Bootstrapping

(and a bit more)

Andrew Lampinen

Psych 252, Winter 2019

Outline

Why bootstrapping?

What is bootstrapping anyway?

Bootstrap confidence intervals

Bootstrap (and permutation) tests

Bootstrap power analyses

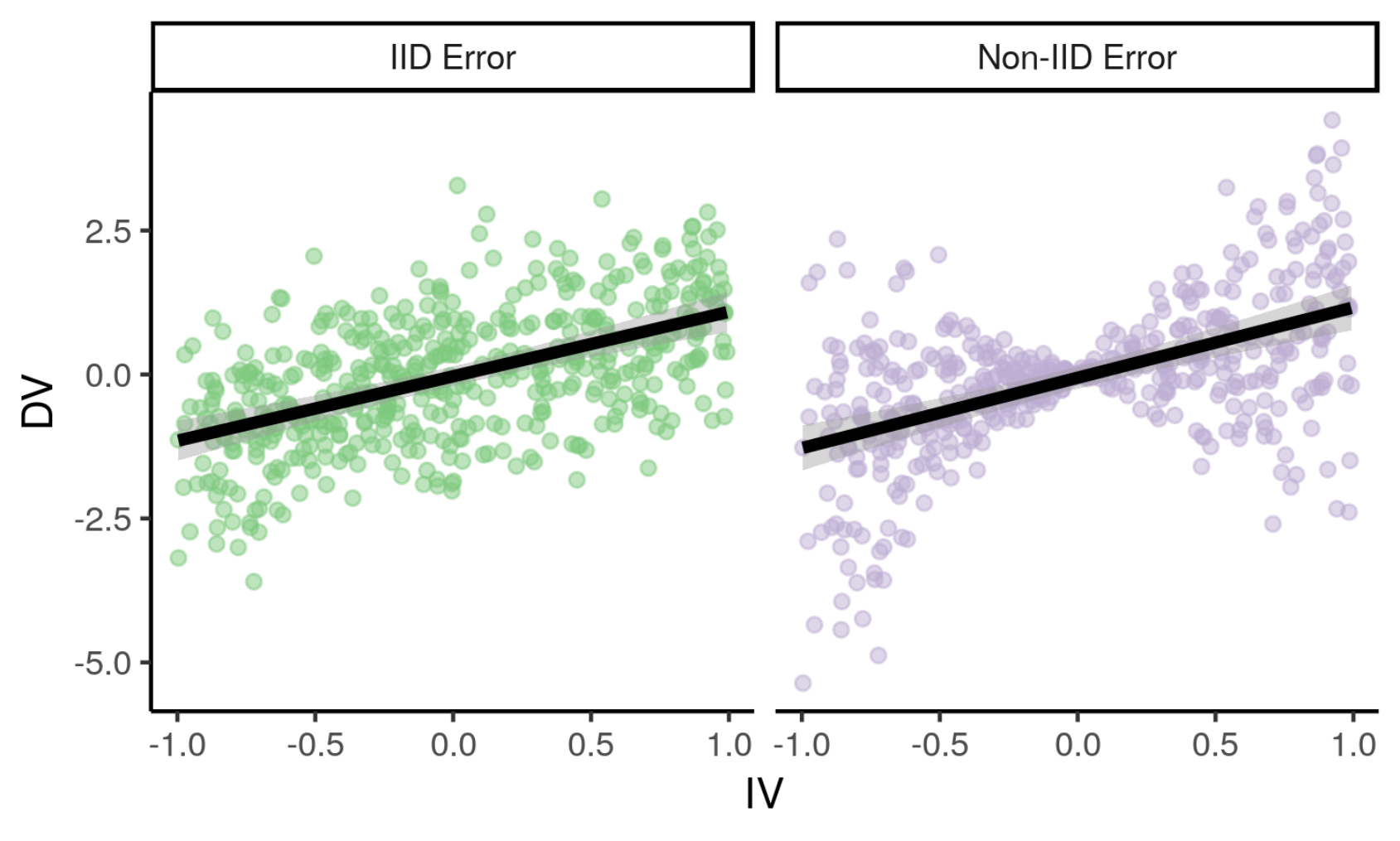
Wrapping up

These slides contain bullet points that summarize the argument.

Why bootstrapping?

Bootstrapping makes fewer assumptions than parametric methods.

Non-IID noise (heteroscedasticity)

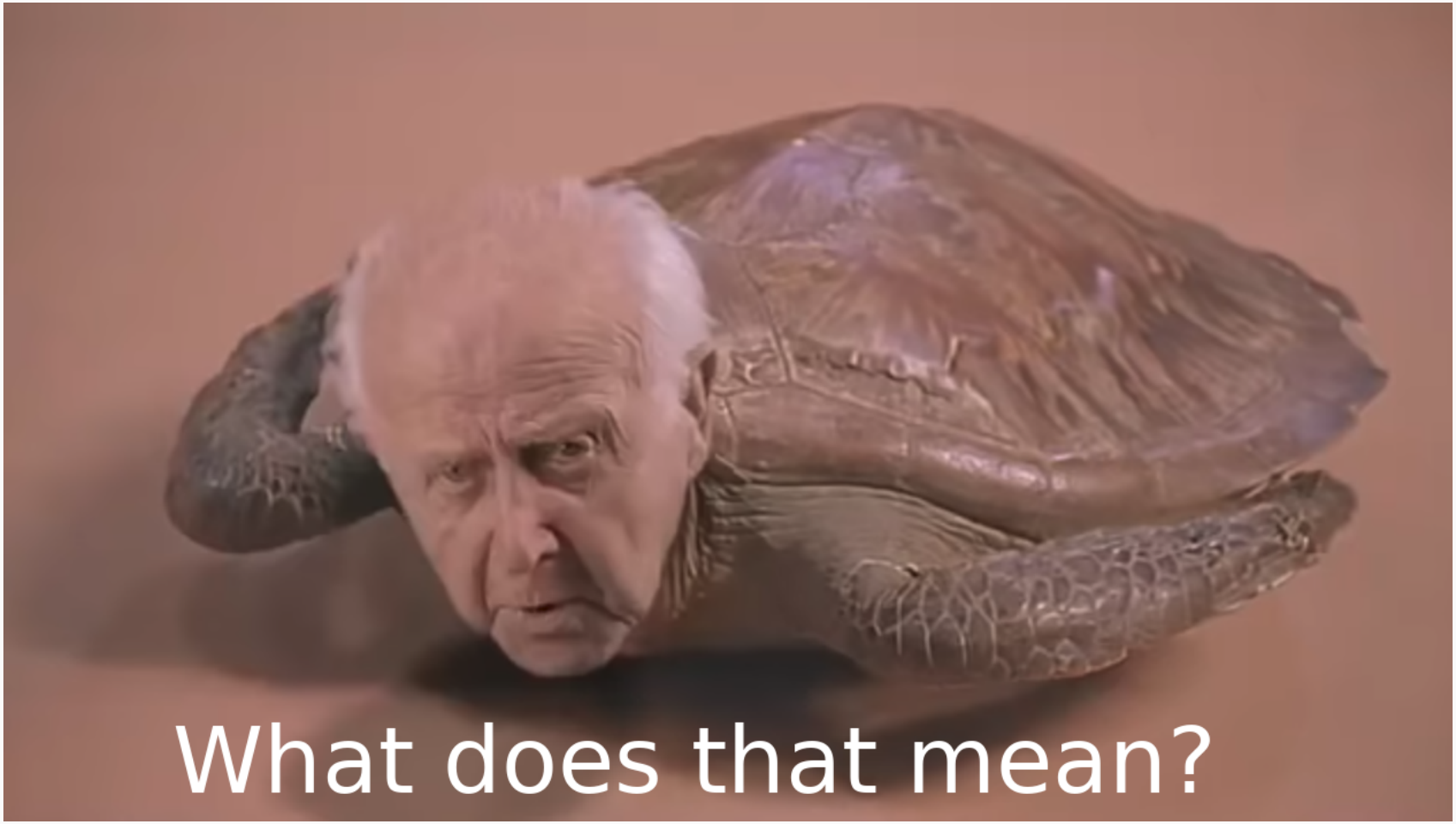


Non-IID noise (heteroscedasticity)

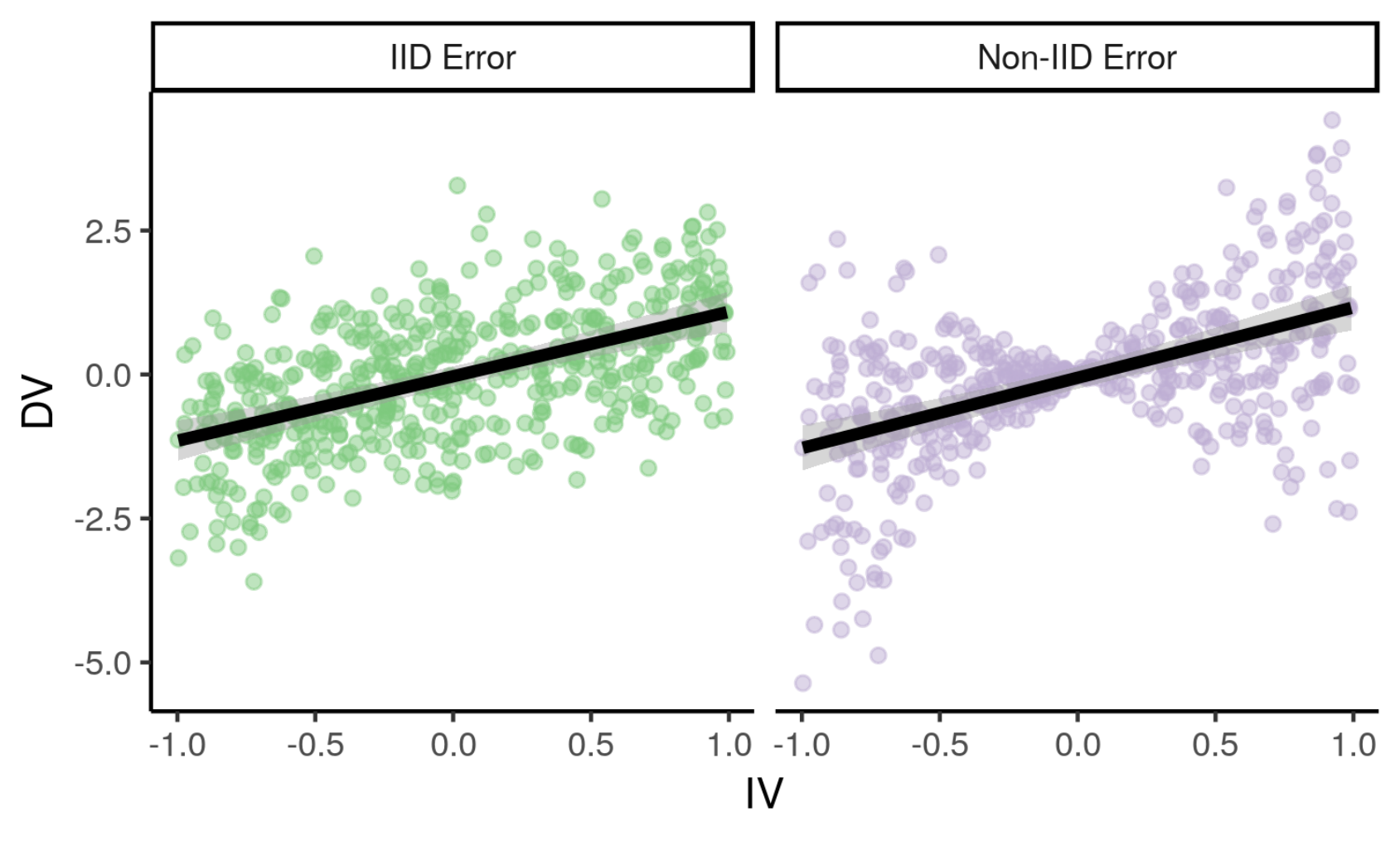
Consequences [edit]

One of the assumptions of the classical linear regression model is that there is no heteroscedasticity. Breaking this assumption means that the Gauss–Markov theorem does not apply, meaning that OLS estimators are not the Best Linear Unbiased Estimators (BLUE) and their variance is not the lowest of all other unbiased estimators. Heteroscedasticity does *not* cause ordinary least squares coefficient estimates to be biased, although it can cause ordinary least squares estimates of the variance (and, thus, standard errors) of the coefficients to be biased, possibly above or below the true or population variance. Thus, regression analysis using heteroscedastic data will still provide an unbiased estimate for the relationship between the predictor variable and the outcome, but standard errors and therefore inferences obtained from data analysis are suspect. Biased standard errors lead to biased inference, so results of hypothesis tests are possibly wrong. For example, if OLS is performed on a heteroscedastic data

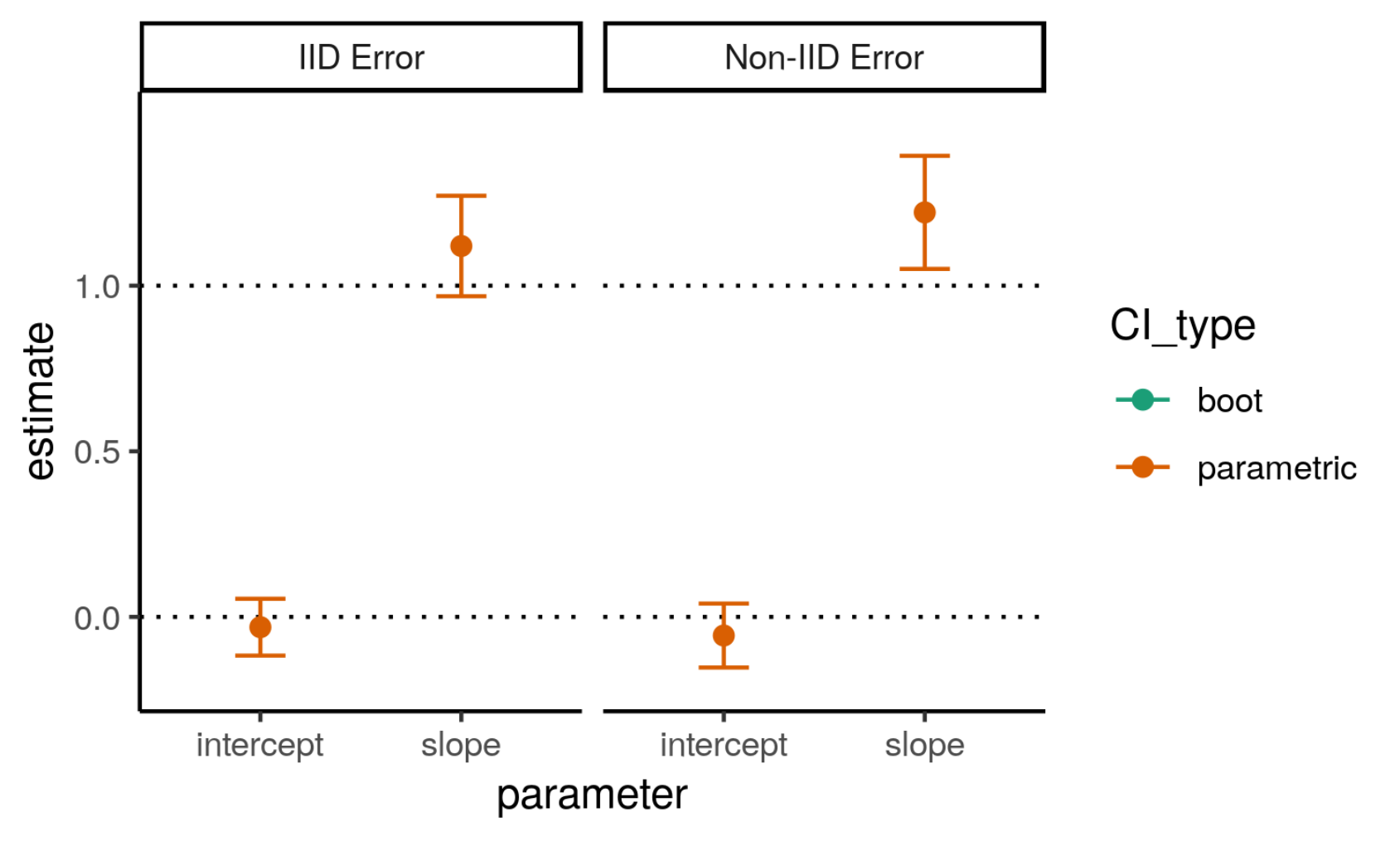
Non-IID noise (heteroscedasticity)



