Three Threads of Modern Applied Statistics

Edward A. Roualdes

2019-09-06

Introduction

Statistics attempts to model real world processes.

All models are wrong, but some are useful.

— attributed to George E. P. Box

Outline

1. Part I
   - plot all your data
2. Part II
   - example
   - p-values are nigh meaningless if assumptions are broken
   - rehearse proper interpretation of p-values
3. Part III
   - checking model suitability by simulating data

Part I

Plot all of your data, when reasonable. Chances are it's reasonable.

Bar Charts

A relatively recent revolt against bar charts

- Show the data, don't conceal them (Drummond and Vowler 2011)
- Kick the bar chart habit (2014, Nature Editorial)
- Beyond bar and line graphs (Weissgerber et al. 2015)
- Dynamite plots must die (Irizarry 2019)

Bar Chart or Dynamite Plot

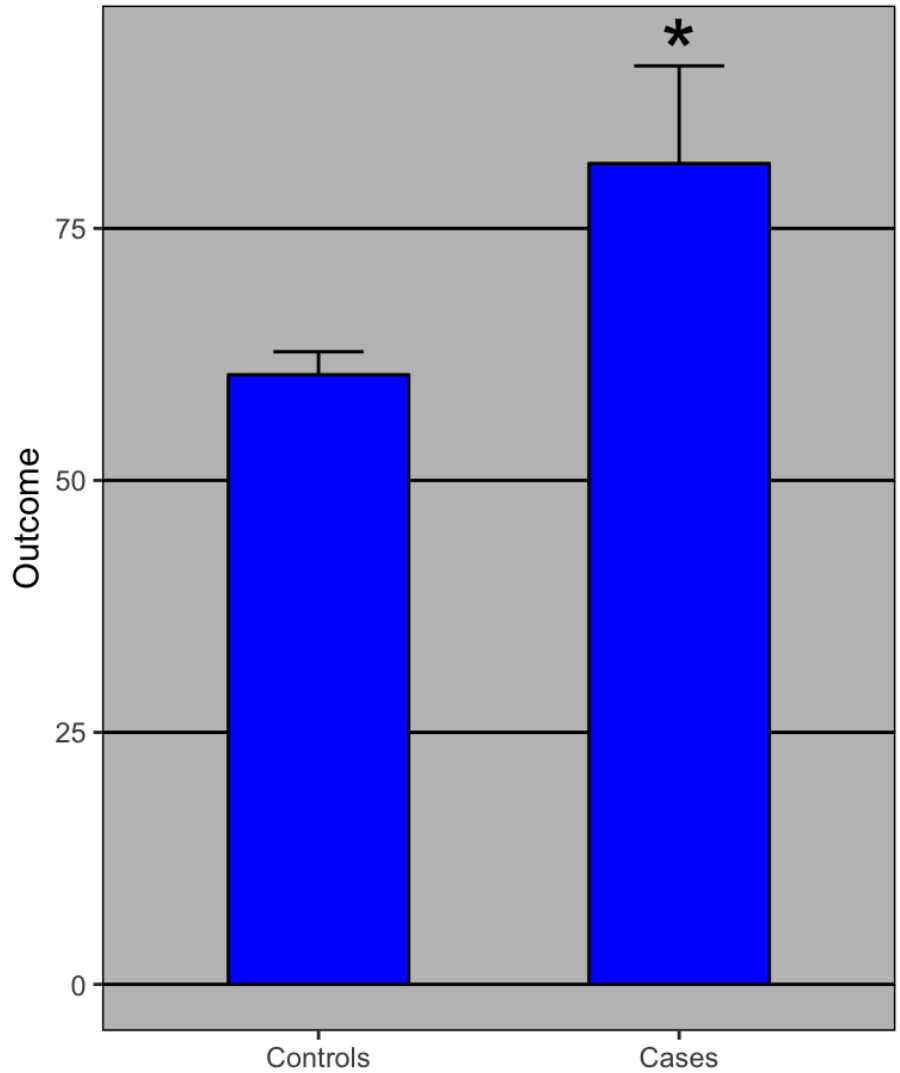


Figure 1

~~Bar Charts~~ Plot All Your Data

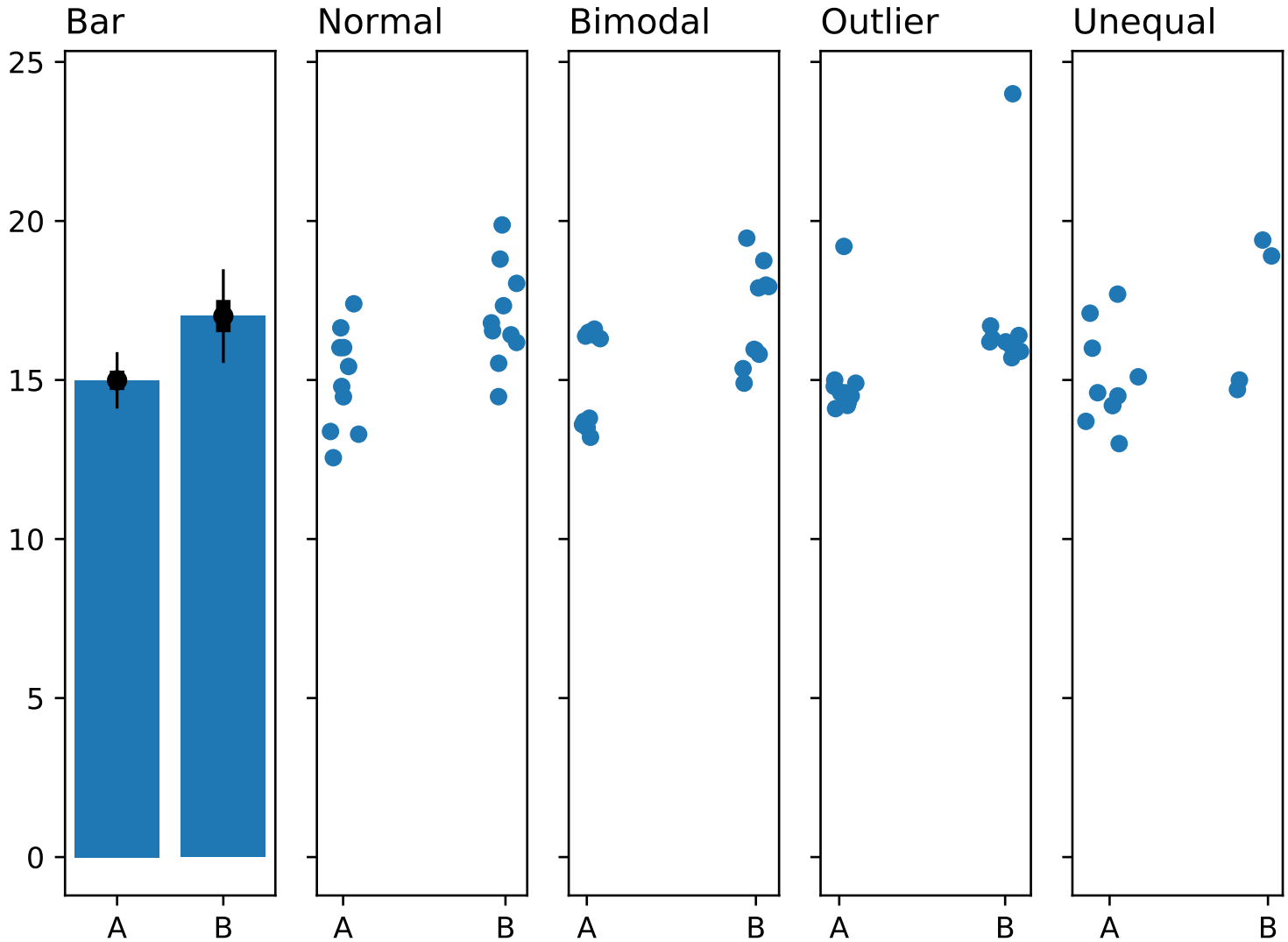
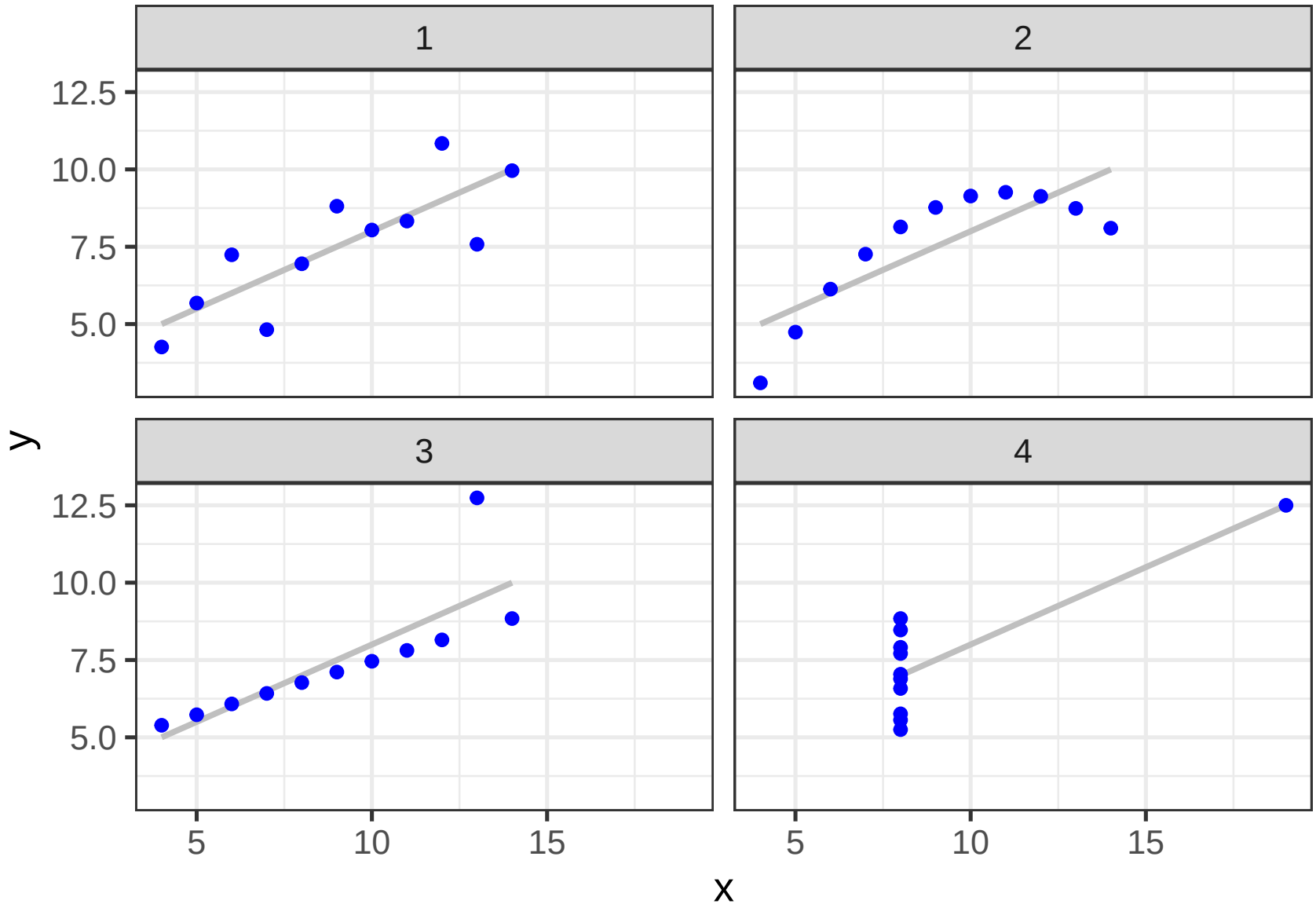


