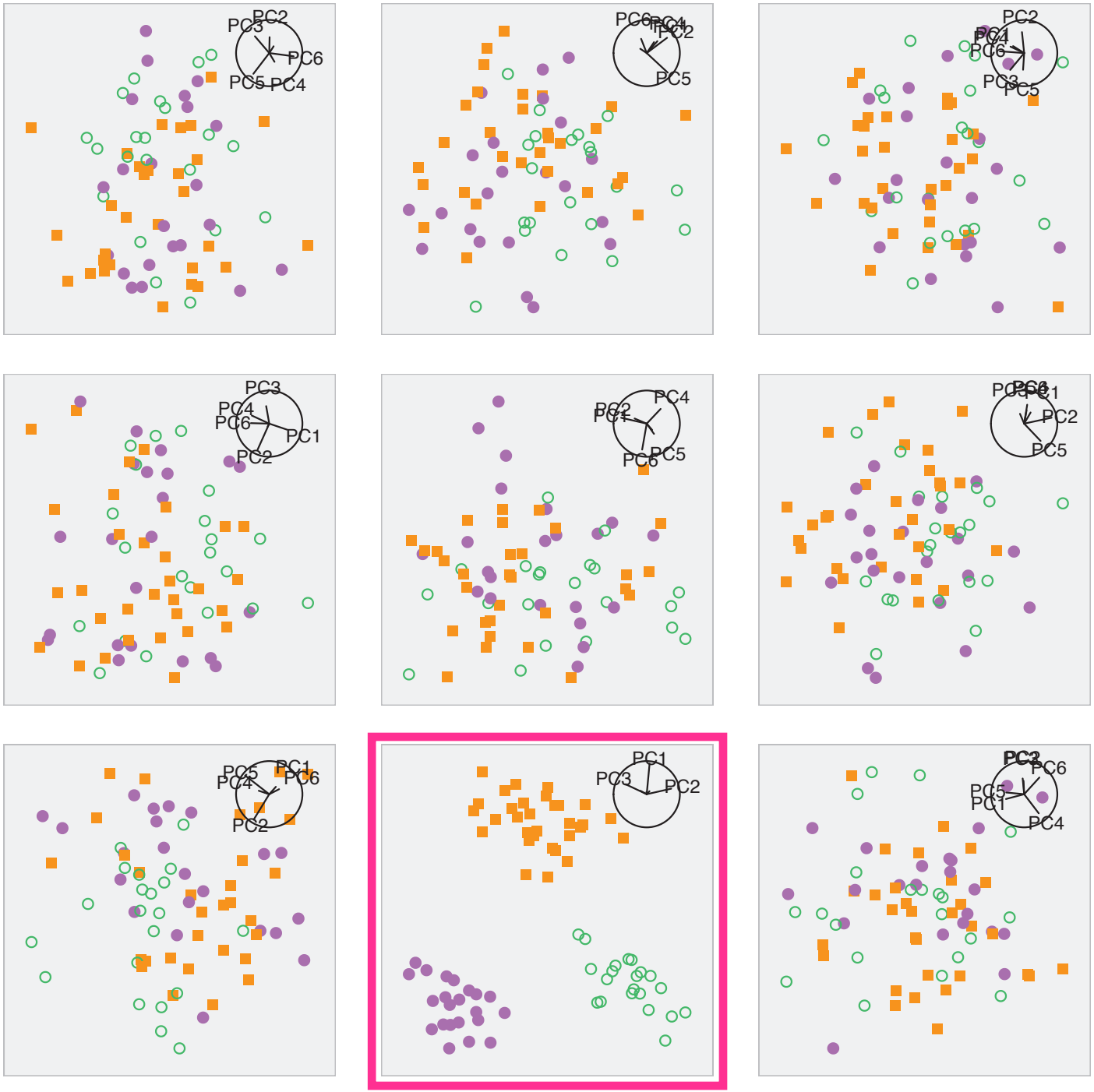






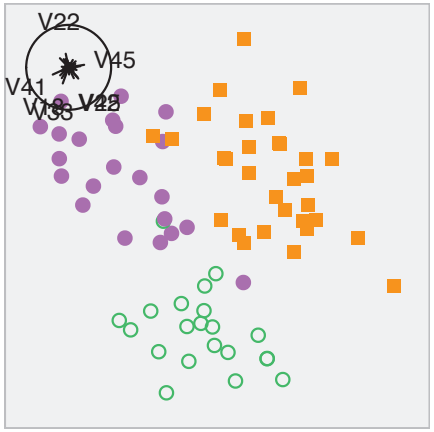
Example: inference for classification

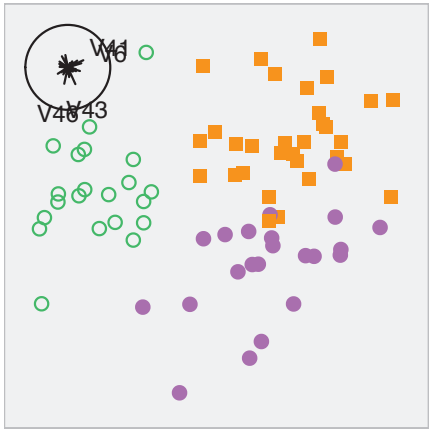
Which of these is different?