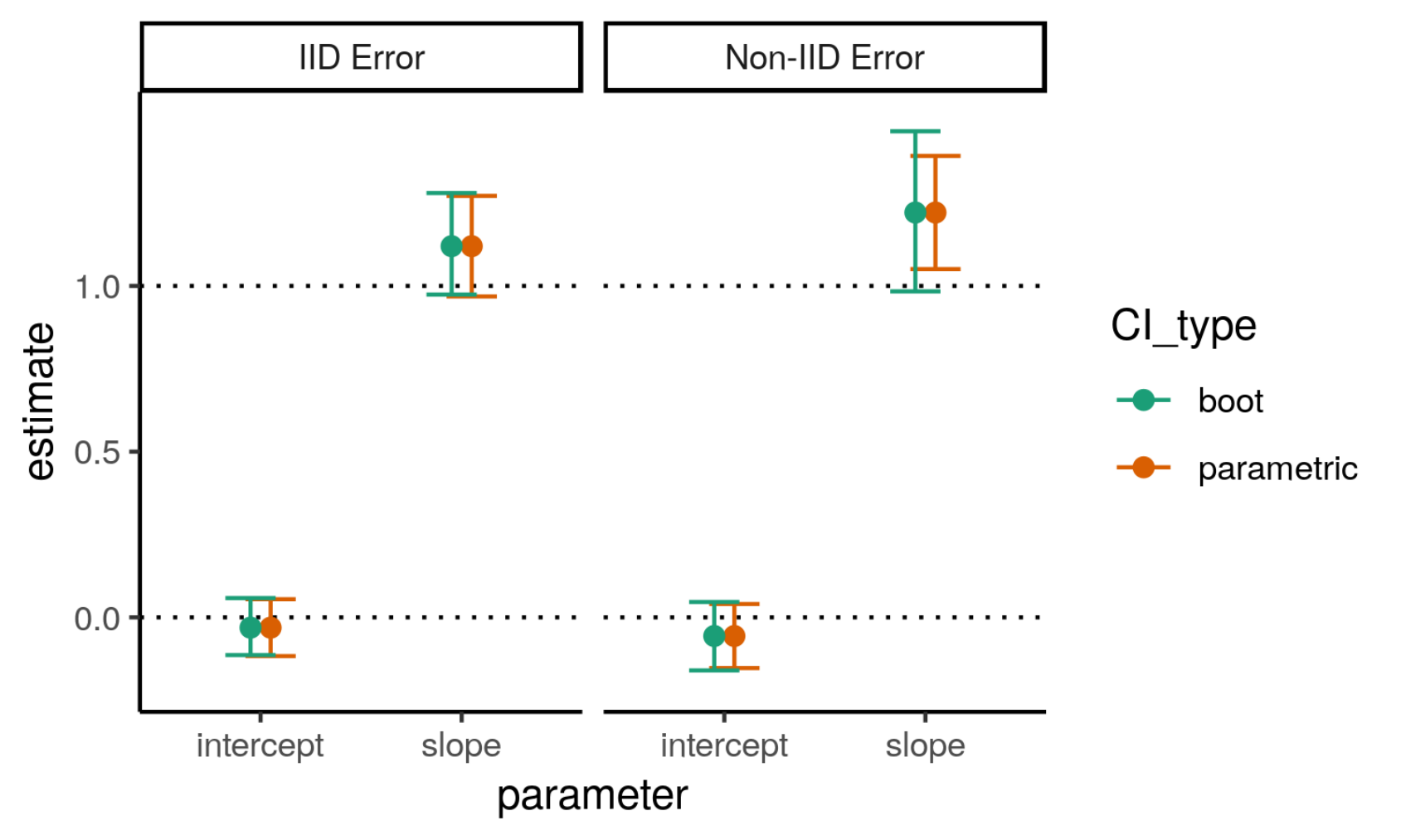
Non-IID noise (heteroscedasticity)



Non-IID noise (heteroscedasticity)

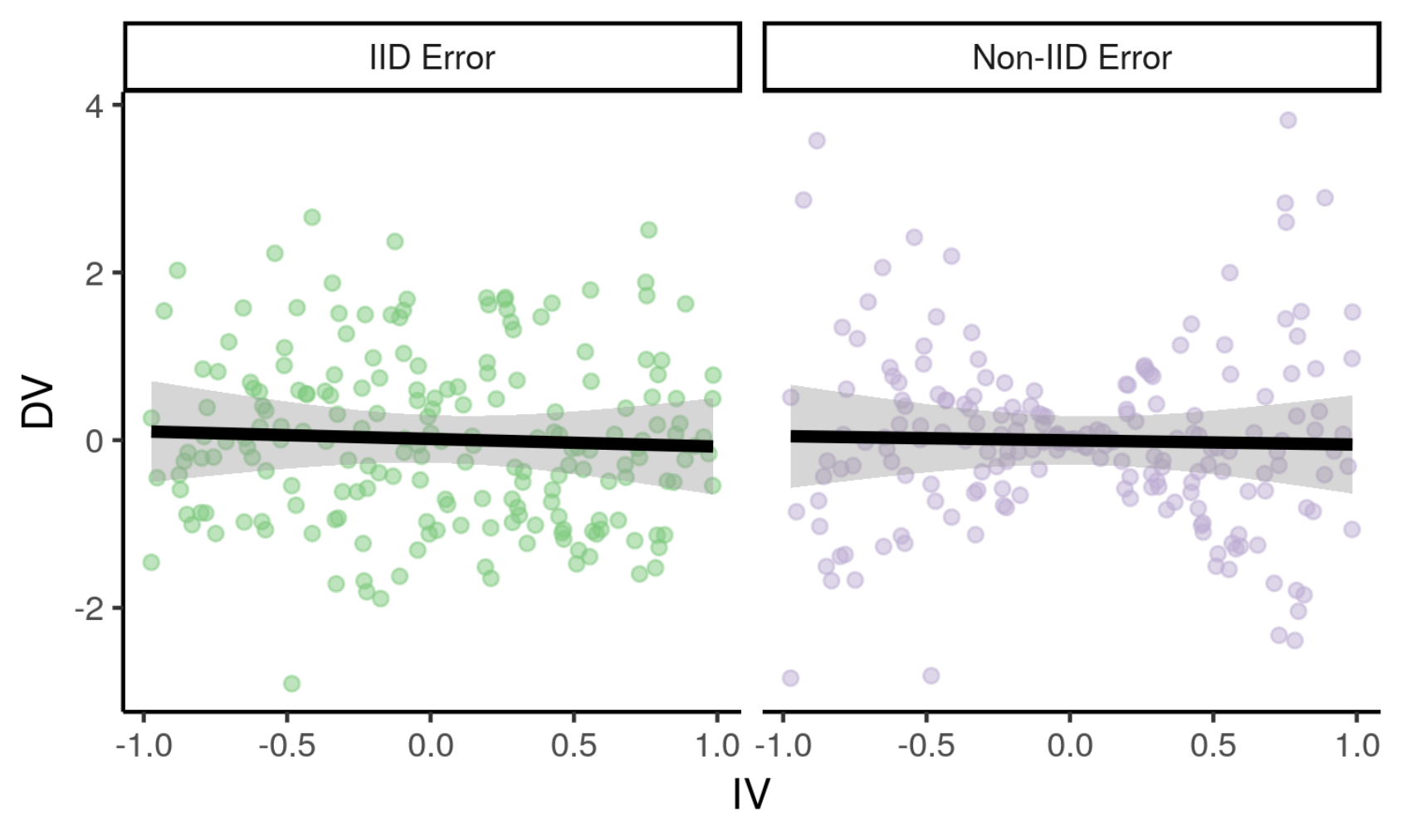


Non-IID noise (heteroscedasticity)



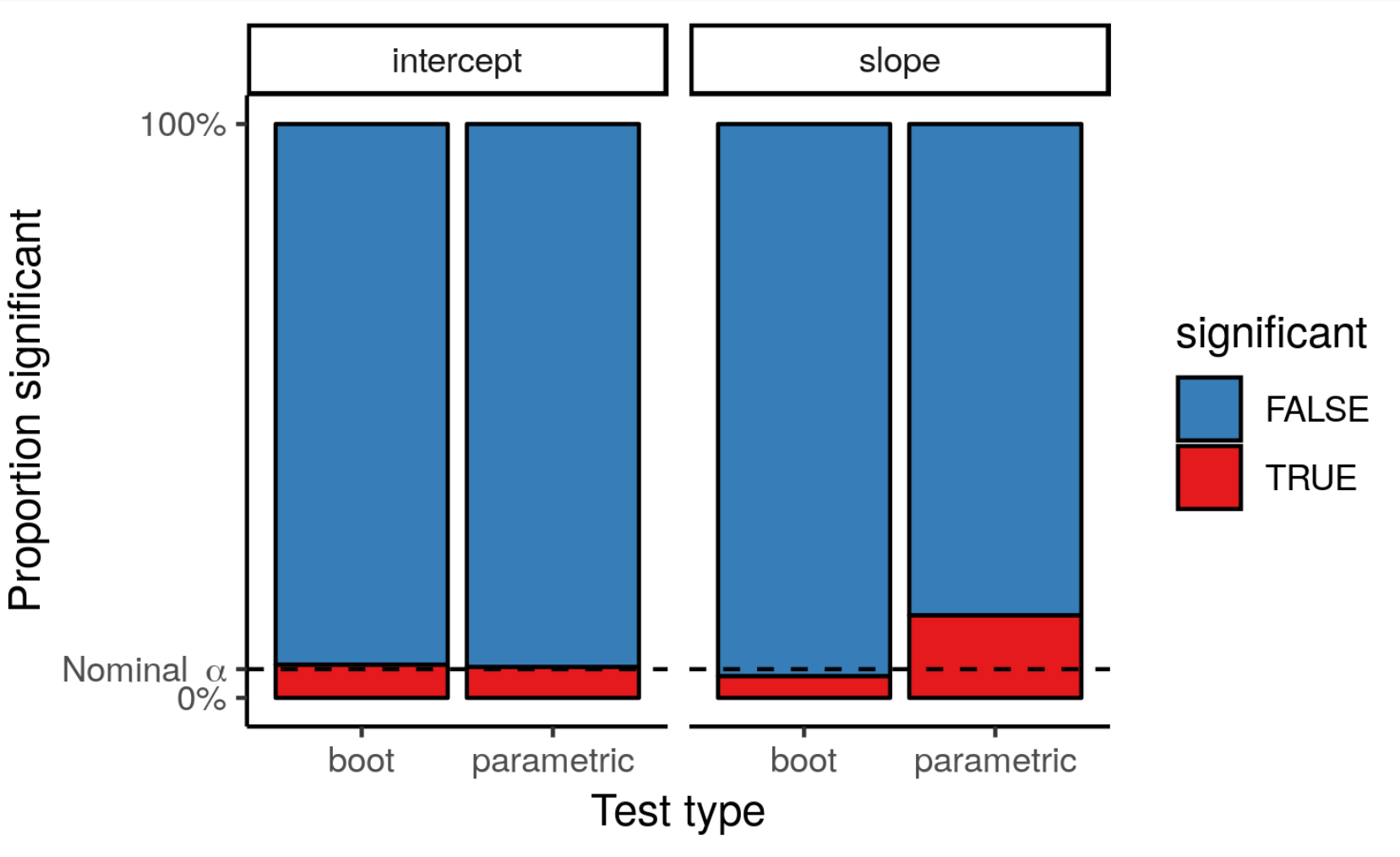
Non-IID noise (heteroscedasticity)

What if the null is true?



Non-IID noise (heteroscedasticity)

What if the null is true?

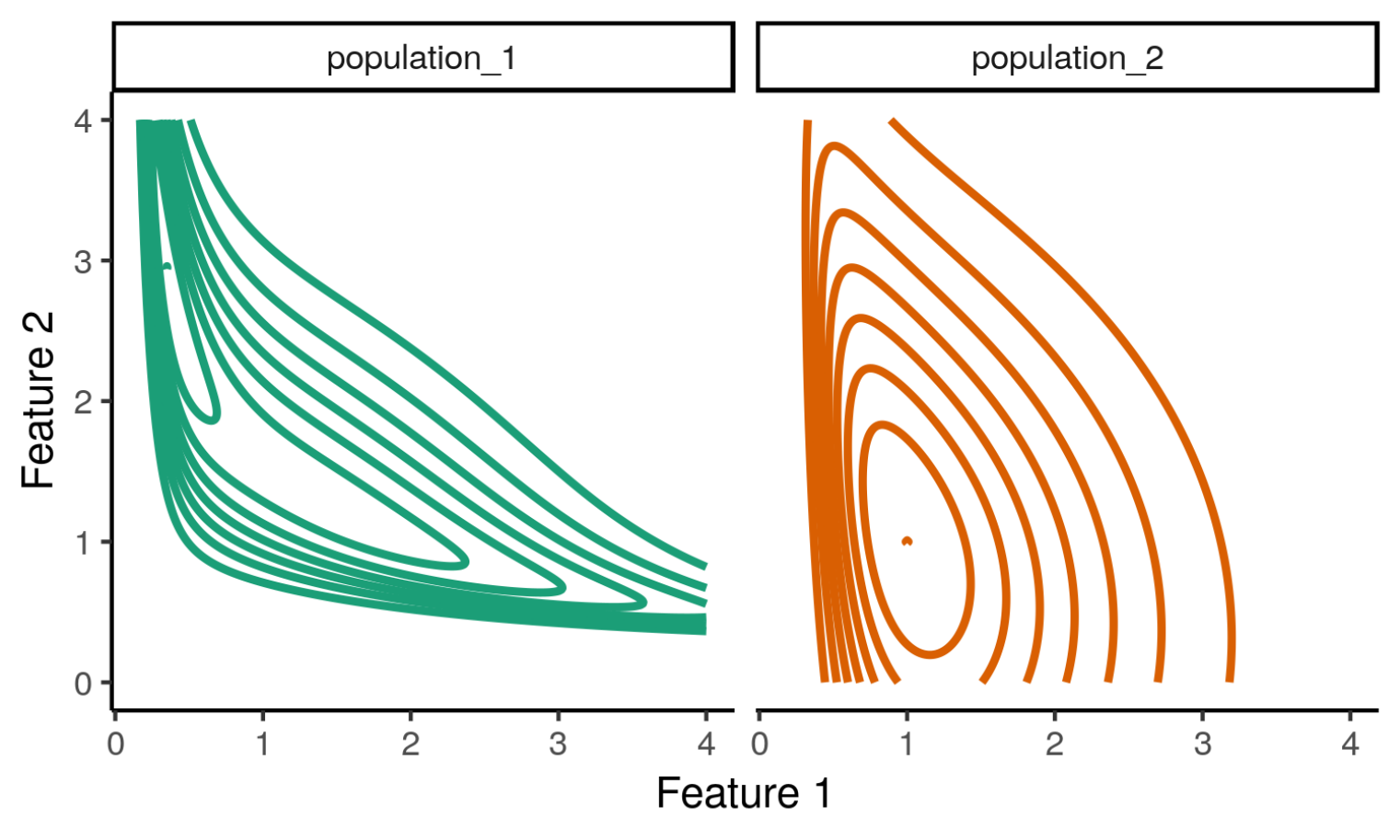


Non-IID noise (heteroscedasticity)

- Linear models assume IID noise.
- Therefore heteroscedasticity leads to incorrect confidence intervals **and** hypothesis tests.
- Bootstrapping avoids this assumption, and so behaves more appropriately if heteroscedasticity is present.

Bootstrapping makes fewer assumptions than parametric methods, so it works better when those assumptions don't hold.