Figure 2

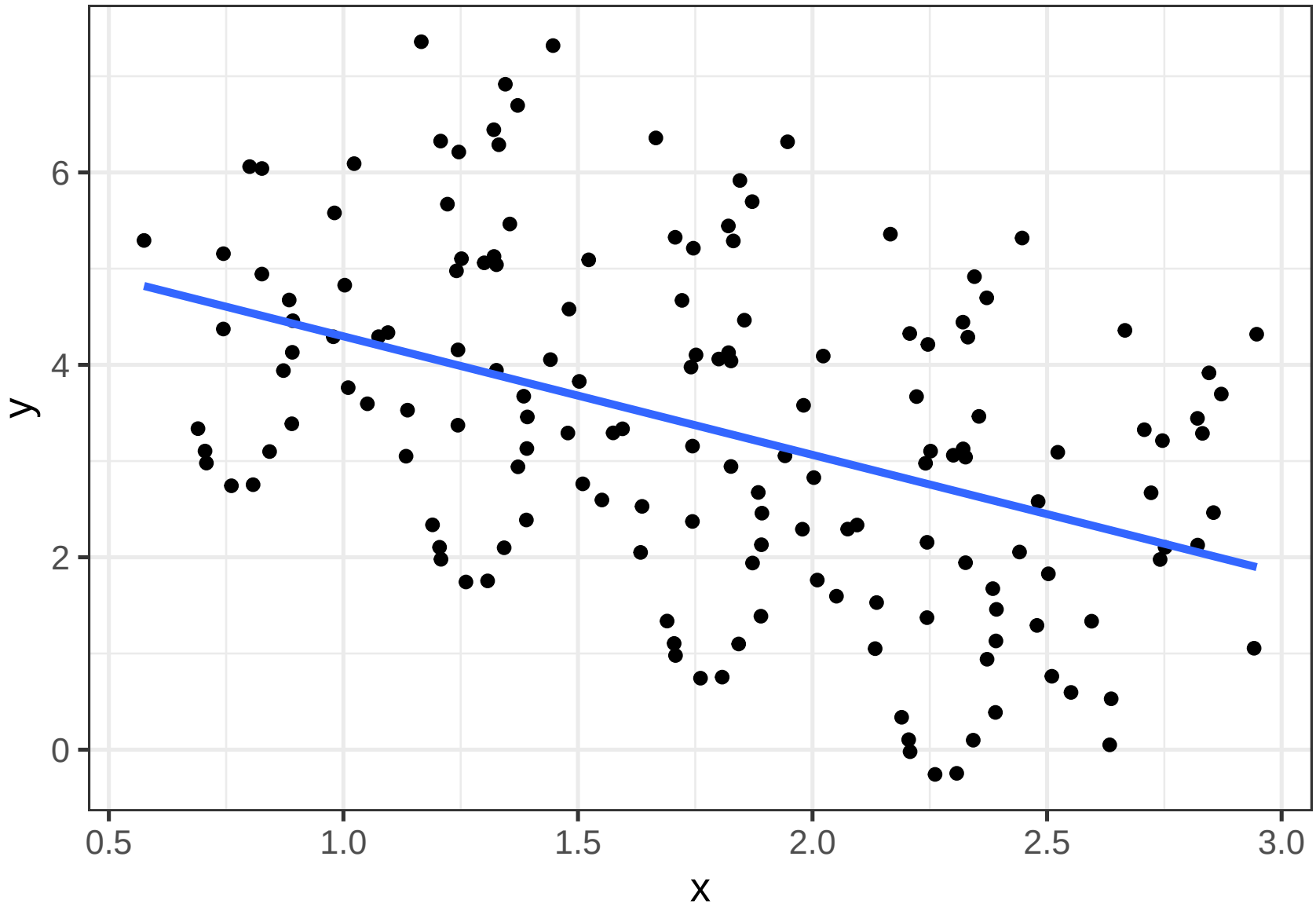
Anscombe's Quartet, 1973



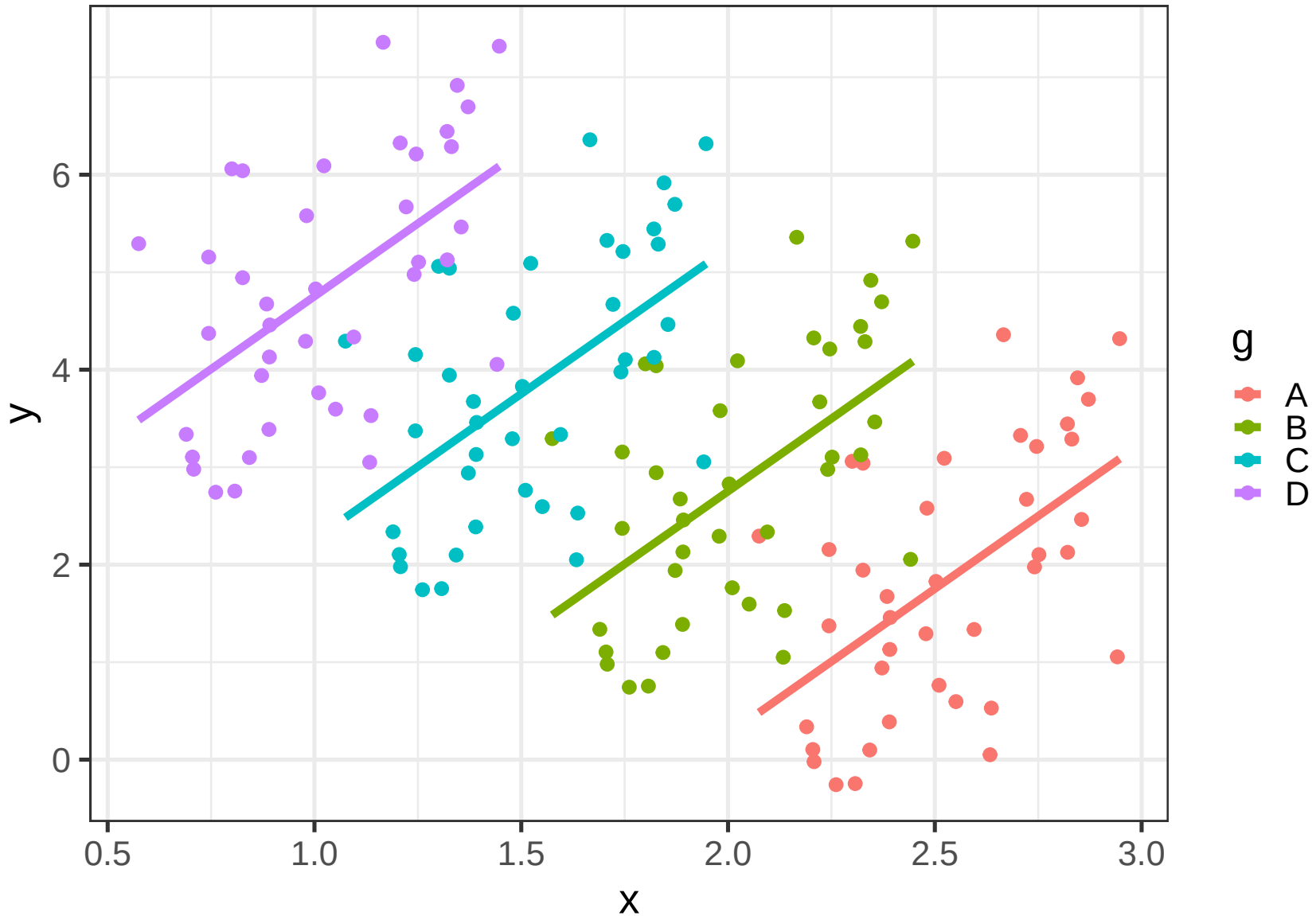
Same Stats, Different Graphs

Datasaurus

Simpson's ~~Paradox~~ data by groups can be annoying



Simpson's ~~Paradox~~ data by groups can be annoying



Part 2

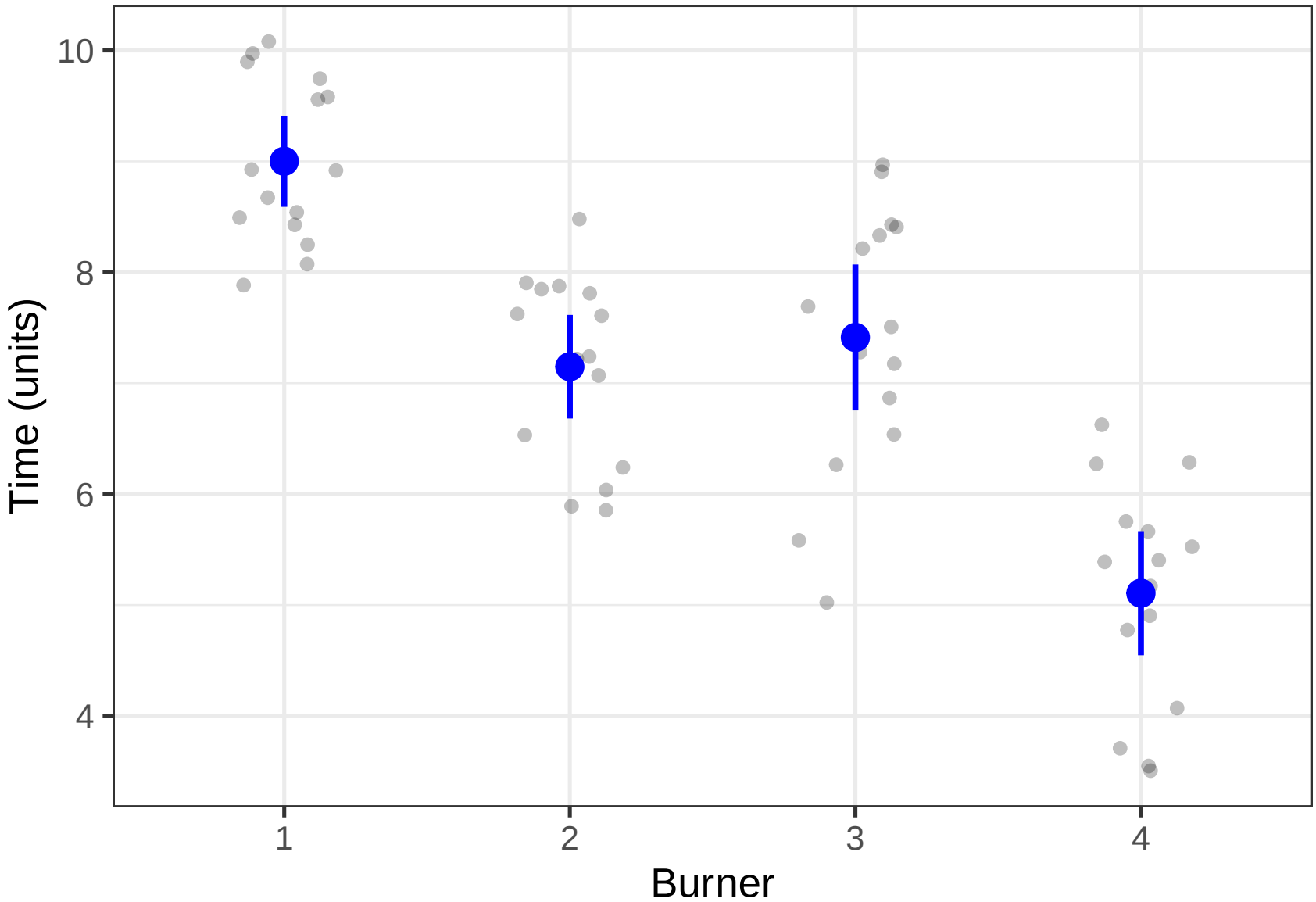
~~Reproducibility~~ Applied statistics is hard.

"Ninety-seven percent of original studies had significant results ($P < .05$). Thirty-six percent of replications had significant results." (Collaboration and others 2015)

Example

Suppose I'm interested in the rate at which the four burners on my stove-top can boil 6 cups of water.