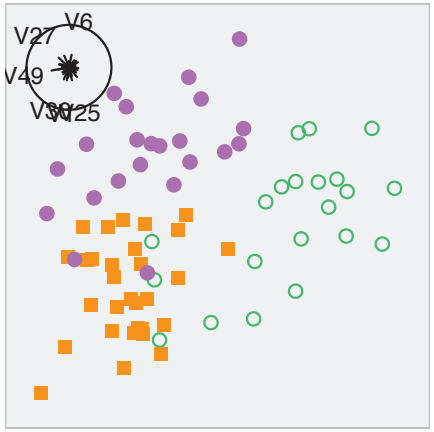


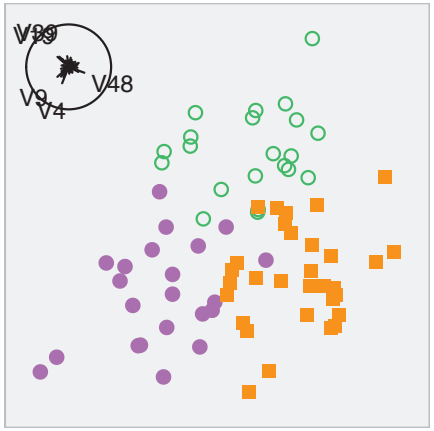
Example: inference for classification

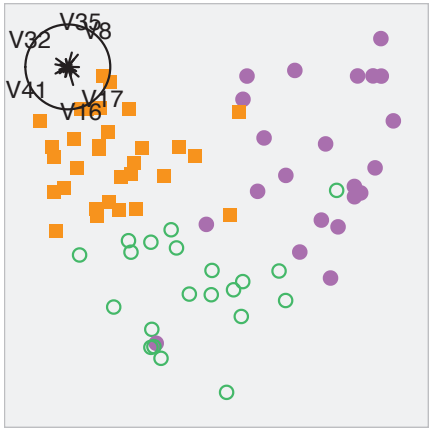
Which of these is different?