A case study in weird tests



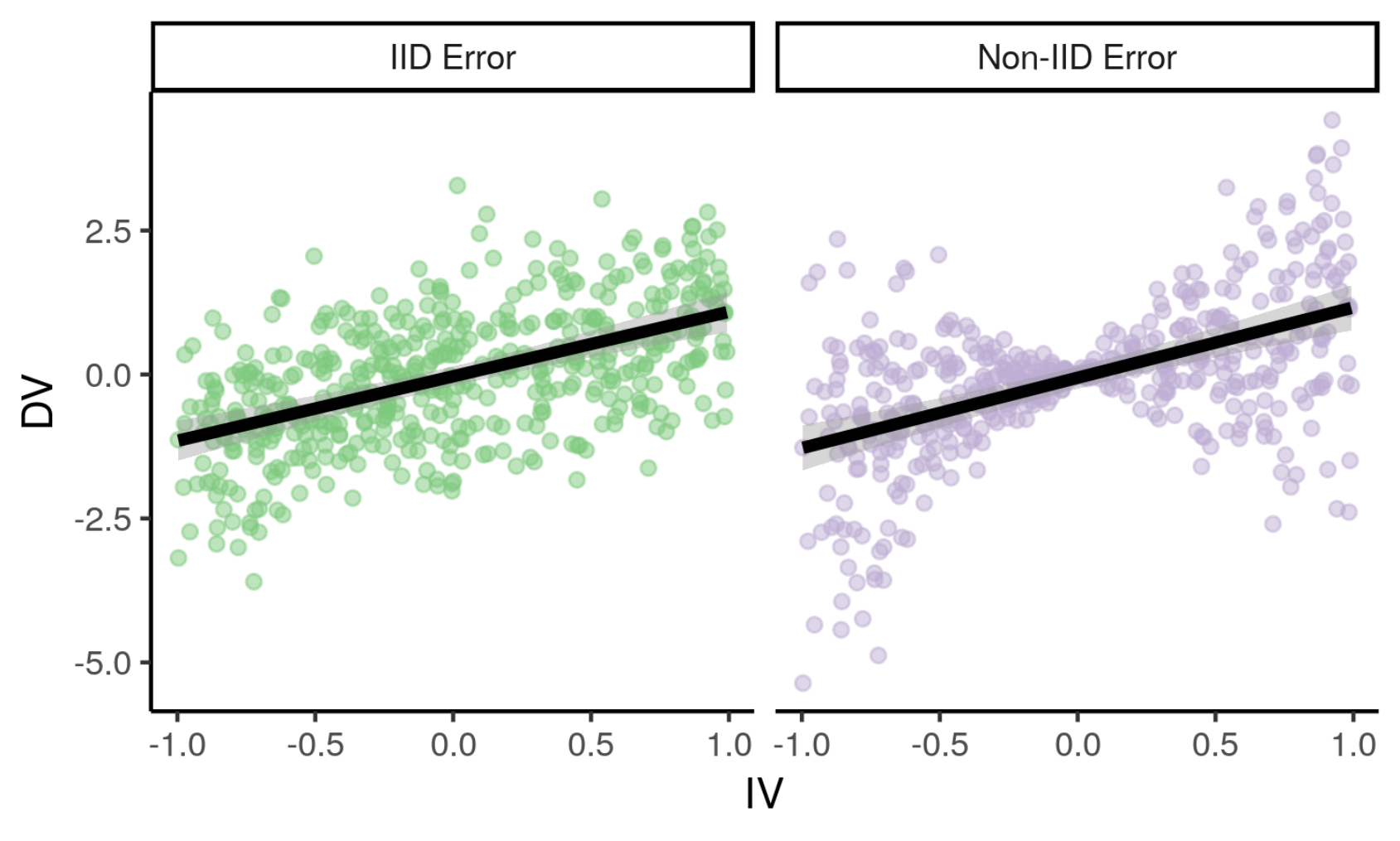
$$\text{Test statistic} = \frac{D_{\text{KL}}\left(P_1 || \frac{P_1+P_2}{2}\right) + D_{\text{KL}}\left(P_2 || \frac{P_1+P_2}{2}\right)}{2}$$

Bootstrapping makes fewer assumptions than parametric methods, so it can be applied in cases where parametric sampling distributions aren't known.

What is bootstrapping anyway?

Fundamental idea: Your best (or only) estimate of what's happening in the population is what's happening in your sample.

Non-IID noise (heteroscedasticity)



So instead of making assumptions about the population, we'll use what we actually know from the sample.

Bootstrap resampling

- We want to understand our uncertainty about a statistic we are estimating (e.g. a difference of means).

Bootstrap resampling

- We want to understand our uncertainty about a statistic we are estimating (e.g. a difference of means).
- The correct (frequentist) way to do this would be to run many experiments, and see how much that parameter varies, but usually this is infeasible.

Bootstrap resampling

- We want to understand our uncertainty about a statistic we are estimating (e.g. a difference of means).
- The correct (frequentist) way to do this would be to run many experiments, and see how much that parameter varies, but usually this is infeasible.
- Instead, we assume our sample distribution closely approximates the population distribution.

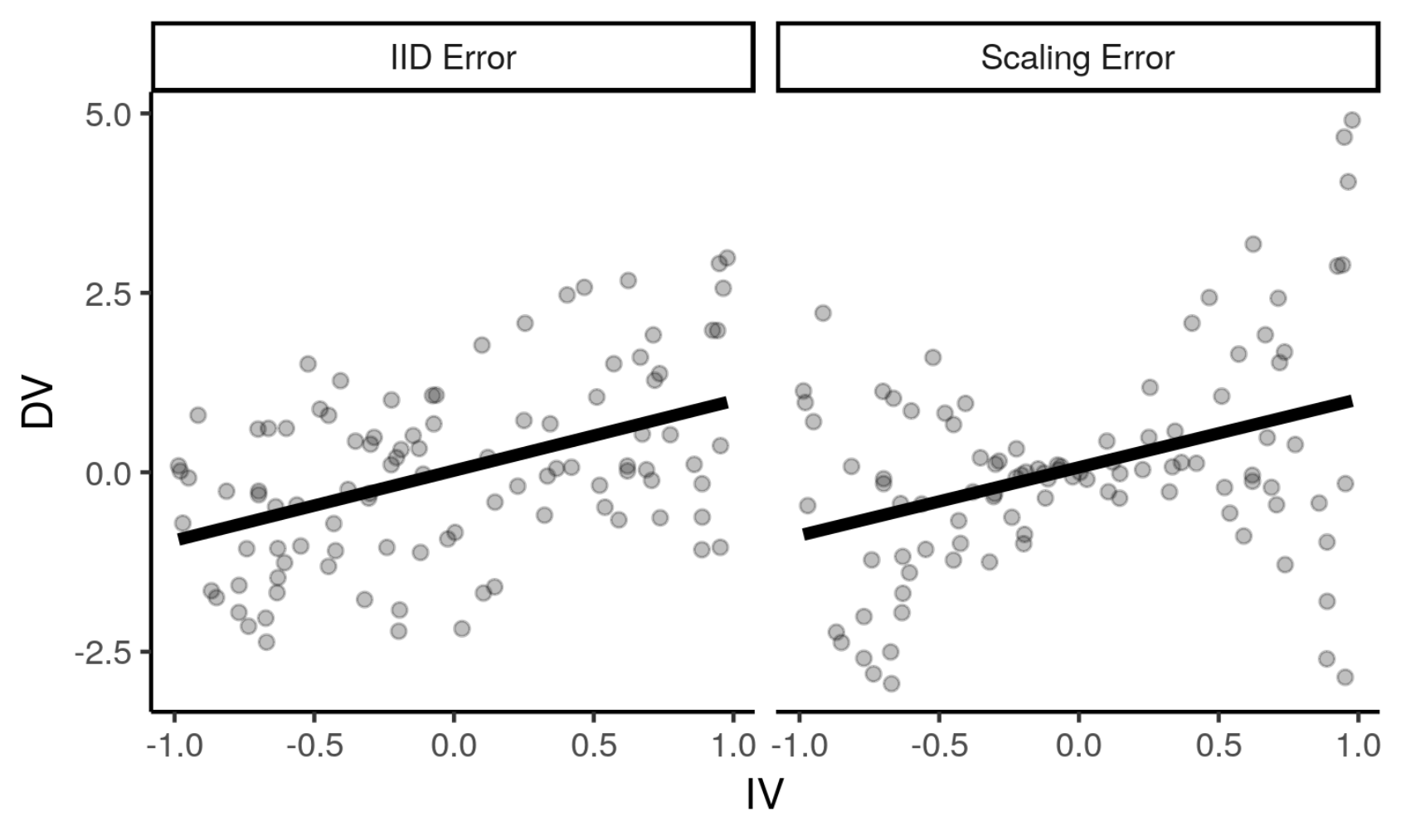
Bootstrap resampling

- We want to understand our uncertainty about a statistic we are estimating (e.g. a difference of means).
- The correct (frequentist) way to do this would be to run many experiments, and see how much that parameter varies, but usually this is infeasible.
- Instead, we assume our sample distribution closely approximates the population distribution.
- Therefore we can **simulate** these repeated experiments by **resampling from our sample with replacement.**