Example



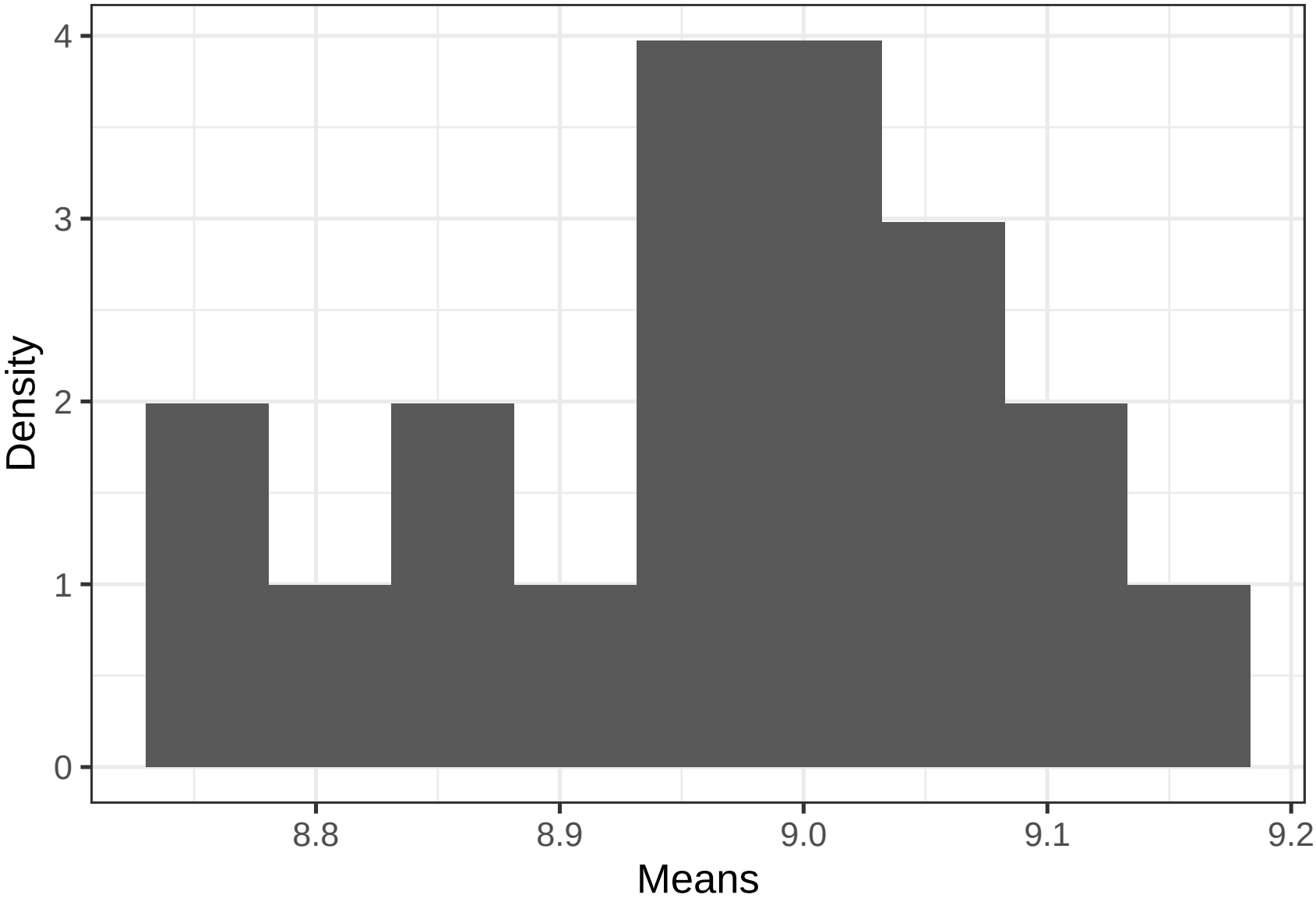
Example

Just consider burner 1. Let's say I tested (even though I didn't) burner 1 five times a day for 100 days in a row. On each day I recorded the mean time to boil 6 cups of water.

```
## [1] 8.7 9.0 9.1 9.0 8.9 9.1
```

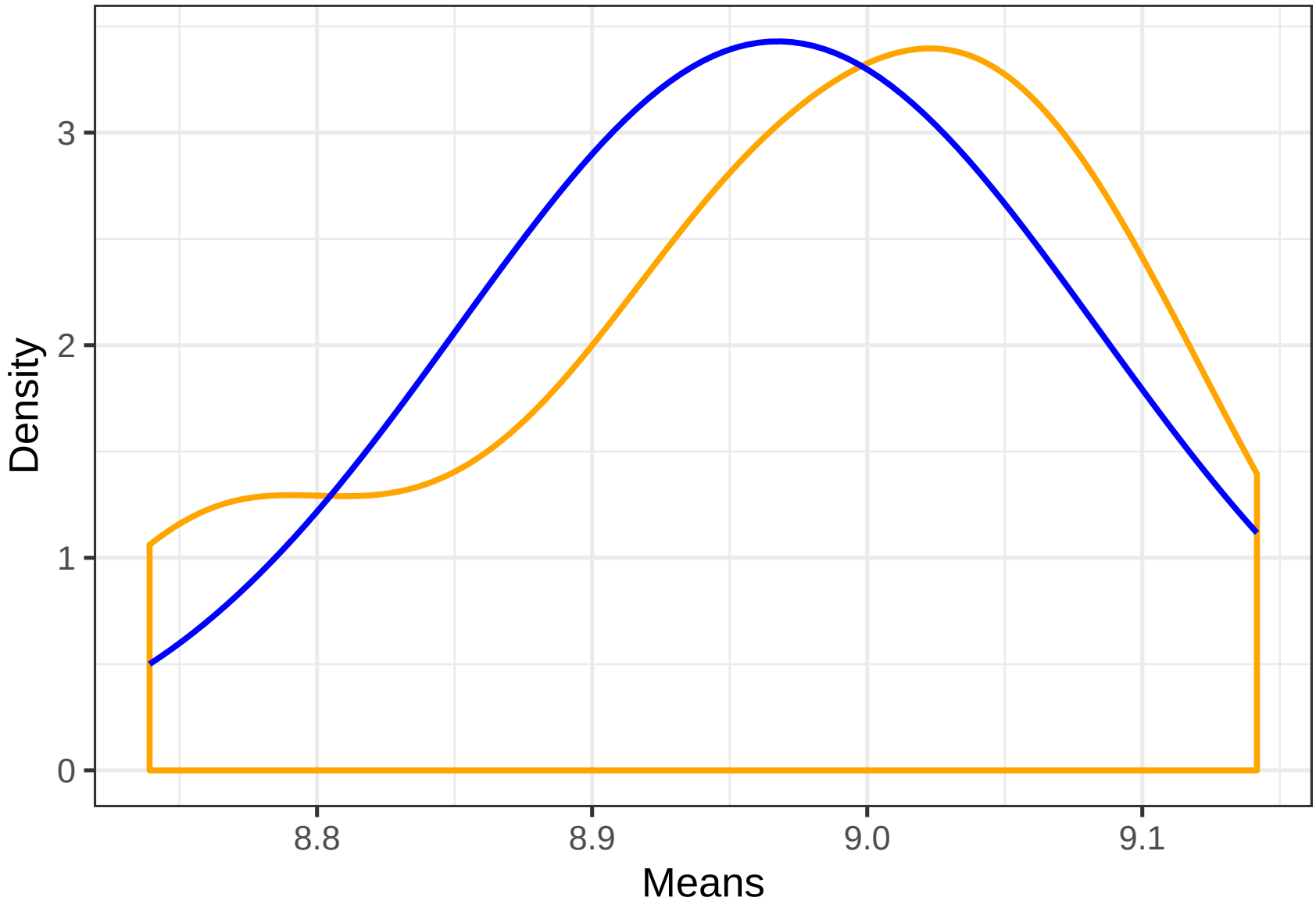
Example

Often interested in the shape of the (theoretically many) means.



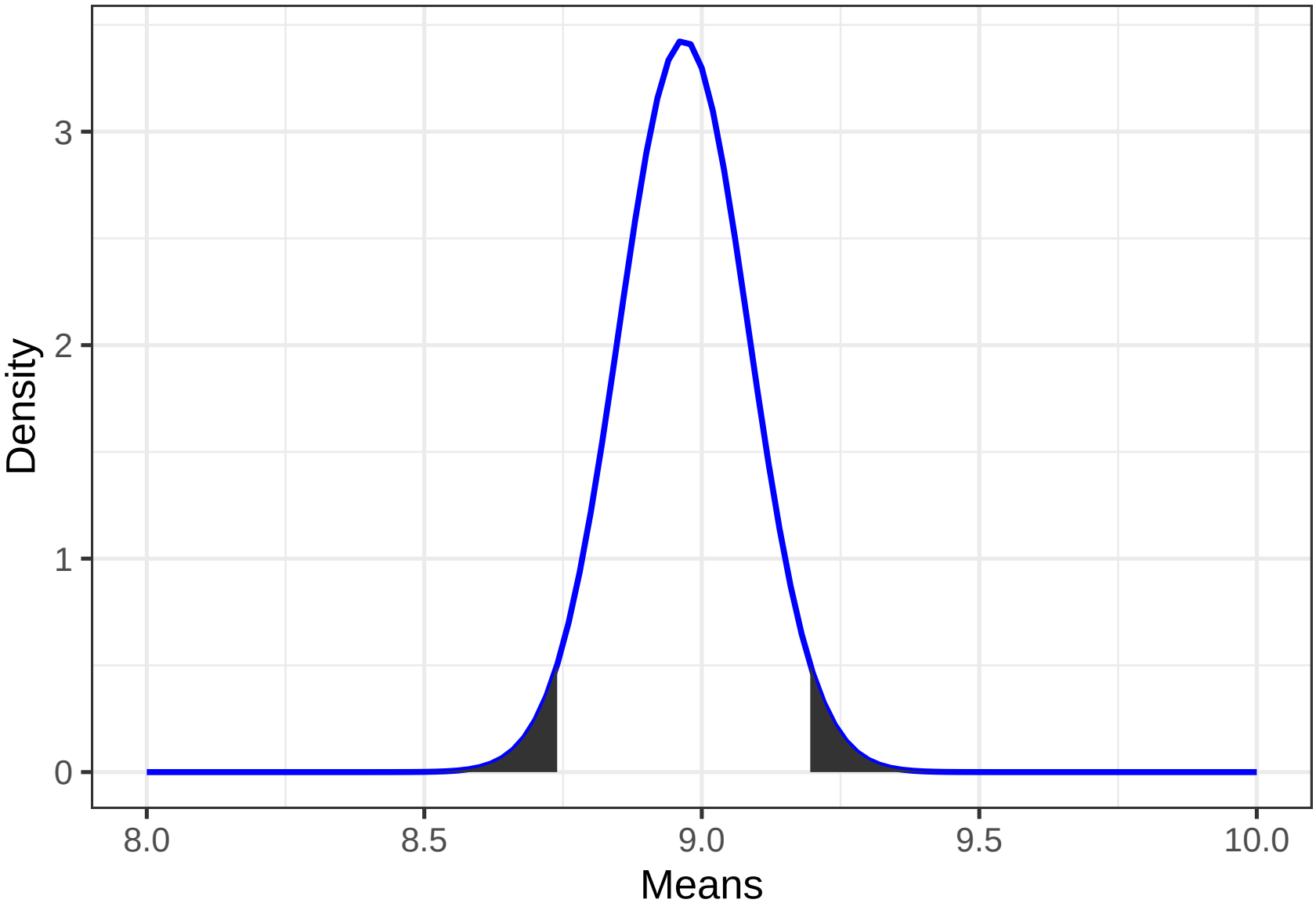
Example

Statistics dictates the shape of the distribution of the sample means.