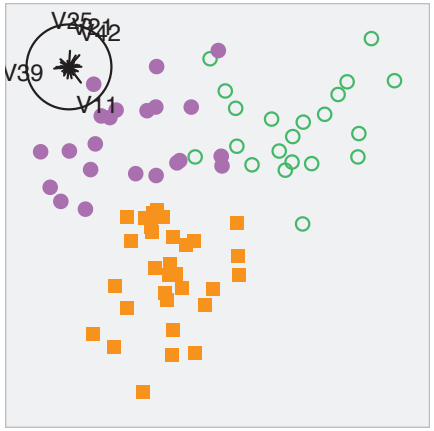


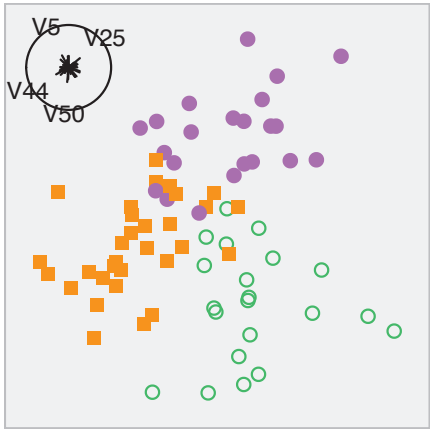


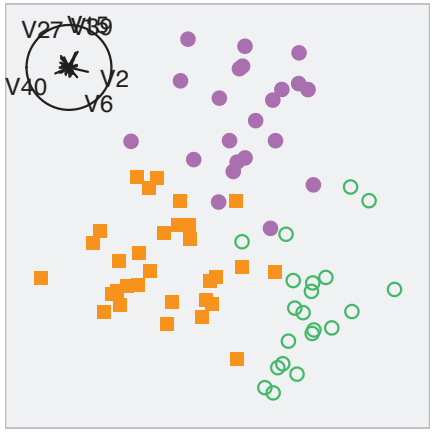




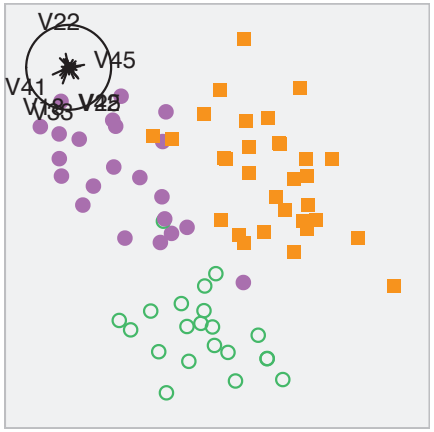


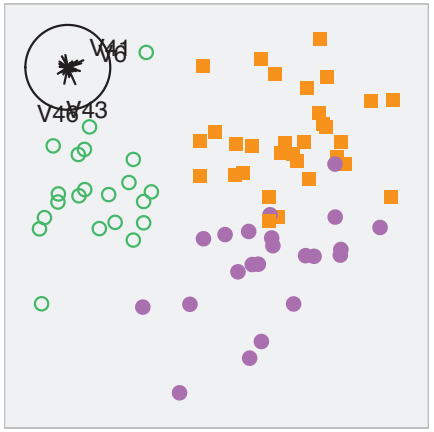


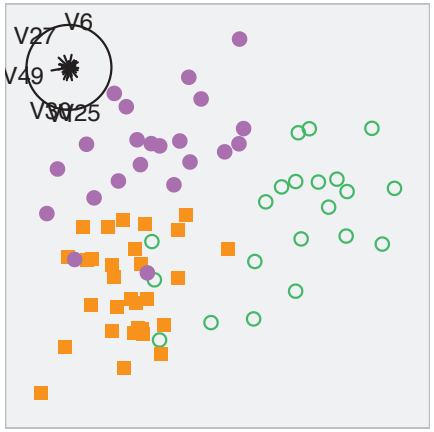


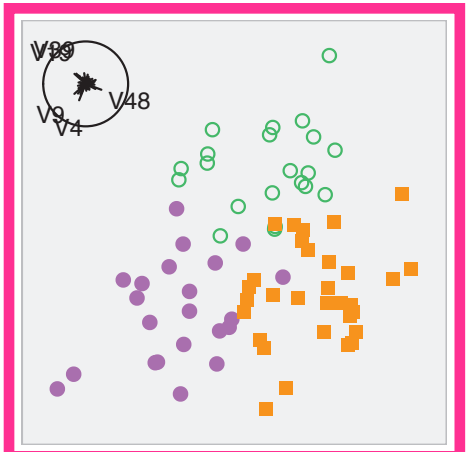


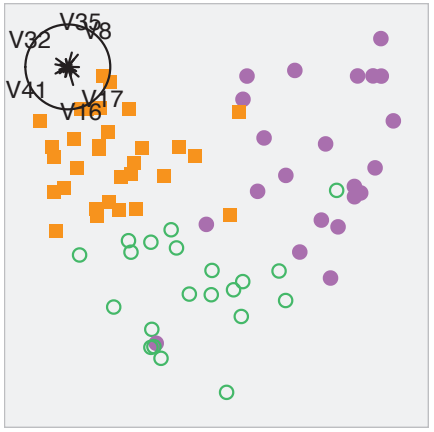
Which one of these is the real data?

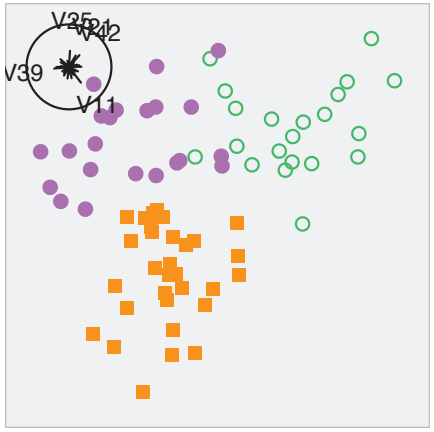


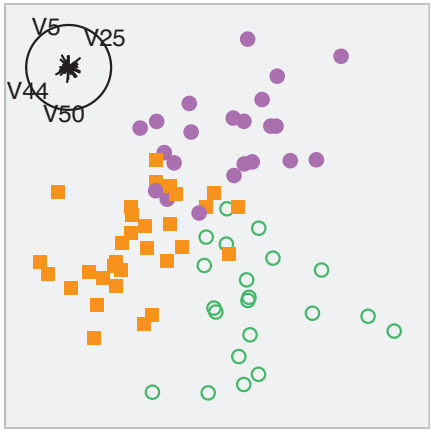


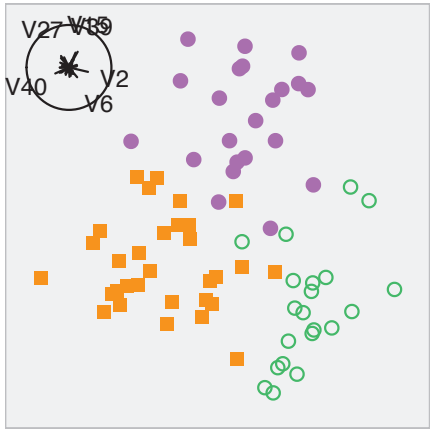












Which one of these is the real data?

Can't tell?

Real data is 50D with no real clusters. Apparent clusters occur because we have large P small n.

Ways to test for significance

- Independence between variables can be tested using permutation methods, where the appropriate columns of data values are shuffled to break the association in the ordered tuples.
- Distributional forms can tested by simulating samples from the distribution using parameters estimated from the sample.