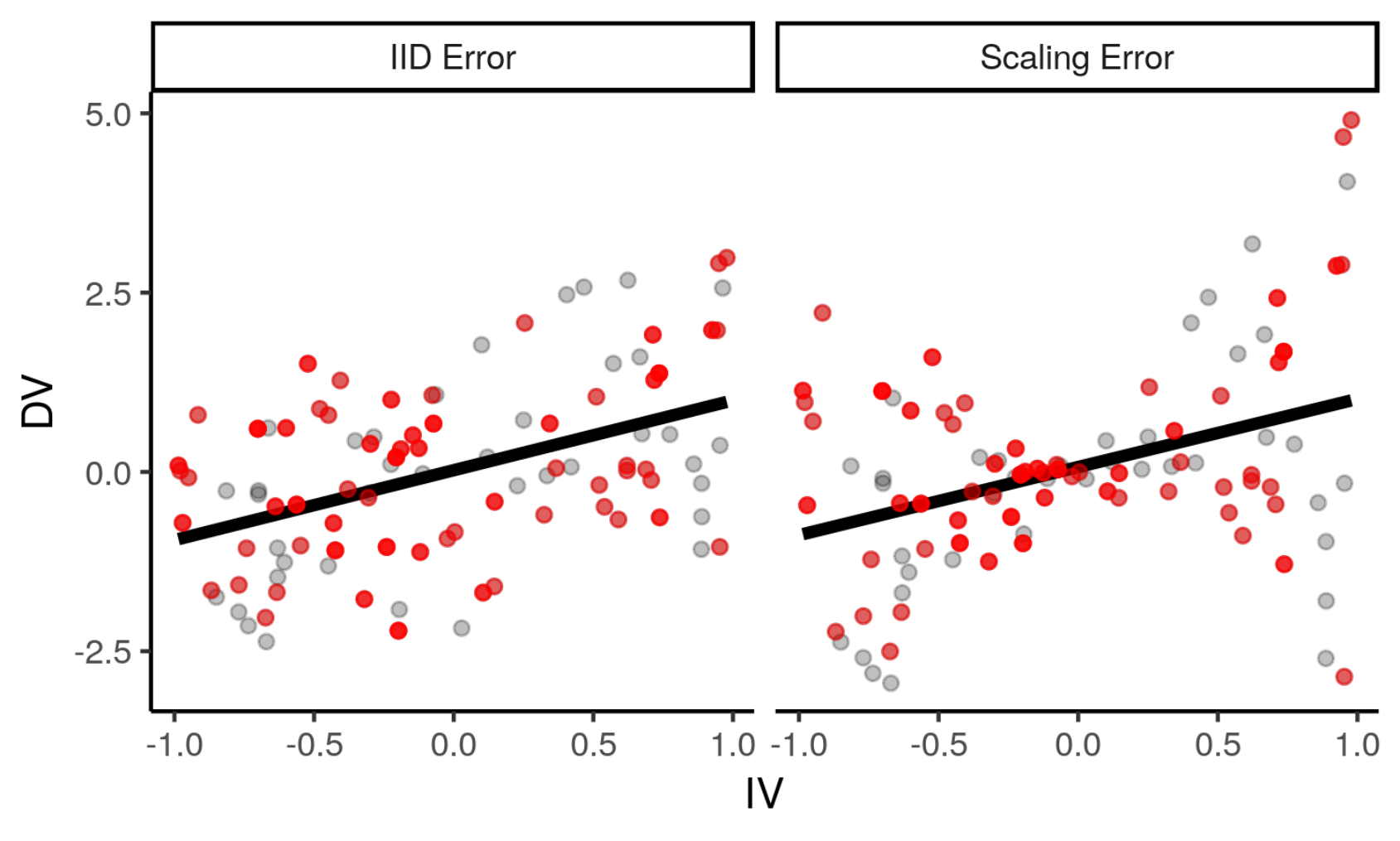
Bootstrap resampling demo I

http://wise.cgu.edu/portfolio/bootstrapping/

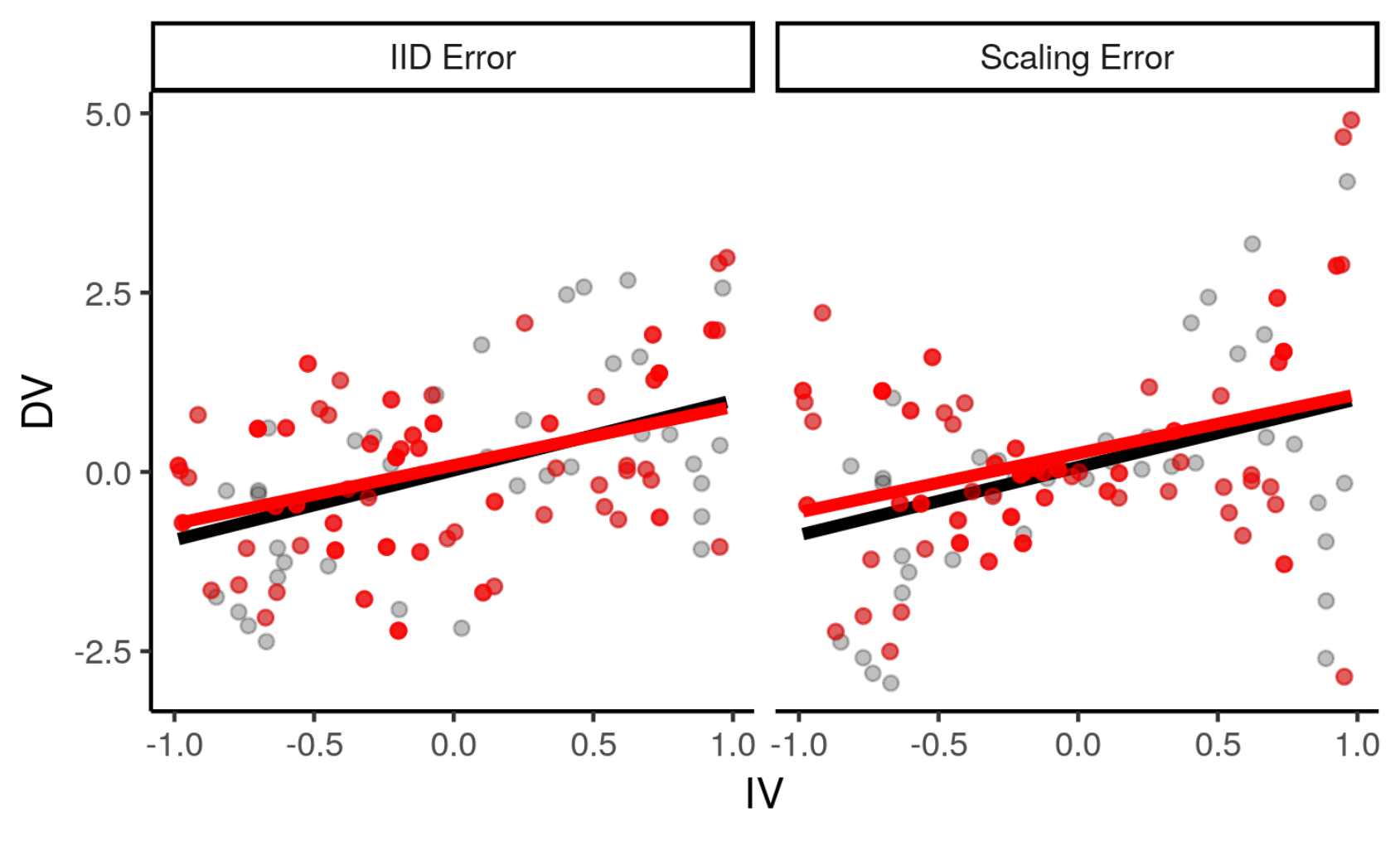
Bootstrap resampling demo II



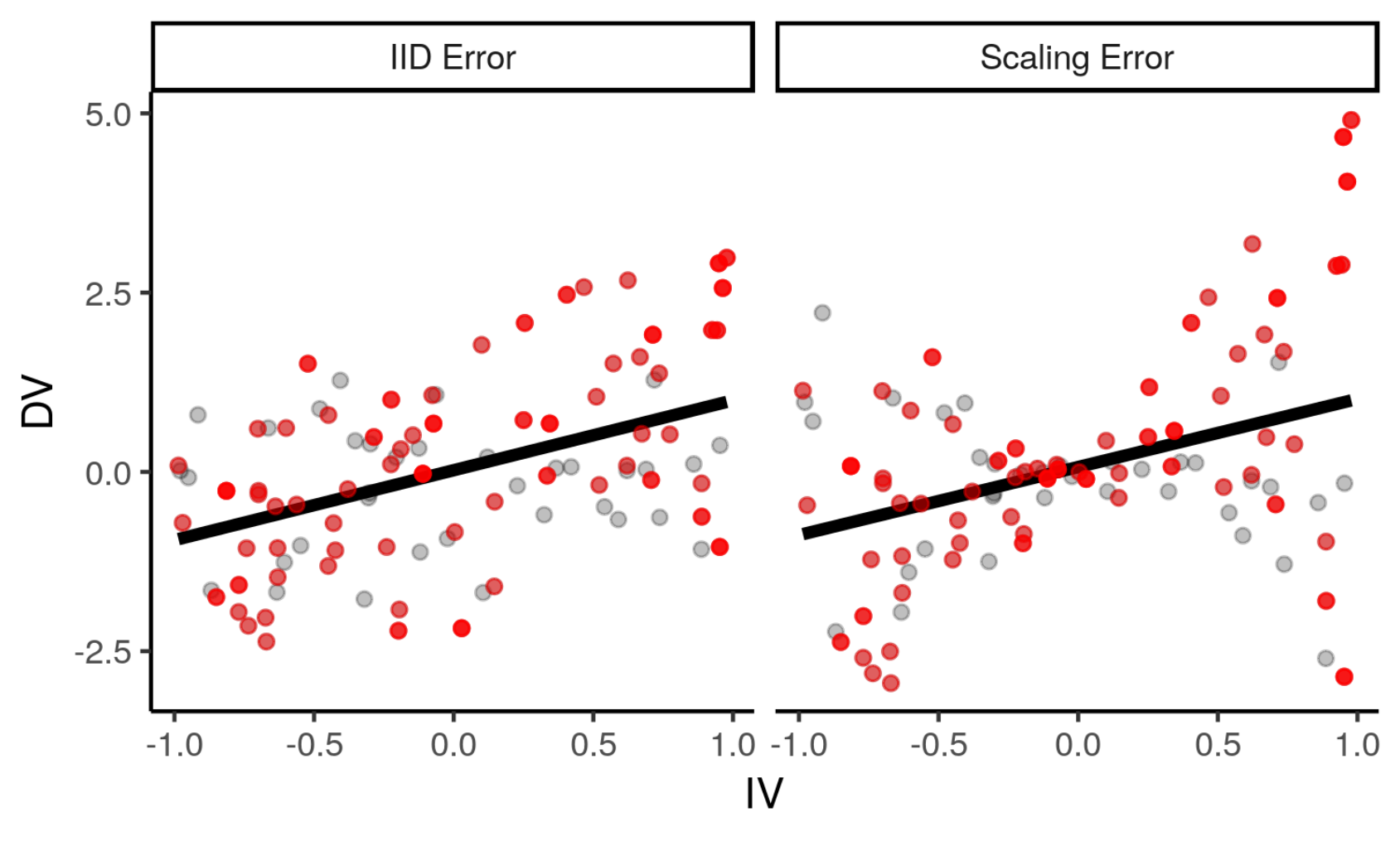
Bootstrap resampling demo II



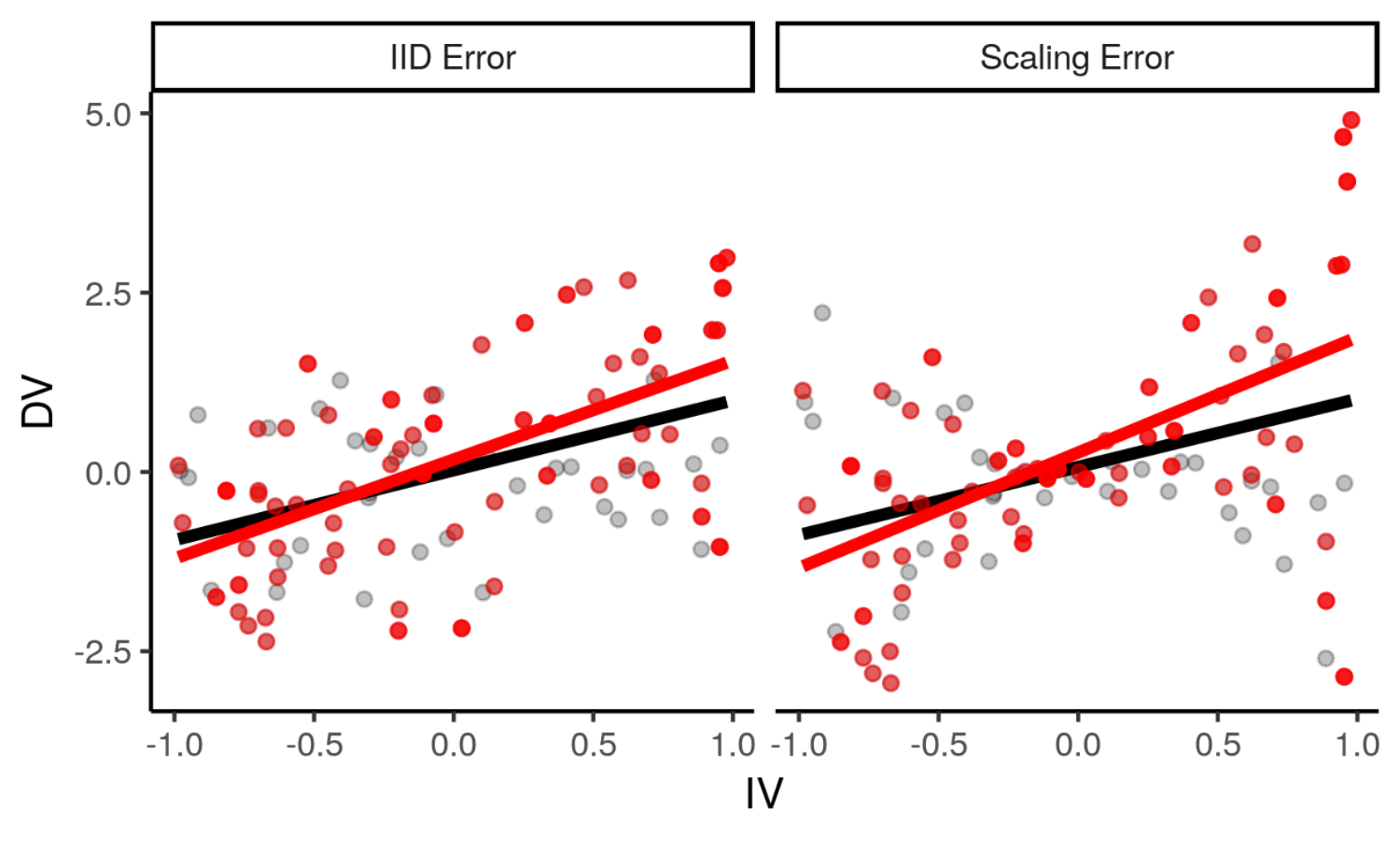
Bootstrap resampling demo II



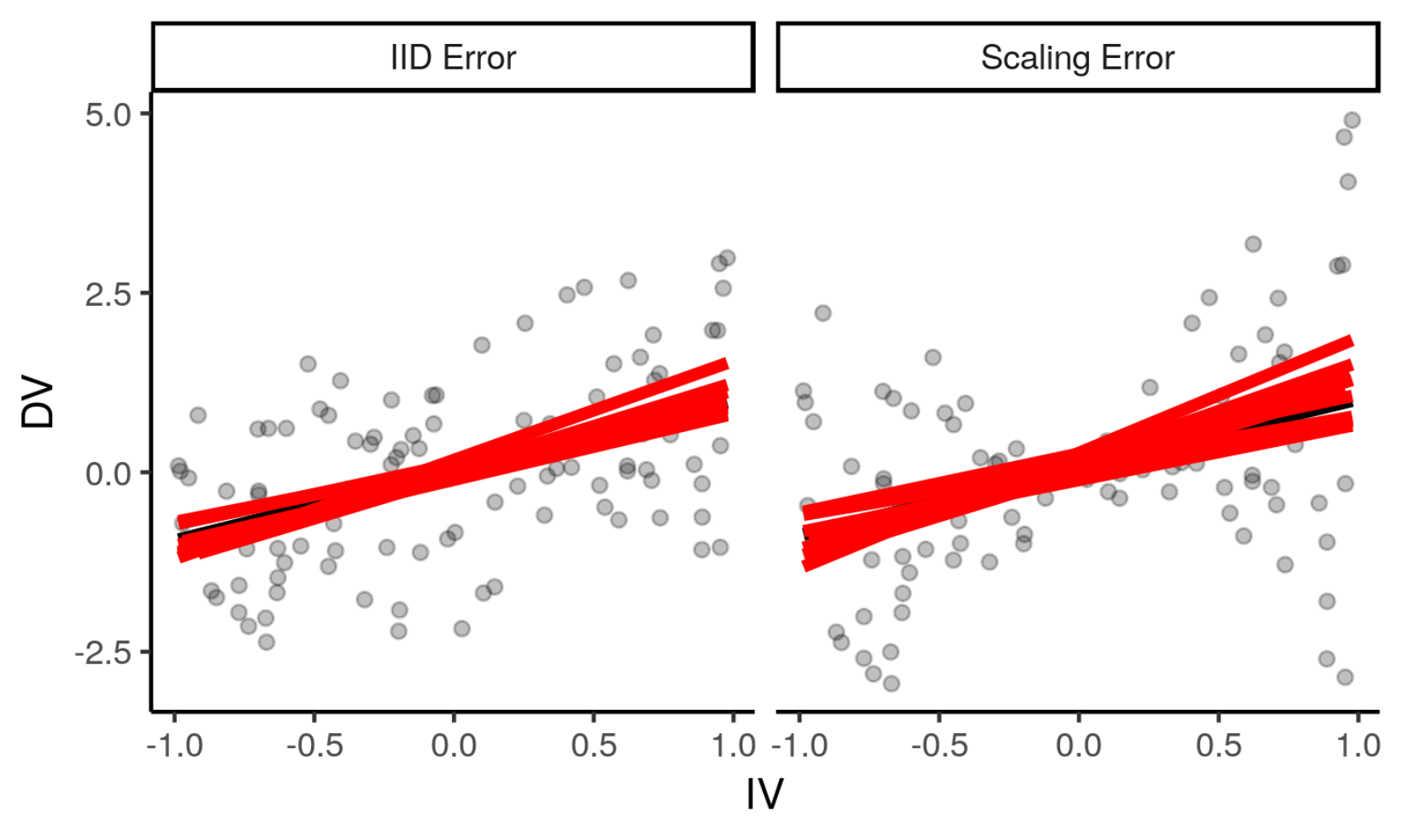
Bootstrap resampling demo II



Bootstrap resampling demo II



Bootstrap resampling demo II



We'll estimate a sampling distribution for our statistic of interest by resampling (with replacement) from our experimental sample.

Bootstrap confidence intervals

Bootstrap confidence intervals

A simple application:

Bootstrap confidence intervals

A simple application:

- We compute a statistic (such as the mean) from some data, and want to understand our uncertainty about it.

Bootstrap confidence intervals

A simple application:

- We compute a statistic (such as the mean) from some data, and want to understand our uncertainty about it.

- Remember: The bootstrap idea is that we can **simulate** these repeated experiments by **resampling from our sample with replacement.**

Bootstrap confidence intervals

A simple application:

- We compute a statistic (such as the mean) from some data, and want to understand our uncertainty about it.
- Remember: The bootstrap idea is that we can **simulate** these repeated experiments by **resampling from our sample with replacement.**
- Now, we'll use the sampling distribution that we "observe" in these "experiments" to construct a confidence interval.

Bootstrap confidence intervals

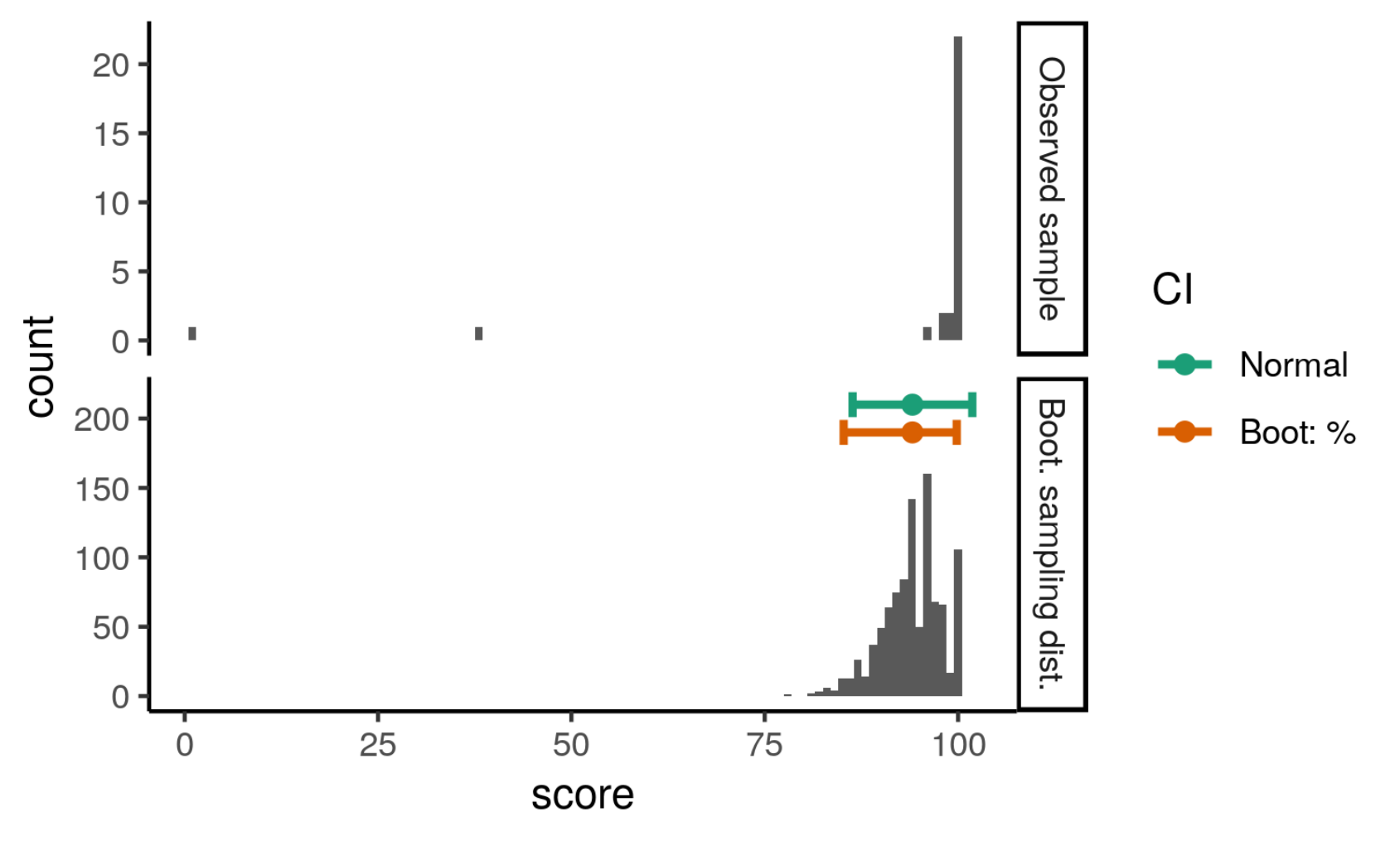
A simple application:

- We compute a statistic (such as the mean) from some data, and want to understand our uncertainty about it.
- Remember: The bootstrap idea is that we can **simulate** these repeated experiments by **resampling from our sample with replacement.**
- Now, we'll use the sampling distribution that we "observe" in these "experiments" to construct a confidence interval.
- The simplest method ("percentile") just uses the 2.5% and 97.5% quantiles of the bootstrap sampling distribution as the CI endpoints. (This is what `mean_cl_boot` does).