Example

P-values come from the theoretical shape, not the actual shape.



p-values don't...

...tell us why they're small (or big).

"A small P-value is the net result of some combination of random variation and violations of model assumptions, but does not indicate which (if any) assumption is violated.".

— [@Amrhein:2019]

p-values don't. . .

. . . prove anything. At best, they provide evidence.

"Getting evidence of a genuine, repeatable effect is at most a necessary but not a sufficient condition for evidence of a substantive theory."

— attributed to Deborah Mayo

p-values don't...

...tell us the probability that a hypothesis is true.

p-values are. . .

. . . hard to interpret.

p-values are. . .

. . . not nor have ever meant to be dichotomized.

"No isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon."

— [@Fisher:1960]

p-values are. . .

. . . conditional on the model, its assumptions, your (null) hypothesis, and the data.

"To avoid the dichotomization . . . , one could report a measure of refutational evidence, such as a traditional P-value in a continuous fashion (as recommended by classic texts on testing such as Lehmann (1986), Testing Statistical Hypotheses (2nd ed.), p. 71), reporting an observed P-value as a measure of the degree of compatibility between the hypothesis or model it tests and the data."

— [@Amrhein:2019]

p-values are. . .

. . . measures of compatibility between model, its assumptions, your hypothesis, and the data.

p-values are. . .

. . . a continuous measure of compatibility.

p-values are. . .

. . . sometimes less than 0.05 because

- data collection had a hidden bias,
- the model is inappropriate,
- at least one assumption is broken,
- the data are incompatible with the model, its assumptions, and/or the null hypothesis.

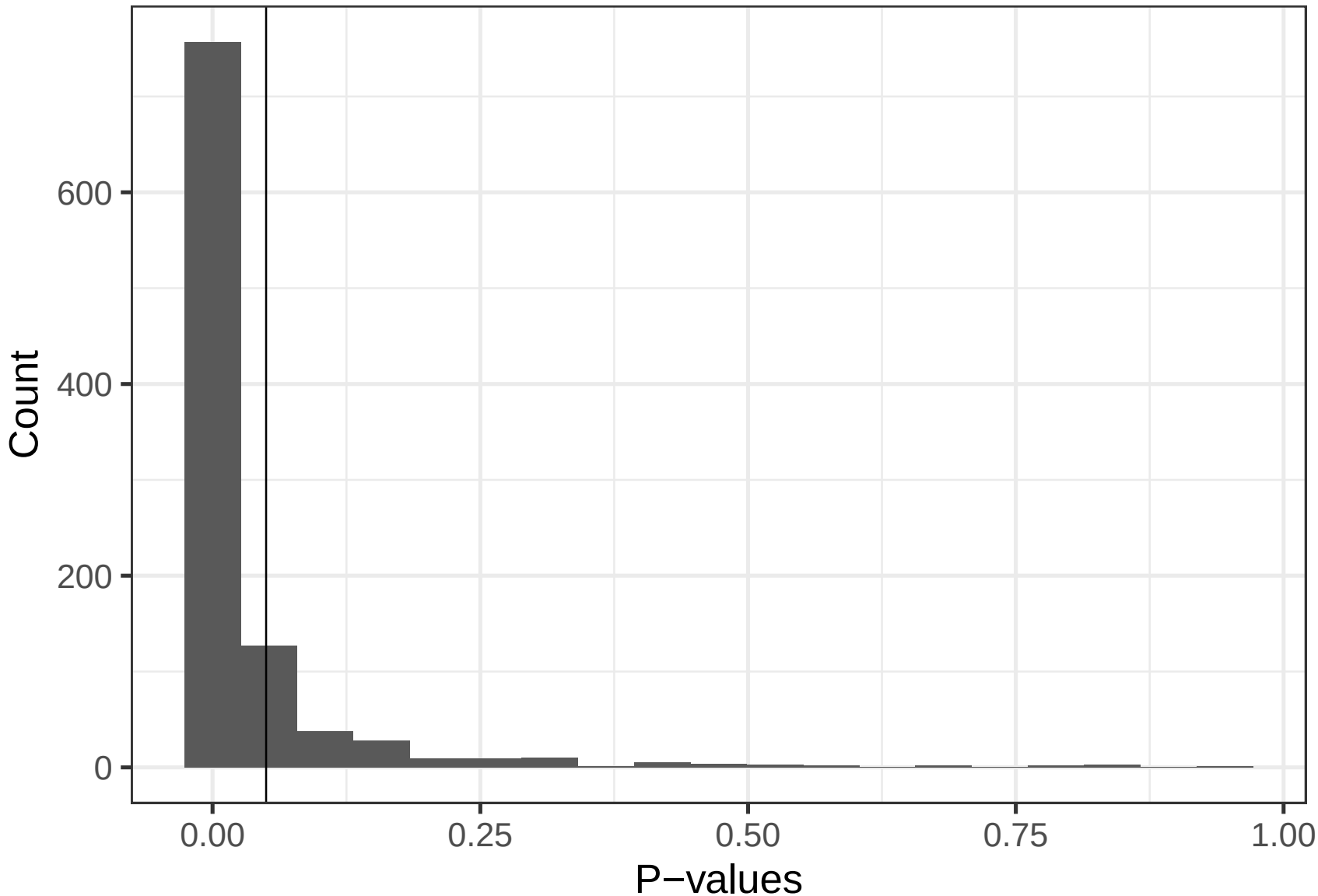
p-values are. . .

. . . sometimes greater than 0.05 because

- data collection had a hidden bias,
- the model is inappropriate,
- at least one assumption is broken,
- the data are compatible with the model, its assumptions, or the null hypothesis.

p-values are. . .

. . . fickle (Halsey et al. 2015).



p-values are. . .

. . . often wrong in various ways (Gelman and Carlin 2014).

Error	Type	Estimated Probability
incorrectly rejct H0	1	0.05
correctly reject H0	2	0.2
wrong sign	S	0.016
magnitude overestimated	M	0.897

p-values are. . .

. . . essentially a swipe-right on tinder.

United States Conference on Teaching Statistics 2019
https://www.causeweb.org/cause/uscots/uscots19/keynote/2

Part 3

How can we better understand when our models are believable?

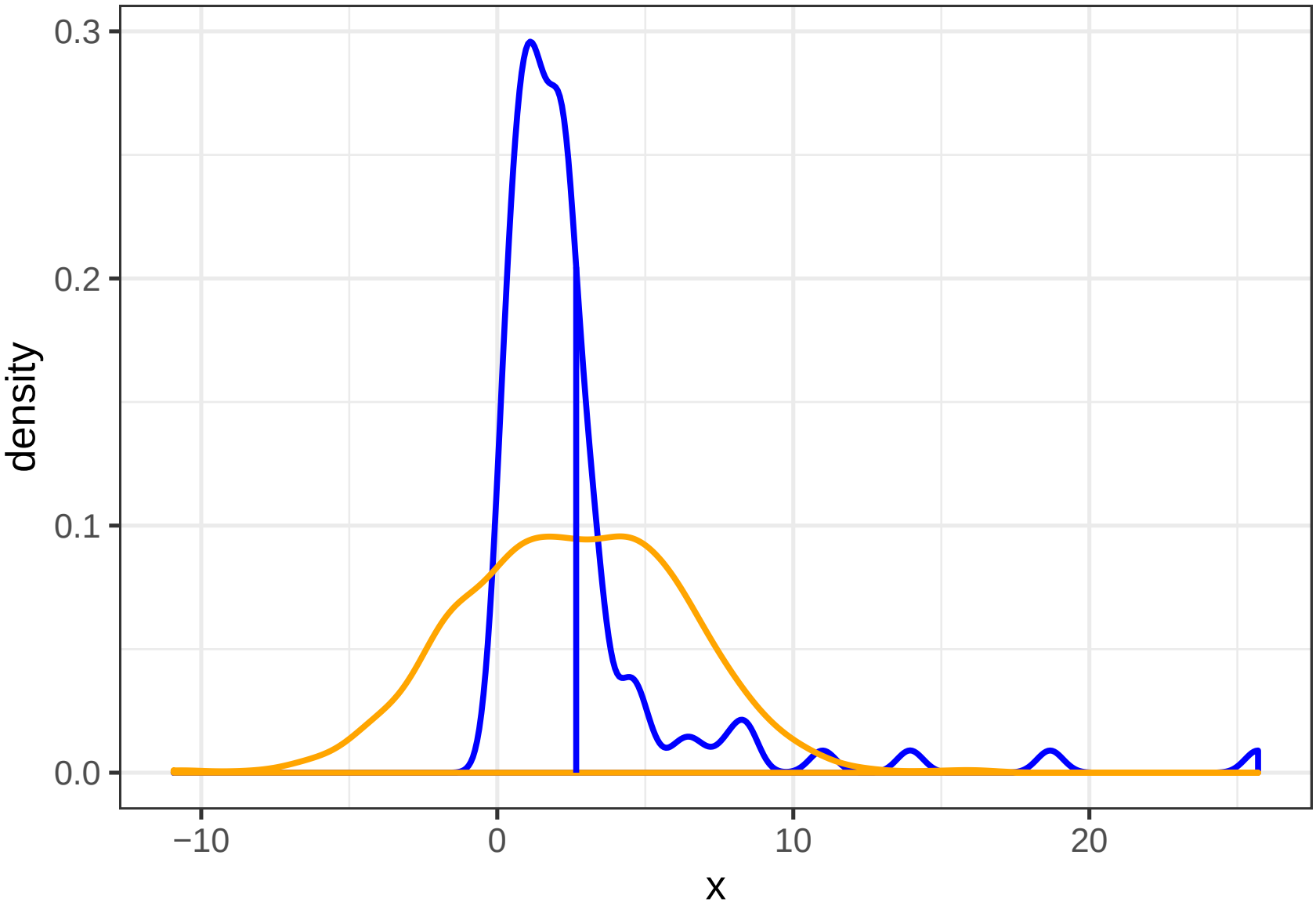
"It's better to solve the right problem approximately than to solve the wrong problem exactly."