Bootstrap CI demo I

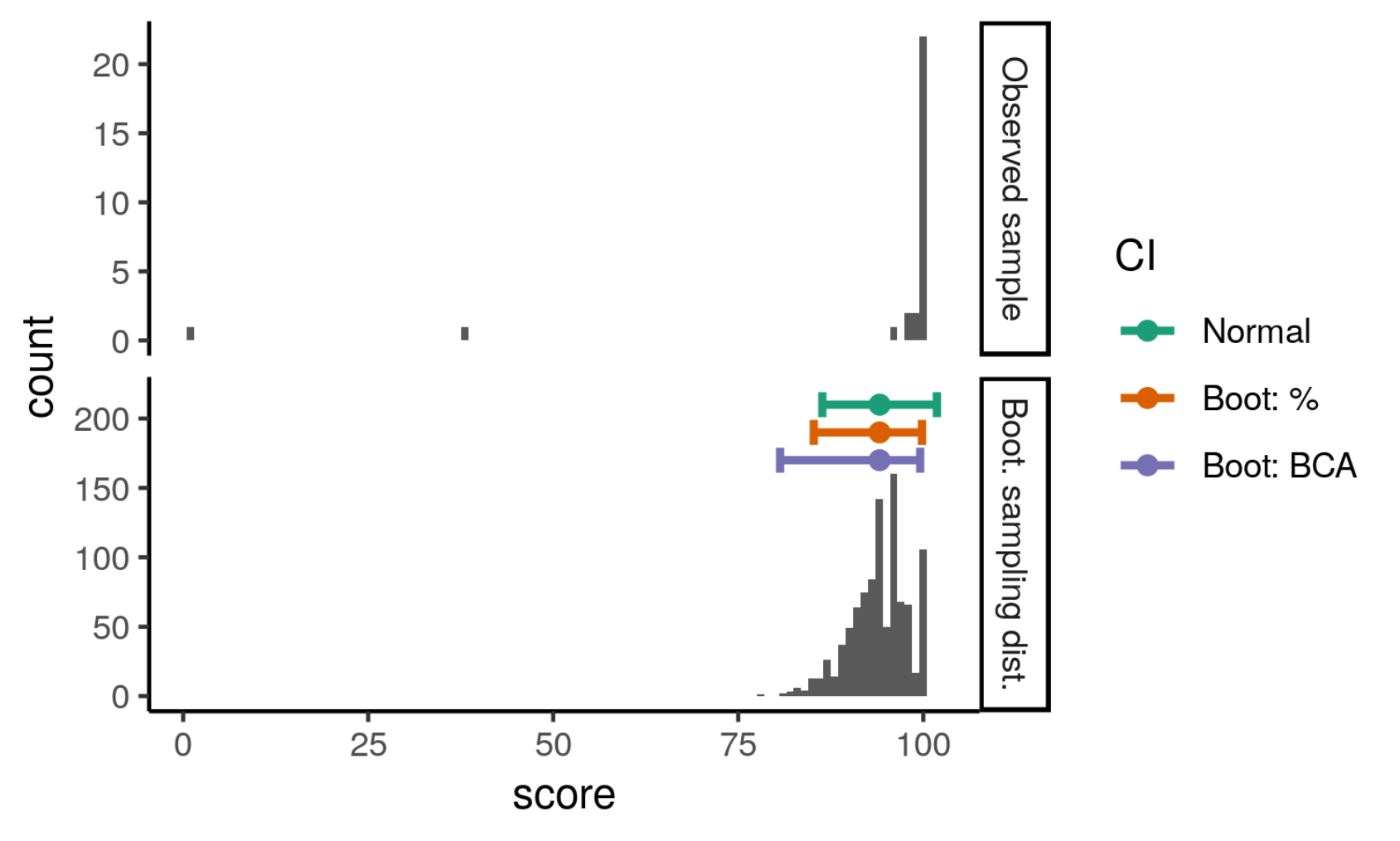
http://wise.cgu.edu/portfolio/bootstrapping/

Bootstrap CI demo II



To compute a bootstrap percentile CI, compute a bootstrap sampling distribution and take the 2.5% and 97.5% quantiles as the CI endpoints.

Bootstrap CI demo II



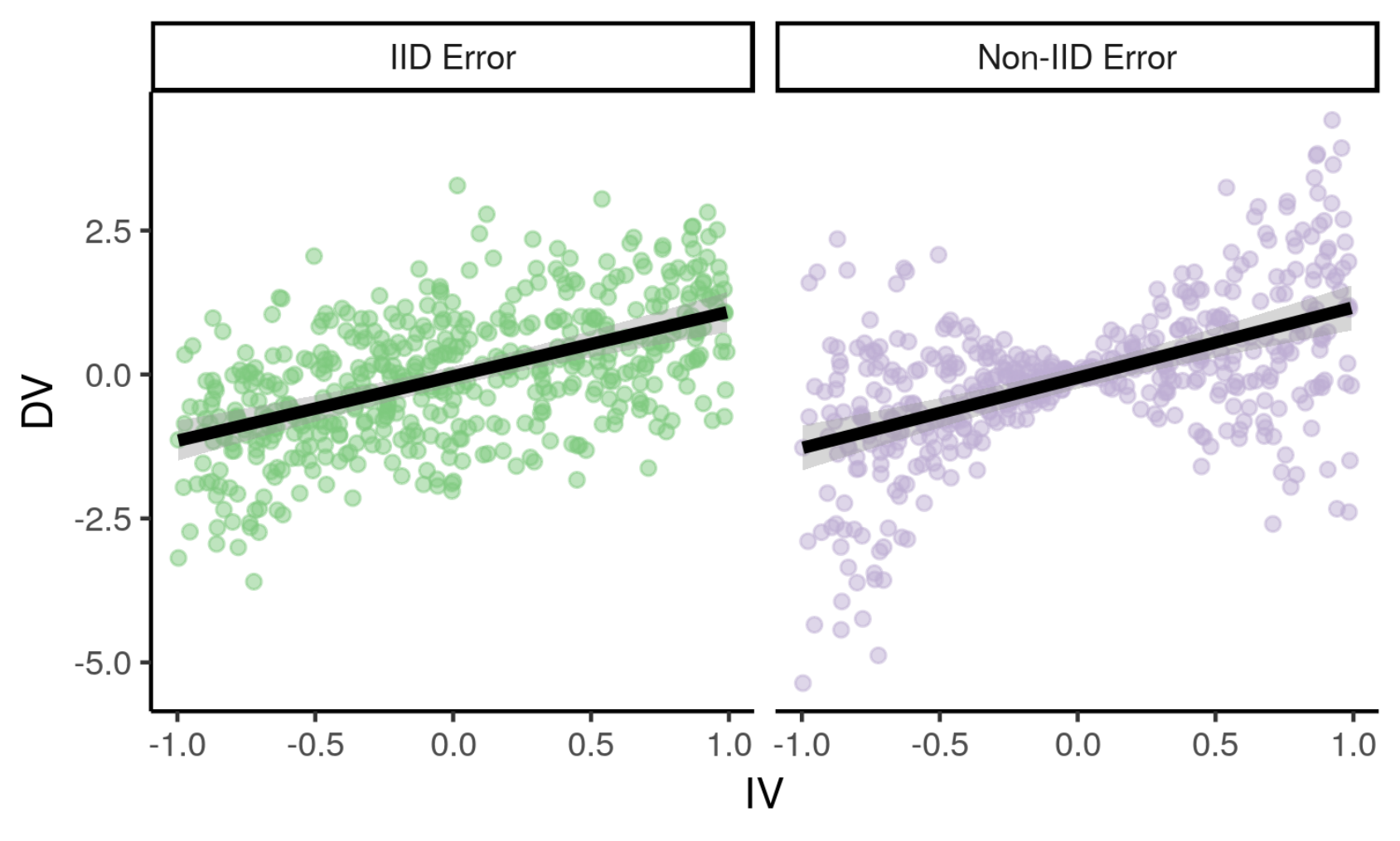
Bootstrap CI demo II

```r
get_mean_score = function(data, indices) {
  return(mean(data[indices,]$score))
}

bootstrap_results = boot(test_score_data, get_mean_score, R=1000)
bootstrap_CIs = boot.ci(bootstrap_results)
bootstrap_CIs
```

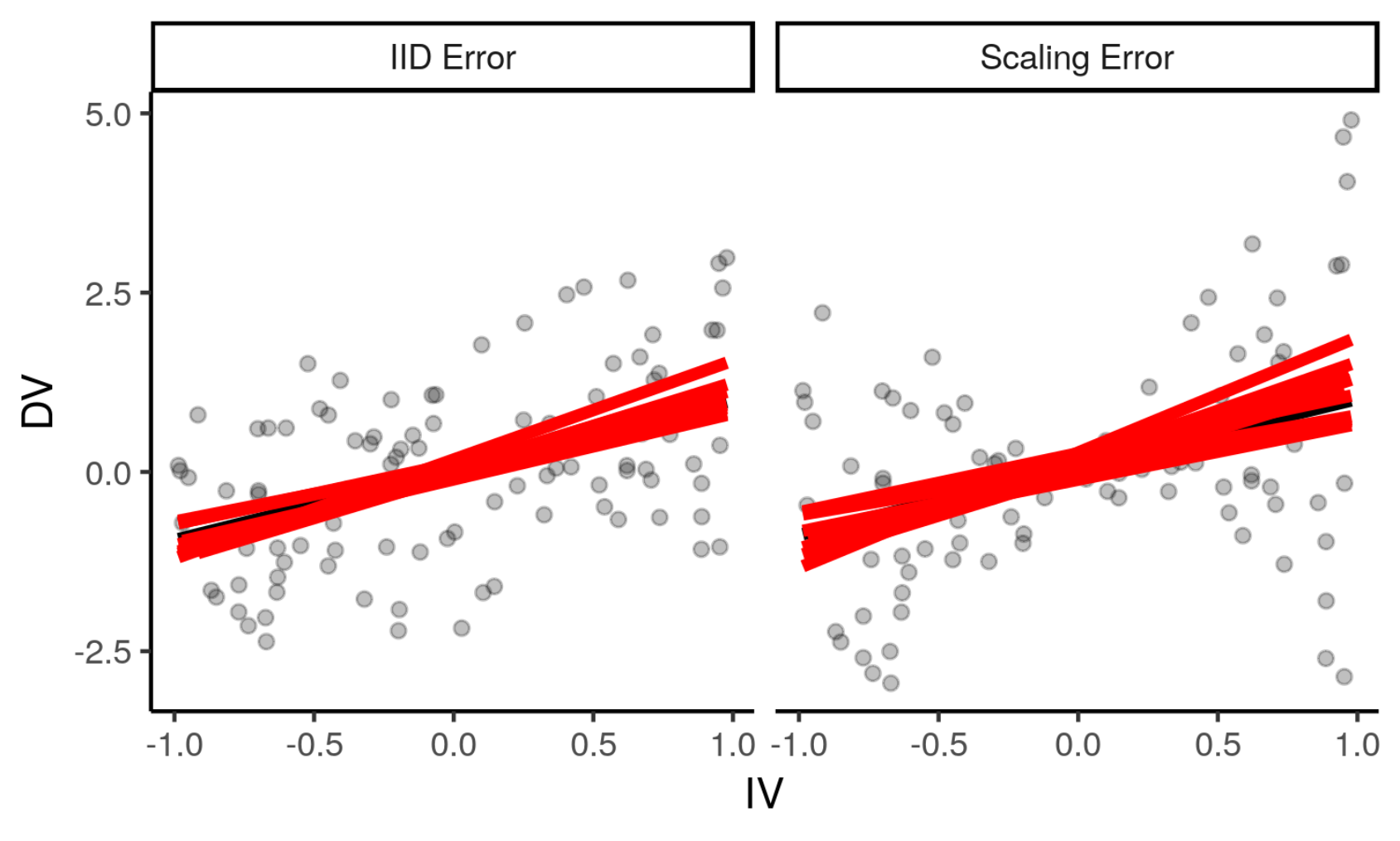
```
bootstrap variances needed for studentized intervalsBOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = bootstrap_results)

Intervals :
Level      Normal              Basic
95%   ( 86.53, 101.70 )   ( 88.38, 103.00 )

Level     Percentile            BCa
95%   (85.21, 99.83 )   (80.66, 99.59 )
Calculations and Intervals on Original Scale
Some BCa intervals may be unstable
```

Generally BCA intervals are superior. You can get them from the boot library in R!

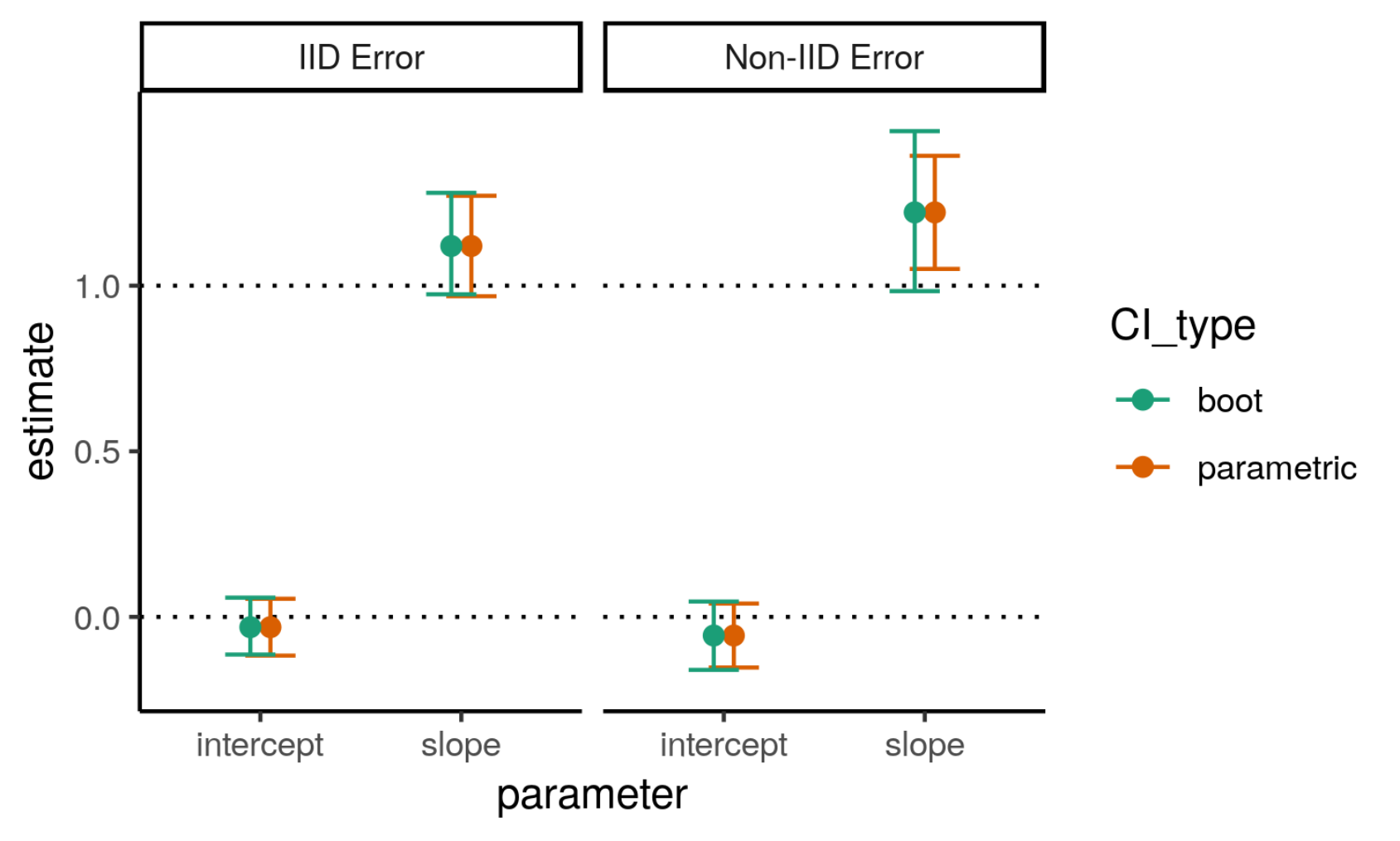
Non-IID noise (heteroscedasticity)



Bootstrap resampling demo II



Non-IID noise (heteroscedasticity)



Bootstrap (and permutation) tests

Estimation is at least as important as testing!

... but sometimes we do need a test for:

Bootstrap hypothesis tests

How can we use bootstrapping for hypothesis testing?

Bootstrap hypothesis tests

How can we use bootstrapping for hypothesis testing?

- Construct a bootstrap $1 - \alpha$ confidence interval.

Bootstrap hypothesis tests

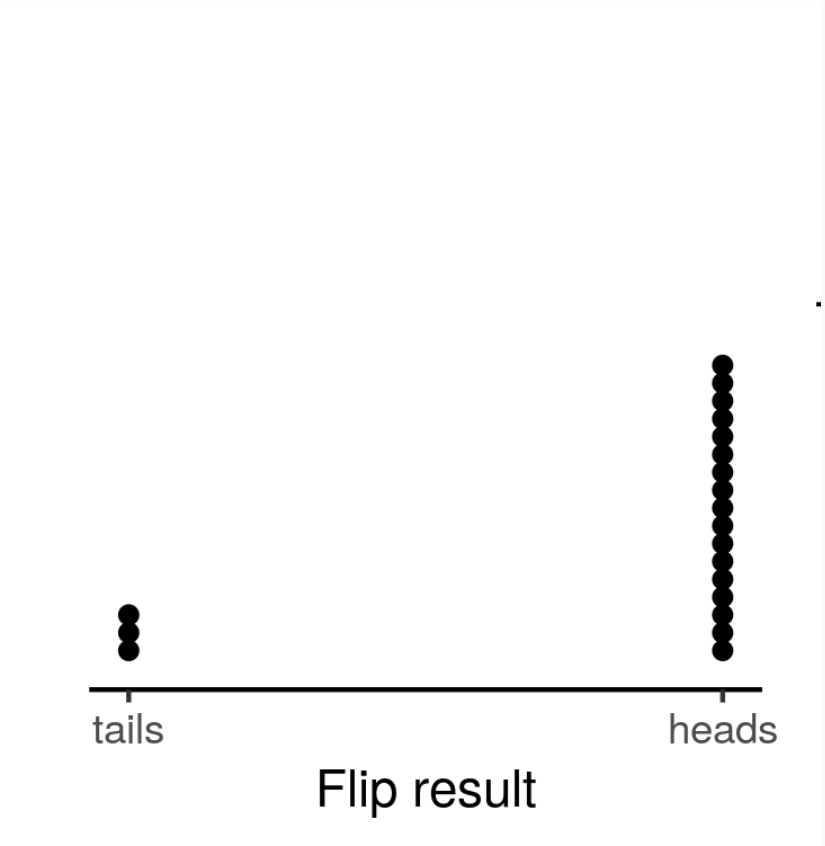
How can we use bootstrapping for hypothesis testing?

- Construct a bootstrap $1 - \alpha$ confidence interval.
- Reject the null if the null hypothesis value of the statistic isn't included in this CI.

Bootstrap hypothesis tests

How can we use bootstrapping for hypothesis testing?

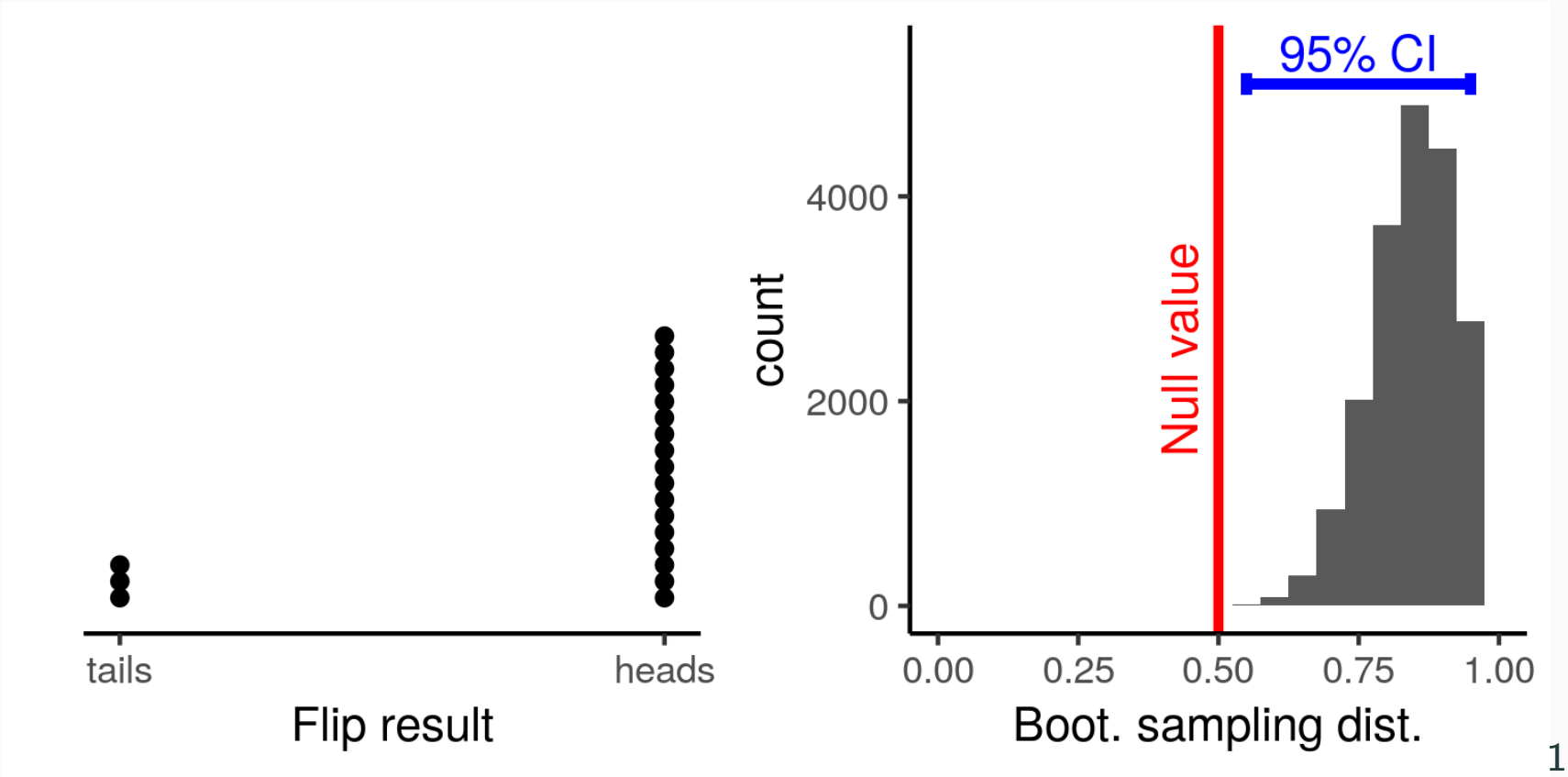
- Construct a bootstrap $1 - \alpha$ confidence interval.
- Reject the null if the null hypothesis value of the statistic isn't included in this CI.



Bootstrap hypothesis tests

How can we use bootstrapping for hypothesis testing?

- Construct a bootstrap $1 - \alpha$ confidence interval.
- Reject the null if the null hypothesis value of the statistic isn't included in this CI.

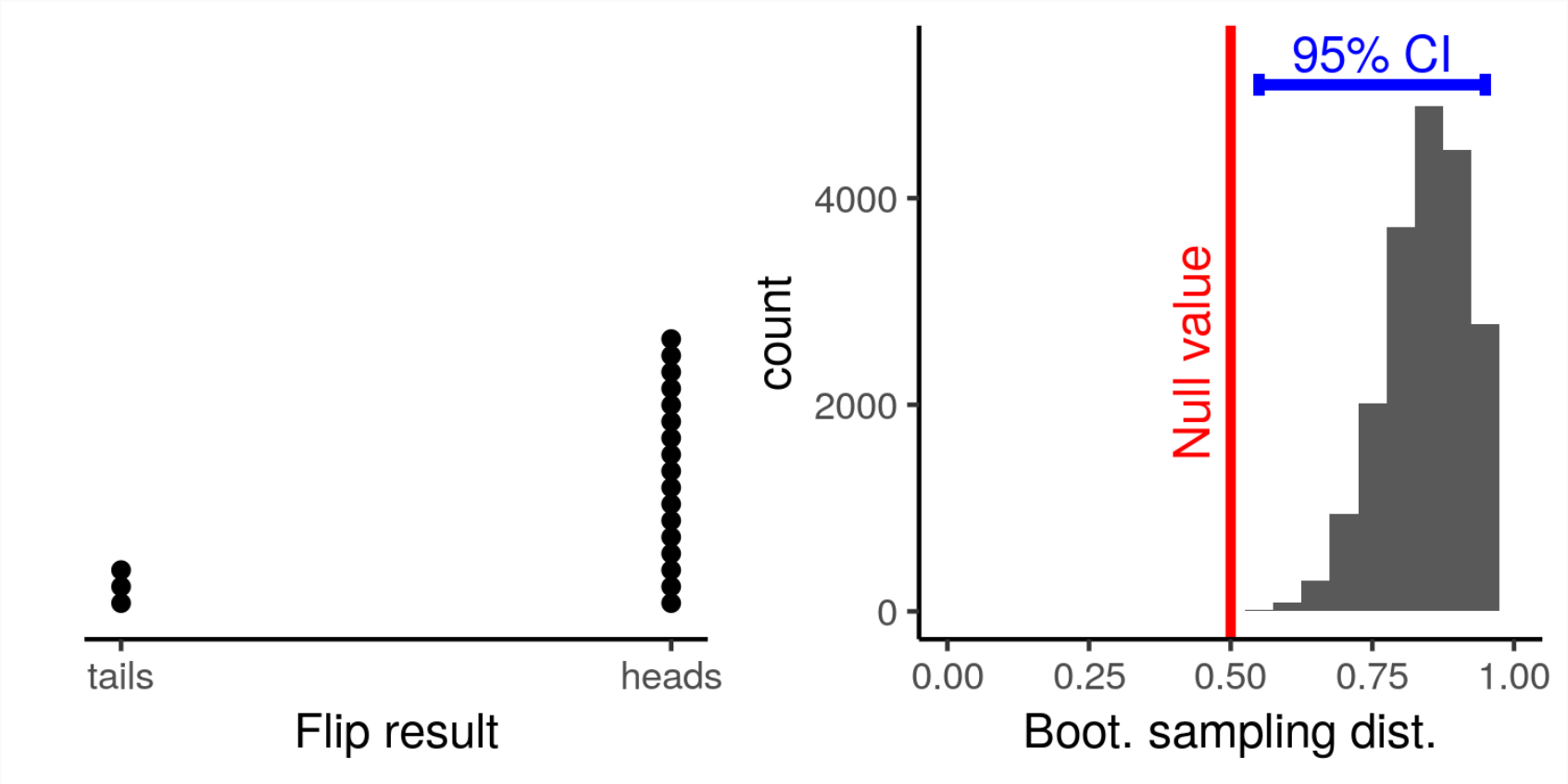


Bootstrap hypothesis testing: reject the null if the null value isn't contained in the 95%-CI.

Bootstrap p-values

How can we get a p-value?

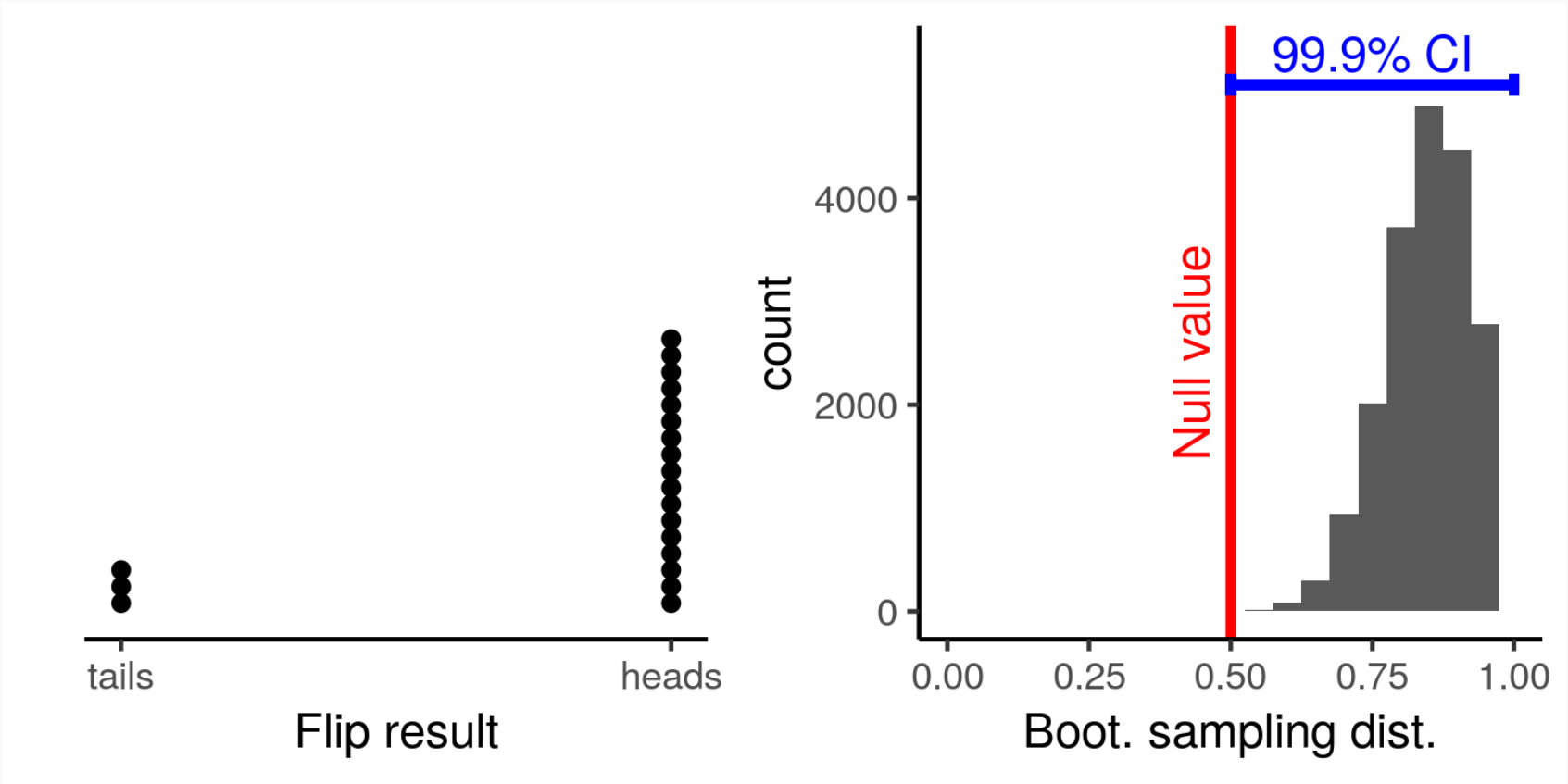
- Find α so that the null value is just at the edge of the $(1 - \alpha)$ CI, and let $p = \alpha$.



Bootstrap p-values

How can we get a p-value?

- Find α so that the null value is just at the edge of the $(1-\alpha)$ CI, and let $p = \alpha$.



Bootstrap p-values: p is the min α such that the null value isn't contained in the $(1-\alpha)$ CI.

Bootstrap tests & p-values: some food for thought

There are some subtle conceptual issues here:

Bootstrap tests & p-values: some food for thought

There are some subtle conceptual issues here:

- In NHST, we calculate the sampling distribution **under the null**, and see whether **our observed statistic** is surprising.

Bootstrap tests & p-values: some food for thought

There are some subtle conceptual issues here:

- In NHST, we calculate the sampling distribution **under the null**, and see whether **our observed statistic** is surprising.
- Here, we calculate the (approximate) sampling distribution under the **(approximate) population distribution,** and see whether the **null value** is surprising.