— attributed to John W. Tukey

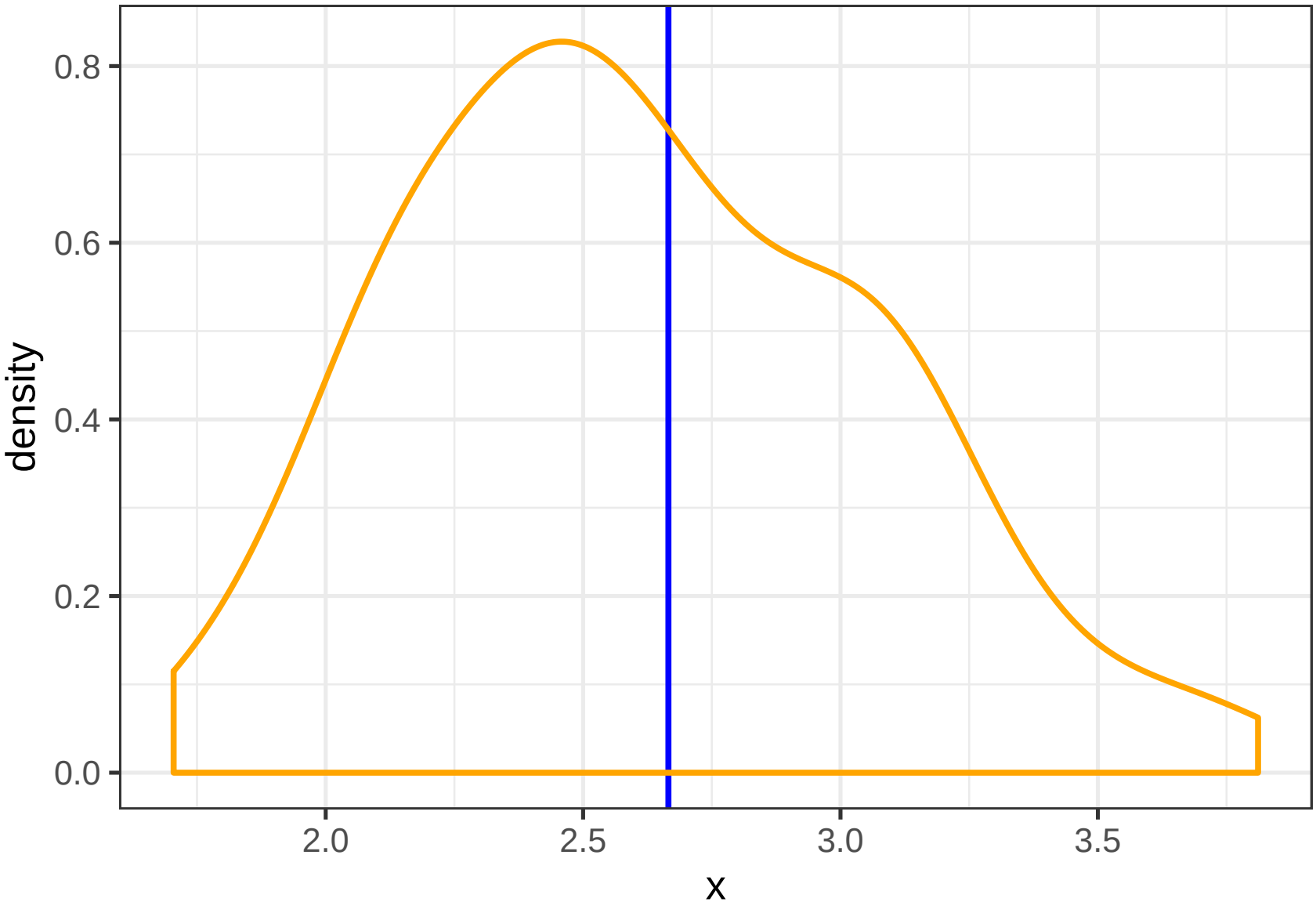
Predictive Models

Models that can at least simulate data similar to what we've already seen.