Bootstrap tests & p-values: some food for thought

There are some subtle conceptual issues here:

- In NHST, we calculate the sampling distribution **under the null**, and see whether **our observed statistic** is surprising.
- Here, we calculate the (approximate) sampling distribution under the **(approximate) population distribution,** and see whether the **null value** is surprising.
- In many cases (such as anytime the null is true, or if the assumptions of e.g. linear regression hold), bootstrap tests are **exactly** standard NHST (at least asymptotically).

Bootstrap tests & p-values: some food for thought

There are some subtle conceptual issues here:

- In NHST, we calculate the sampling distribution **under the null**, and see whether **our observed statistic** is surprising.
- Here, we calculate the (approximate) sampling distribution under the **(approximate) population distribution,** and see whether the **null value** is surprising.
- In many cases (such as anytime the null is true, or if the assumptions of e.g. linear regression hold), bootstrap tests are **exactly** standard NHST (at least asymptotically).
- More generally, the bootstrap testing procedure **is valid**, in the sense that it has the nominal false-positive rate, but the interpretation of the p-value is not standard.

Bootstrap p-values: conceptually tricky, but so are regular frequentist ones, and reviewers probably won't understand either well enough to argue with you.

Permutation tests

- Sometimes we **can** actually nonparametrically sample from the null.

Permutation tests

- Sometimes we **can** actually nonparametrically sample from the null.

- In a two-sample t-test, the null is that the two groups really are samples from the same distribution.

Permutation tests

- Sometimes we **can** actually nonparametrically sample from the null.
- In a two-sample t-test, the null is that the two groups really are samples from the same distribution.
- We can actually sample from the null by just shuffling the labels.

Permutation tests

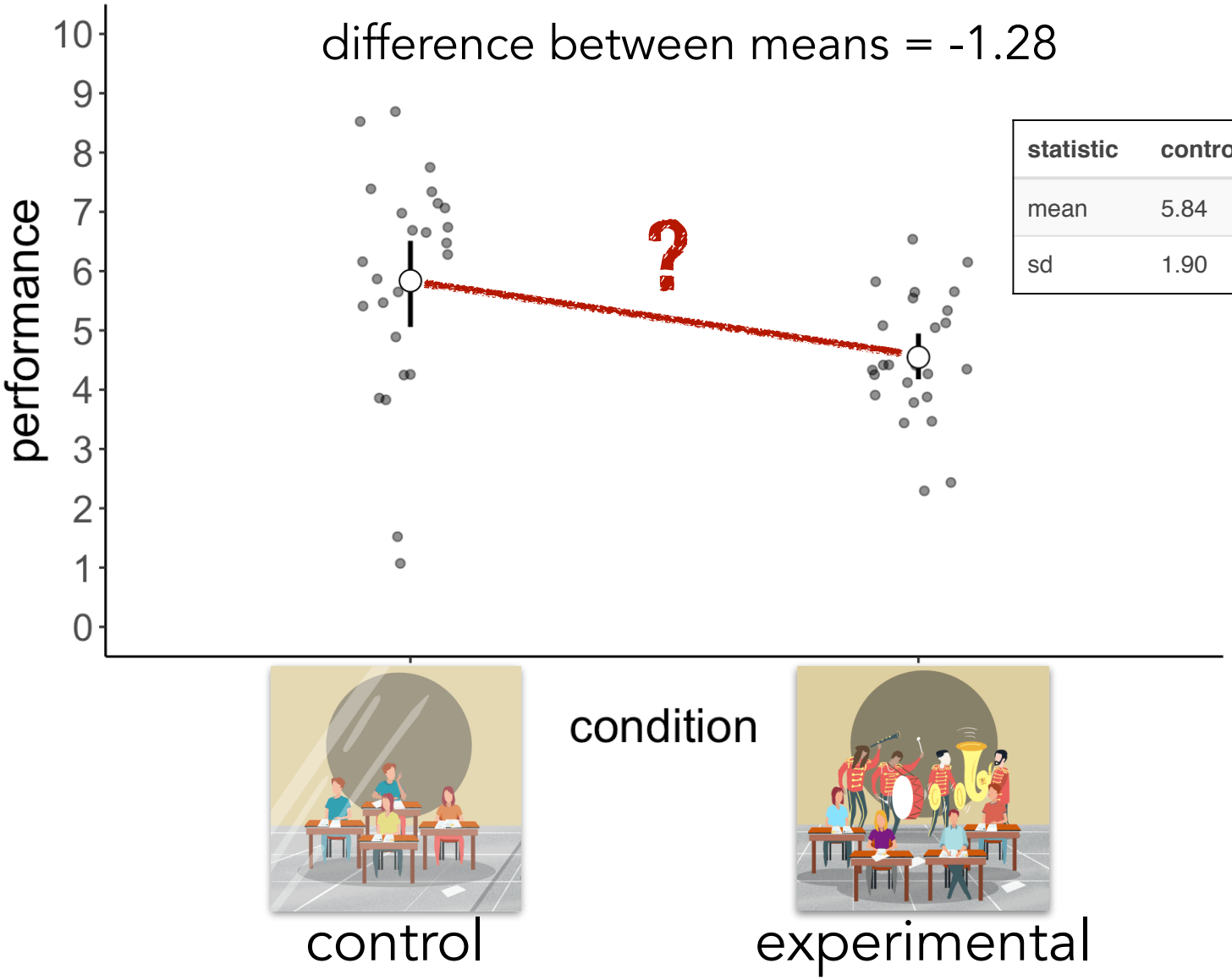
- Sometimes we **can** actually nonparametrically sample from the null.
- In a two-sample t-test, the null is that the two groups really are samples from the same distribution.
- We can actually sample from the null by just shuffling the labels.
- (This also works more generally, e.g. if you want to test median differences rather than mean.)

Permutation tests

- Sometimes we **can** actually nonparametrically sample from the null.
- In a two-sample t-test, the null is that the two groups really are samples from the same distribution.
- We can actually sample from the null by just shuffling the labels.
- (This also works more generally, e.g. if you want to test median differences rather than mean.)
- (I will now steal some slides from Tobi.)

Permutation test

Is the difference in performance statistically significant?



statistic	control	experimental
mean	5.84	4.55
sd	1.90	1.06

Permutation test

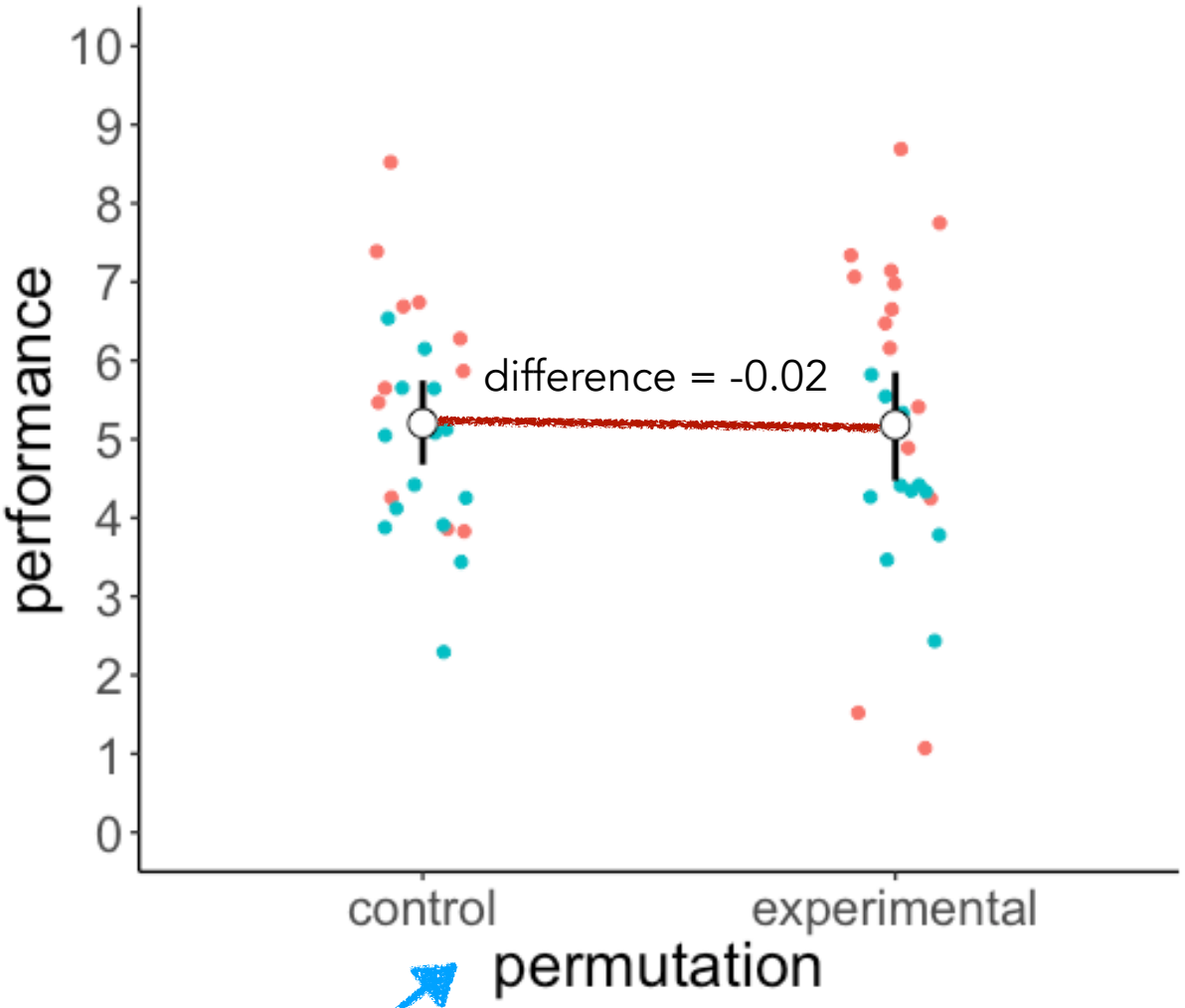
observed data

participant	condition	performance
1	control	4.25
2	control	5.87
3	control	3.83
4	control	8.69
5	control	6.16
26	experimental	4.42
27	experimental	4.27
28	experimental	2.29
29	experimental	3.78
30	experimental	5.13

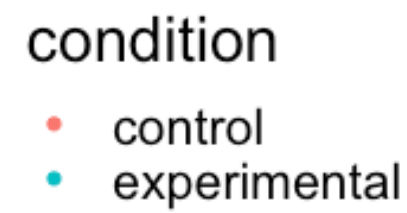
random permutation

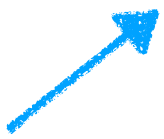
participant	condition	performance
1	control	4.25
2	experimental	5.87
3	control	3.83
4	experimental	8.69
5	control	6.16
26	control	4.42
27	experimental	4.27
28	control	2.29
29	experimental	3.78
30	experimental	5.13

Permutation test



in the actual experiment





in the permutation

Permutation test

observed data

participant	condition	performance
1	control	4.25
2	control	5.87
3	control	3.83
4	control	8.69
5	control	6.16
26	experimental	4.42
27	experimental	4.27
28	experimental	2.29
29	experimental	3.78
30	experimental	5.13

1

participant	condition	performance
1	experimental	4.25
2	control	5.87
3	control	3.83
4	experimental	8.69
5	experimental	6.16
26	control	4.42
27	experimental	4.27
28	control	2.29
29	control	3.78
30	experimental	5.13

2

participant	condition	performance
1	experimental	4.25
2	control	5.87
3	experimental	3.83
4	experimental	8.69
5	experimental	6.16
26	control	4.42
27	control	4.27
28	control	2.29
29	control	3.78
30	experimental	5.13

3

participant	condition	performance
1	control	4.25
2	experimental	5.87
3	control	3.83
4	experimental	8.69
5	control	6.16
26	control	4.42
27	experimental	4.27
28	control	2.29
29	experimental	3.78
30	experimental	5.13

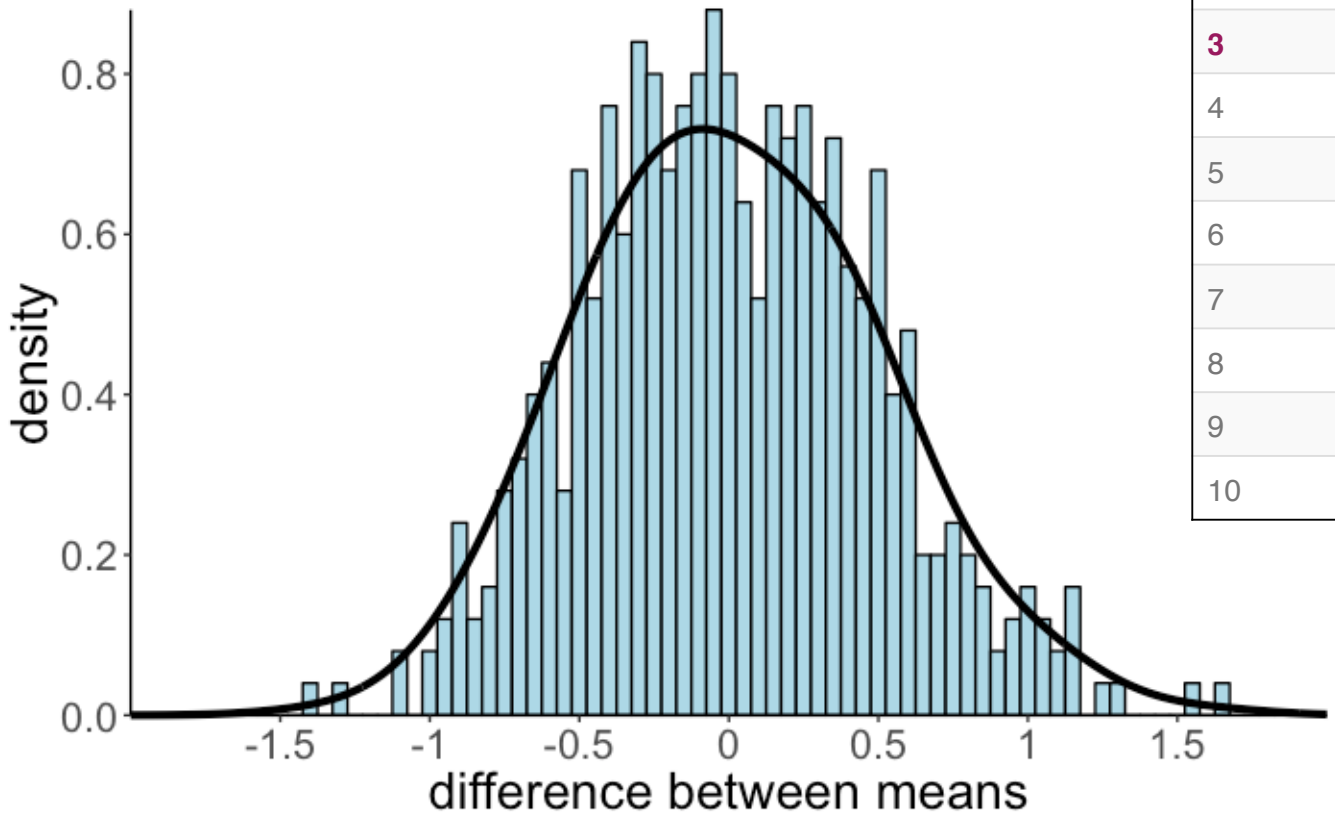
⋮

permutation	mean_difference
1	-0.88
2	-0.26
3	-0.94
4	0.47
5	-0.28
6	1.15
7	0.98
8	0.38
9	-0.08
10	0.31

Permutation test

df.permutations

permutation	mean_difference
1	-0.88
2	-0.26
3	-0.94
4	0.47
5	-0.28
6	1.15
7	0.98
8	0.38
9	-0.08
10	0.31