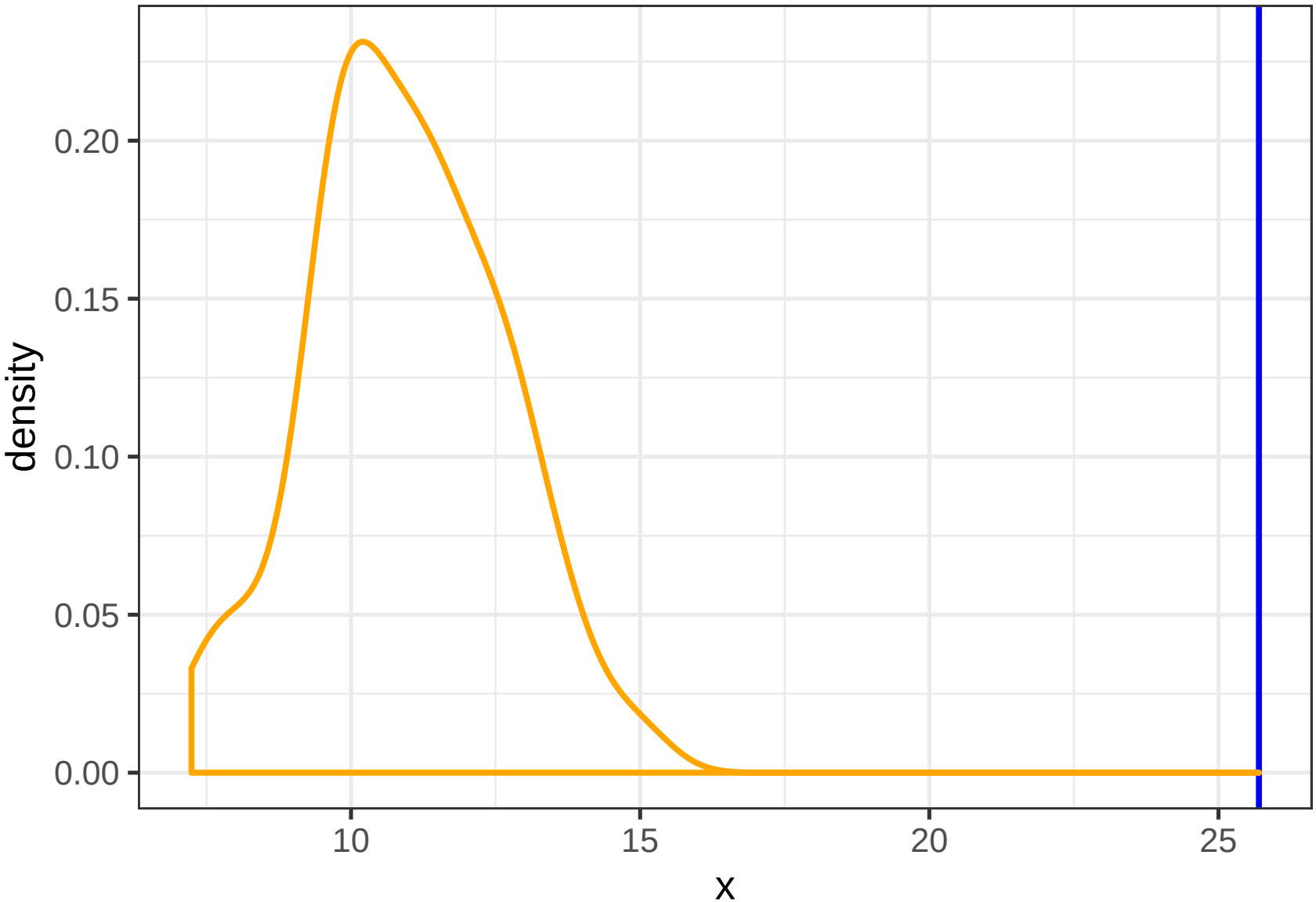
Model Simulated Data, original data



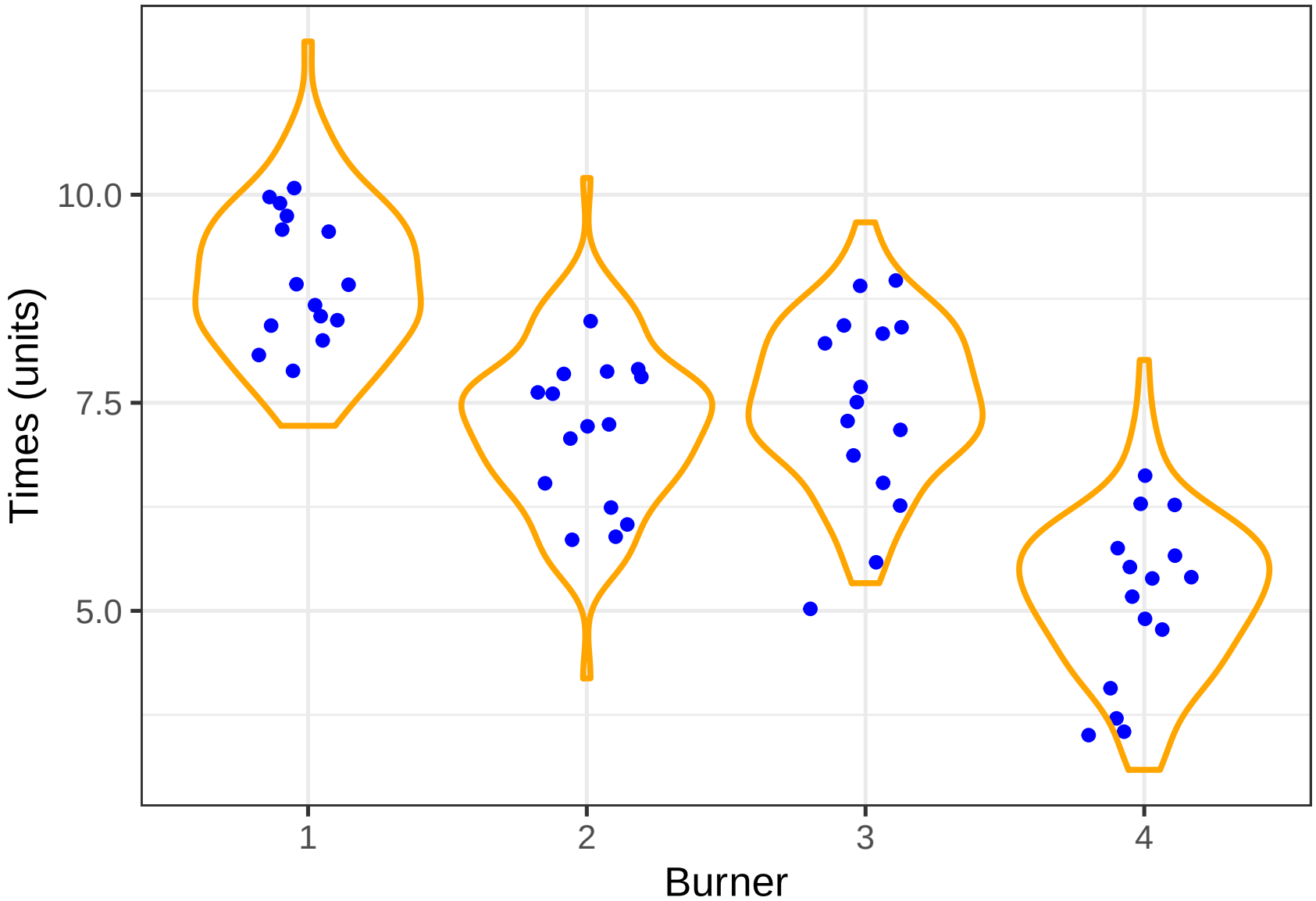
Model Simulated Data, mean



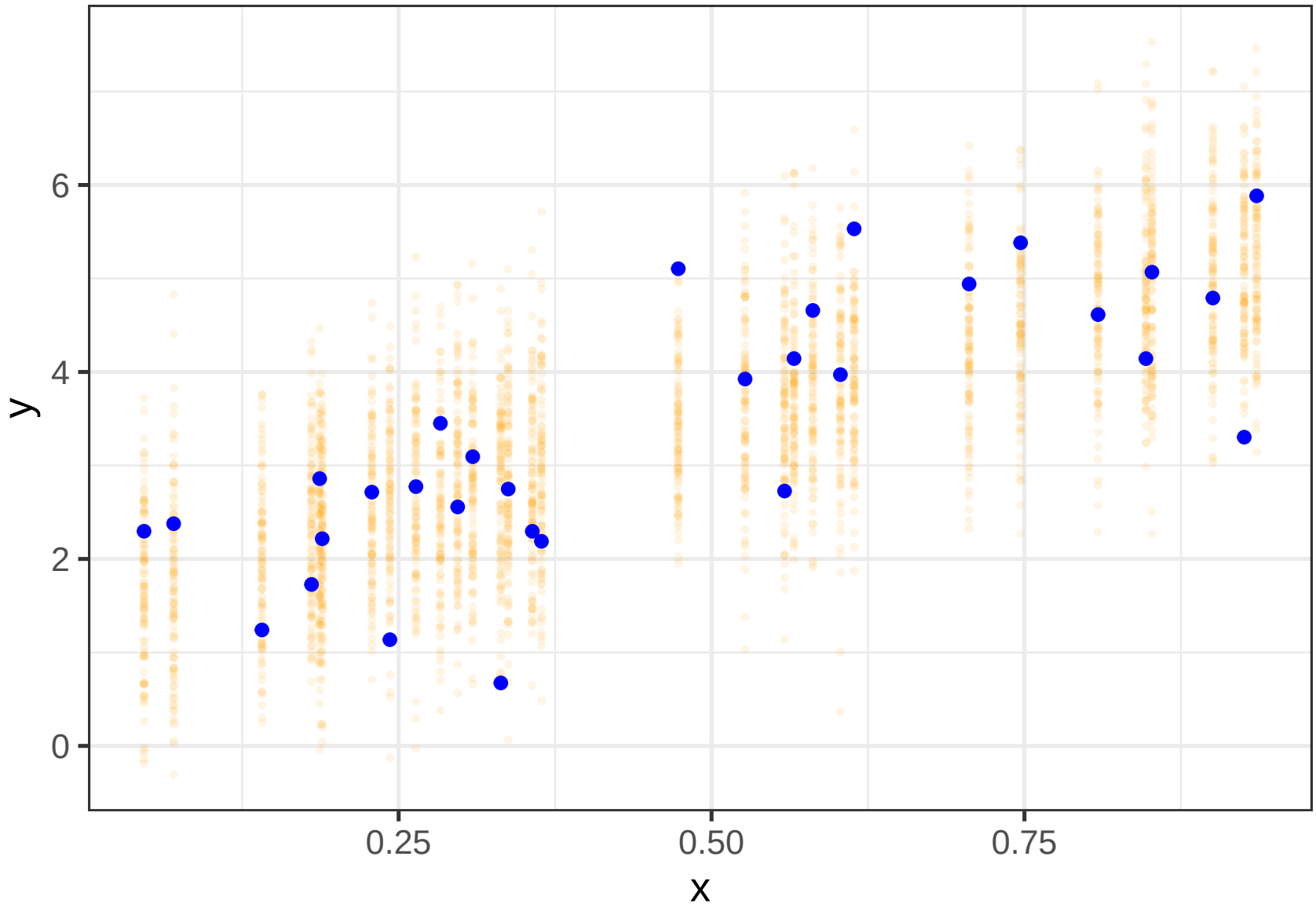
Model Simulated Data, max



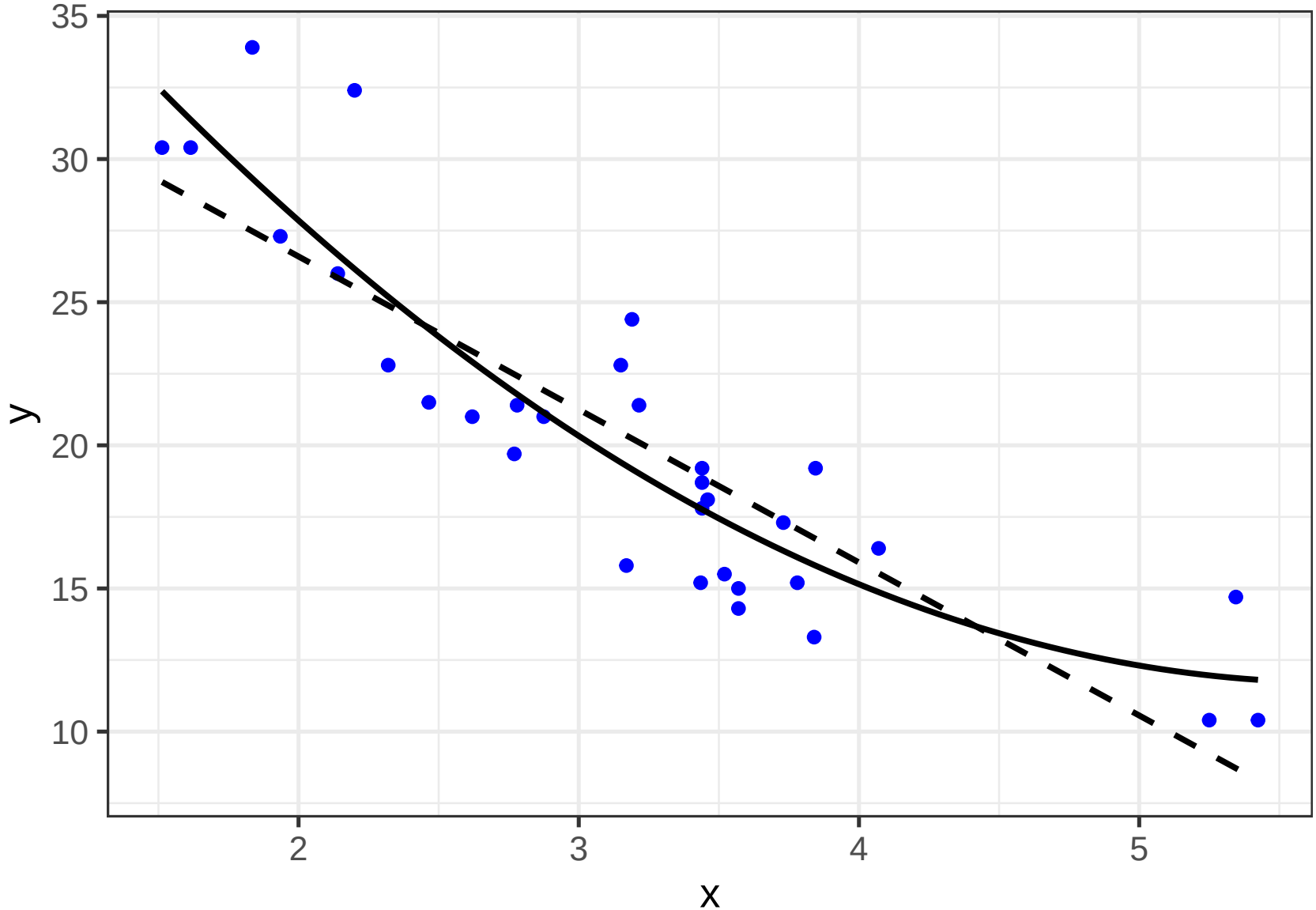
Model Simulated Data, ANOVA



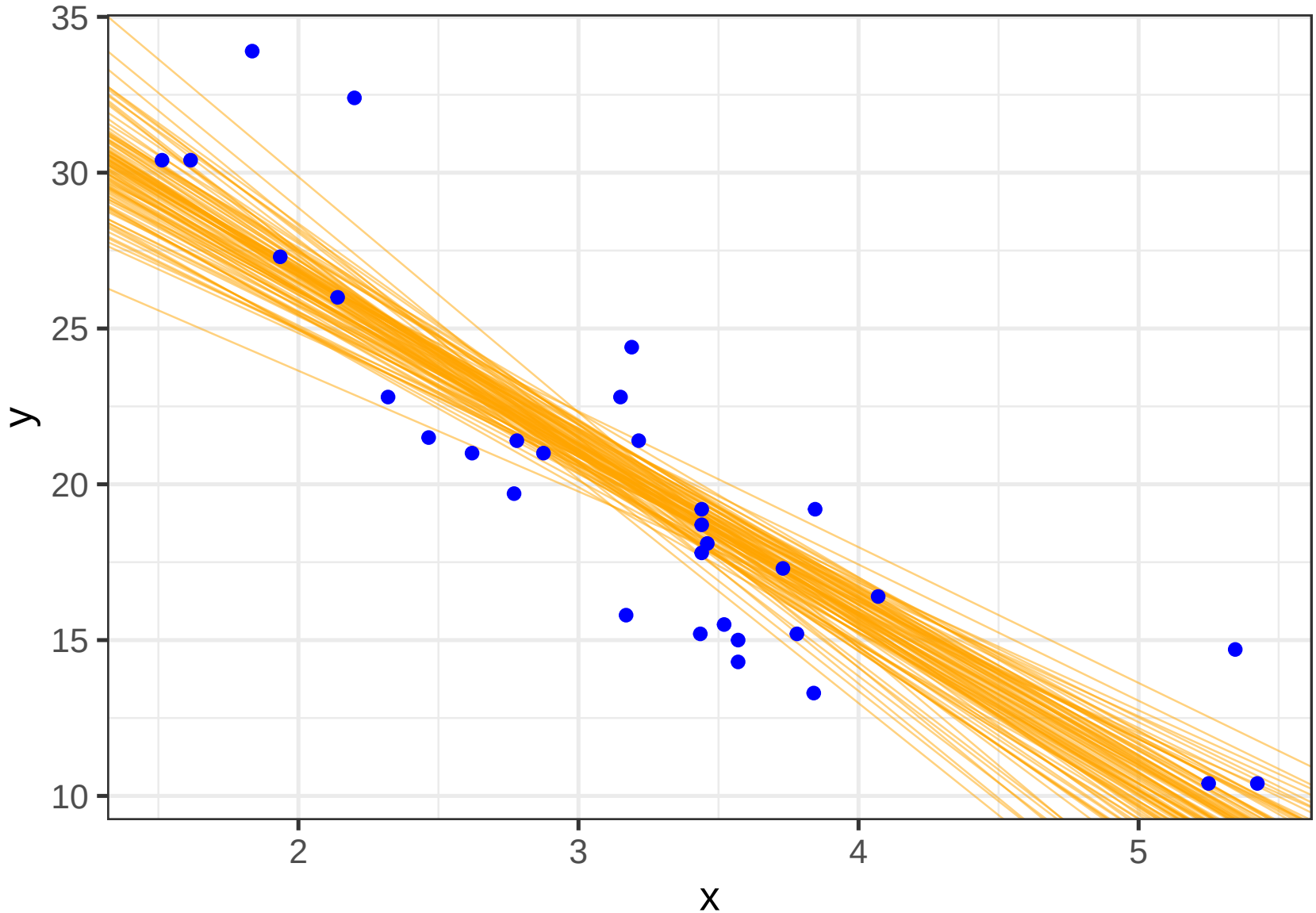
Model Simulated Data, linear regression



Model Simulated Data, linear model can be good enough



Model Simulated Data, linear model can be good enough



Model Simulated Data

The point: If simulated data from a model can't recover the aspects of the original data that are most important to a researcher, then the model probably isn't a reasonable approximation of the underlying process of interest.

End

Thank you.

https://roualdes.us/misc/bio.pdf

References

Anscombe, Francis J. 1973. "Graphs in Statistical Analysis." *The American Statistician* 27 (1). Taylor & Francis Group: 17–21.

Collaboration, Open Science, and others. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349 (6251). American Association for the Advancement of Science: aac4716.

Drummond, Gordon B, and Sarah L Vowler. 2011. "Show the Data, Don't Conceal Them." *Advances in Physiology Education* 35 (2). American Physiological Society Bethesda, MD: 130–32.

Gelman, Andrew, and John Carlin. 2014. "Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors." *Perspectives on Psychological Science* 9 (6). Sage Publications Sage CA: Los Angeles, CA: 641–51.

Halsey, Lewis G, Douglas Curran-Everett, Sarah L Vowler, and Gordon B Drummond. 2015. "The Fickle P Value Generates Irreproducible Results." *Nature Methods* 12 (3). Nature Publishing Group: 179.