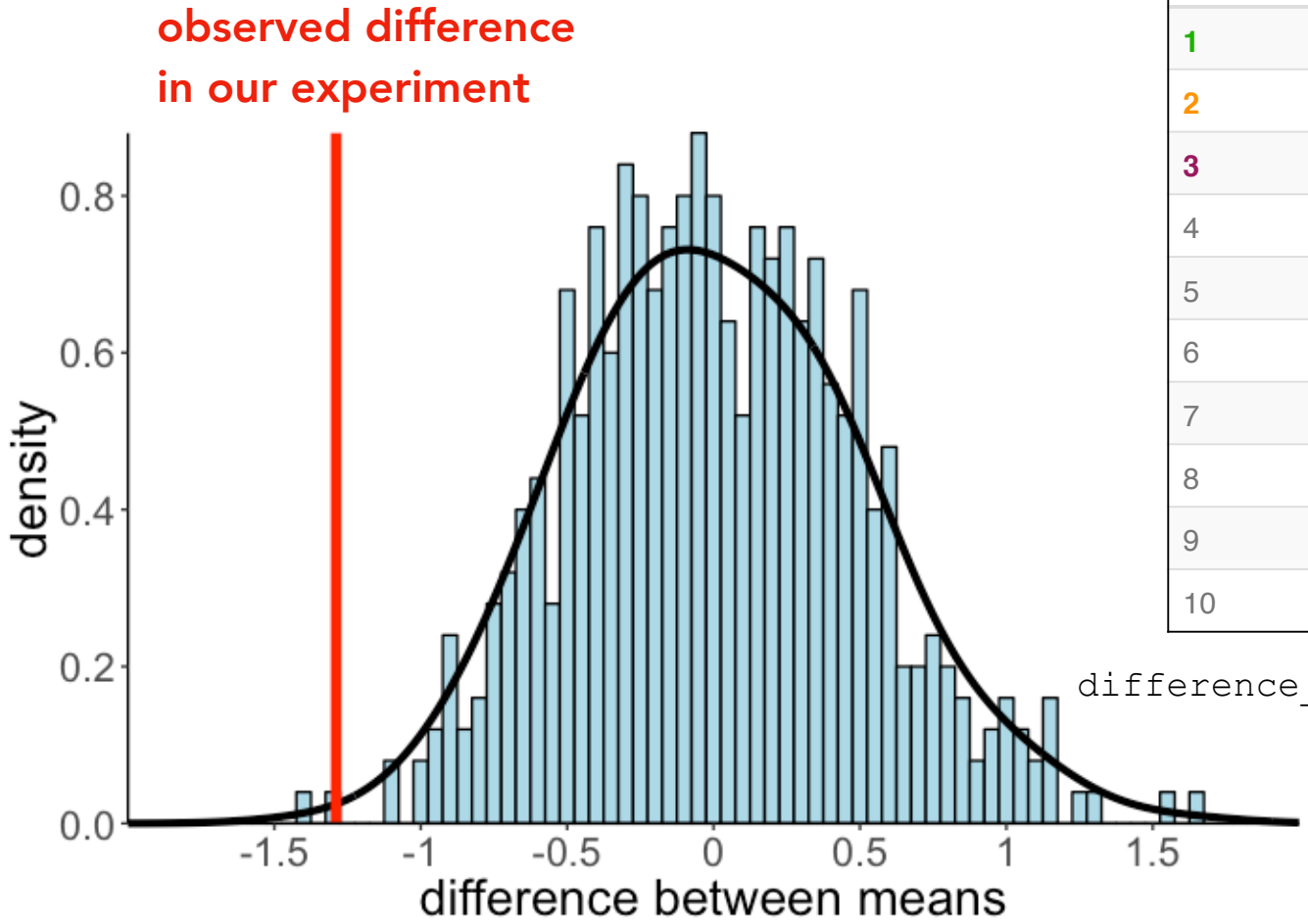


Sampling distribution of differences
(expected differences if the null hypothesis is true)

Permutation test

df.permutations

permutation	mean_difference
1	-0.88
2	-0.26
3	-0.94
4	0.47
5	-0.28
6	1.15
7	0.98
8	0.38
9	-0.08
10	0.31

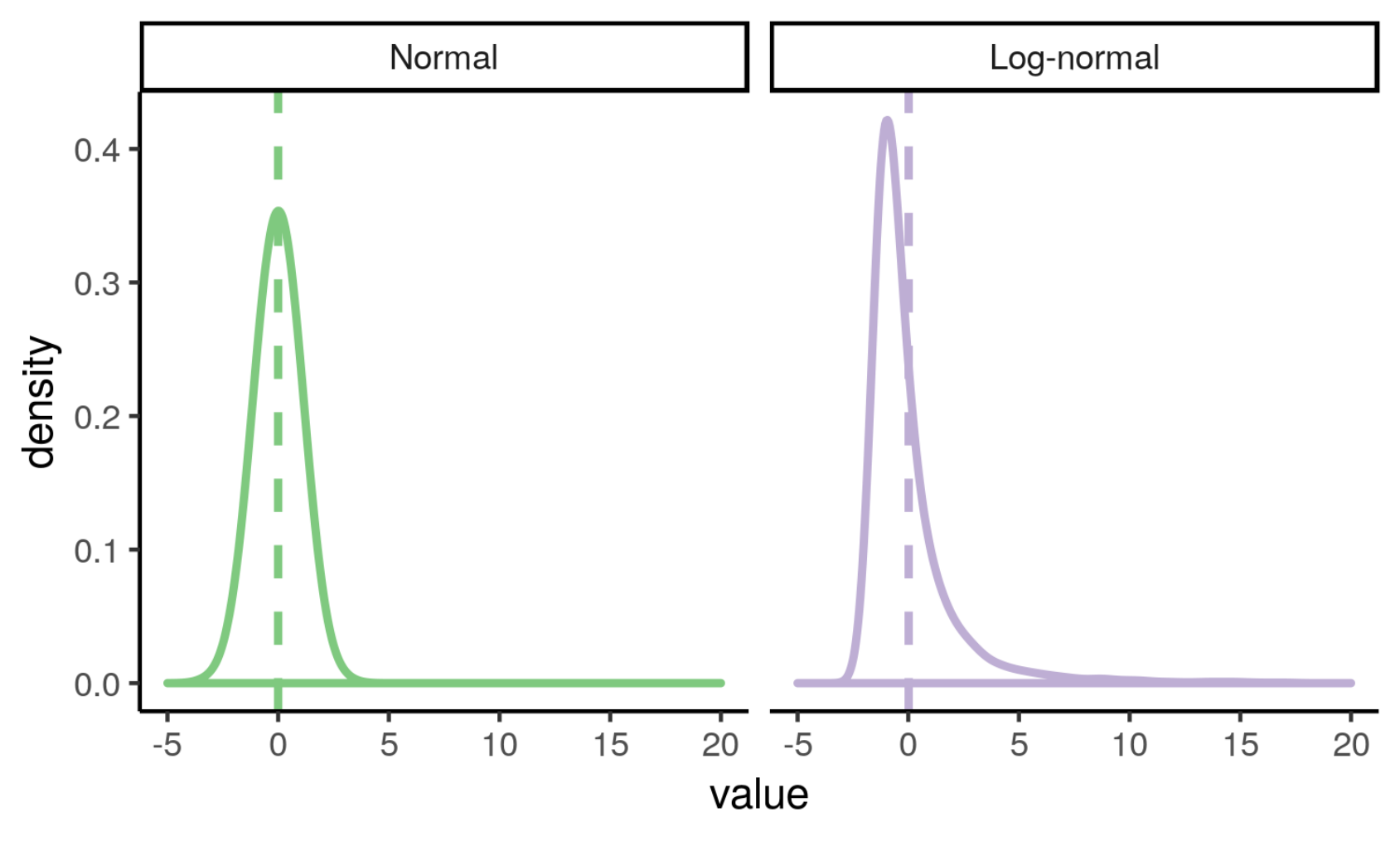


difference_actual = -1.28

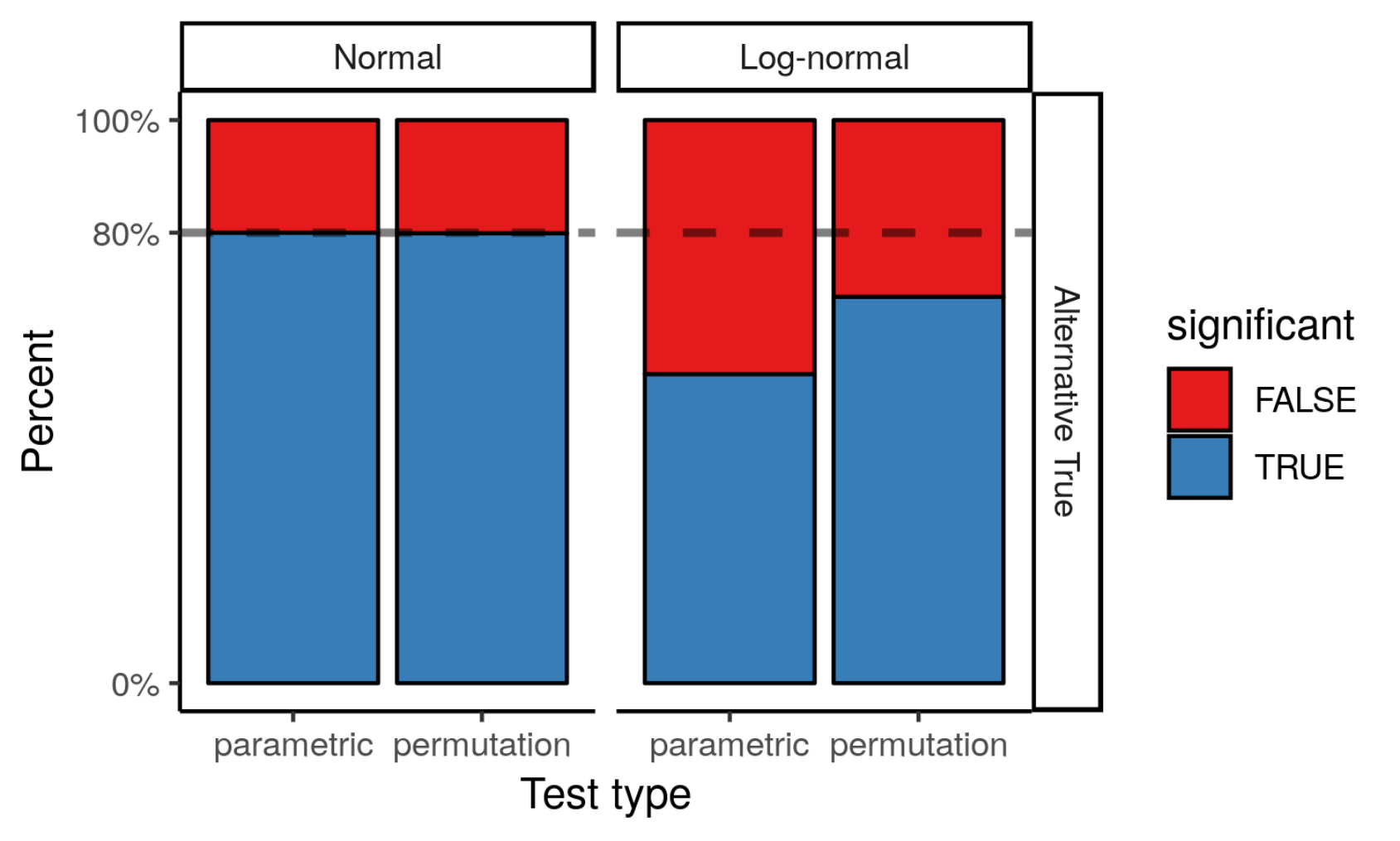
Sampling distribution of differences
(expected differences if the null hypothesis is true)

To run a permutation test on data from two groups, randomly shuffle the group labels many times to generate a null sampling distribution.

Permutation test robustness demo



Permutation test robustness demo



Permutation tests maintain more power if t-test assumptions are violated.

However, both tests benefit from normality, so you should still transform your data if that makes it more normal.

Bootstrap power analyses

Can we use bootstrapping to calculate power?

Can we use bootstrapping to calculate power?

- Yes, if you have pilot or prior experiment data.

Can we use bootstrapping to calculate power?

- Yes, if you have pilot or prior experiment data.
  - Bootstrap resample your pilot data with a resample size of e.g. $n = 50$.

Can we use bootstrapping to calculate power?

- Yes, if you have pilot or prior experiment data.
  - Bootstrap resample your pilot data with a resample size of e.g. $n = 50$.
  - See how many times your desired test (which can be any type, including parametric tests!) comes out significant.

Can we use bootstrapping to calculate power?

- Yes, if you have pilot or prior experiment data.
  - Bootstrap resample your pilot data with a resample size of e.g. $n = 50$.
  - See how many times your desired test (which can be any type, including parametric tests!) comes out significant.
  - Repeat with a different resample size (e.g. $n = 100$).

Can we use bootstrapping to calculate power?

- Yes, if you have pilot or prior experiment data.
  - Bootstrap resample your pilot data with a resample size of e.g. $n = 50$.
  - See how many times your desired test (which can be any type, including parametric tests!) comes out significant.
  - Repeat with a different resample size (e.g. $n = 100$).
  - Iterate until you reach a sample size that achieves the desired power.

Can we use bootstrapping to calculate power?

- Yes, if you have pilot or prior experiment data.
  - Bootstrap resample your pilot data with a resample size of e.g. $n = 50$.
  - See how many times your desired test (which can be any type, including parametric tests!) comes out significant.
  - Repeat with a different resample size (e.g. $n = 100$).
  - Iterate until you reach a sample size that achieves the desired power.
- This will again be more accurate than parametric power calculations when your data don't exactly match the assumed distribution in some way.

To bootstrap a power calculation, run your test on many bootstrap resamples with different resample sizes, and choose the resample size that gives you the desired detection rate.

Wrapping up

A few notes and caveats

A few notes and caveats

A few notes and caveats

- If you can make the data better through experiment design tweaks or data transformation, do that first (it can only improve things).

A few notes and caveats

- If you can make the data better through experiment design tweaks or data transformation, do that first (it can only improve things).
- Samples should not be "too" small ($\gtrsim 20$ points, otherwise you **need** stronger assumptions that probably will be most easily expressed in a parametric framework).

A few notes and caveats

- If you can make the data better through experiment design tweaks or data transformation, do that first (it can only improve things).
- Samples should not be "too" small ($\gtrsim 20$ points, otherwise you **need** stronger assumptions that probably will be most easily expressed in a parametric framework).
- Should compute many bootstrap resamples ($R \approx 1000$ at least, come on, it's the 21st century, modern laptops are more powerful than 20th century supercomputers).

A few notes and caveats

- If you can make the data better through experiment design tweaks or data transformation, do that first (it can only improve things).
- Samples should not be "too" small ($\gtrsim 20$ points, otherwise you **need** stronger assumptions that probably will be most easily expressed in a parametric framework).
- Should compute many bootstrap resamples ($R \approx 1000$ at least, come on, it's the 21st century, modern laptops are more powerful than 20th century supercomputers).
- Biased estimators can still cause problems.

A few notes and caveats

- If you can make the data better through experiment design tweaks or data transformation, do that first (it can only improve things).
- Samples should not be "too" small ($\gtrsim 20$ points, otherwise you **need** stronger assumptions that probably will be most easily expressed in a parametric framework).
- Should compute many bootstrap resamples ($R \approx 1000$ at least, come on, it's the 21st century, modern laptops are more powerful than 20th century supercomputers).
- Biased estimators can still cause problems.
- Nonparametrics generally **do** assume independence, so still need to use mixed(/hierarchical) models, and bootstrapping within these models must respect their dependence structure.

Bootstrapping doesn't fix everything, you still need to be careful!

Summary

- Your sample is a better description of what's going on in the population than whatever you would assume *a priori*.
- So we will approximate samples from the population by resampling with replacement from our sample.
- Using these samples to produce a sampling distribution, we can compute more accurate estimates of important quantities:
  - Confidence intervals.
  - Hypothesis tests.
  - Power calculutations.
- These will generally correspond to the parametric versions when the assumptions of the parametric versions hold, but will be more robust when those assumptions are violated.
- Nonparametrics aren't a panacea, but they help.

Further reading

- The classic:
  - An Introduction to the Bootstrap, Efron & Tibshirani, 1993
- Bootstrap hypothesis testing
  - `https://core.ac.uk/download/pdf/6494364.pdf`
- Permutation tests:
  - `http://faculty.washington.edu/kenrice/sisg/SISG-08-06.pdf`
- More advanced & general:
  - All of Nonparametric Statistics, Larry Wasserstein, 2006 (available electronically through Stanford Libraries!)
- Wikipedia, stackexchange, as always.