浙江大学爱丁堡大学联合学院

ZJU-UoE Institute

Bootstrapping II

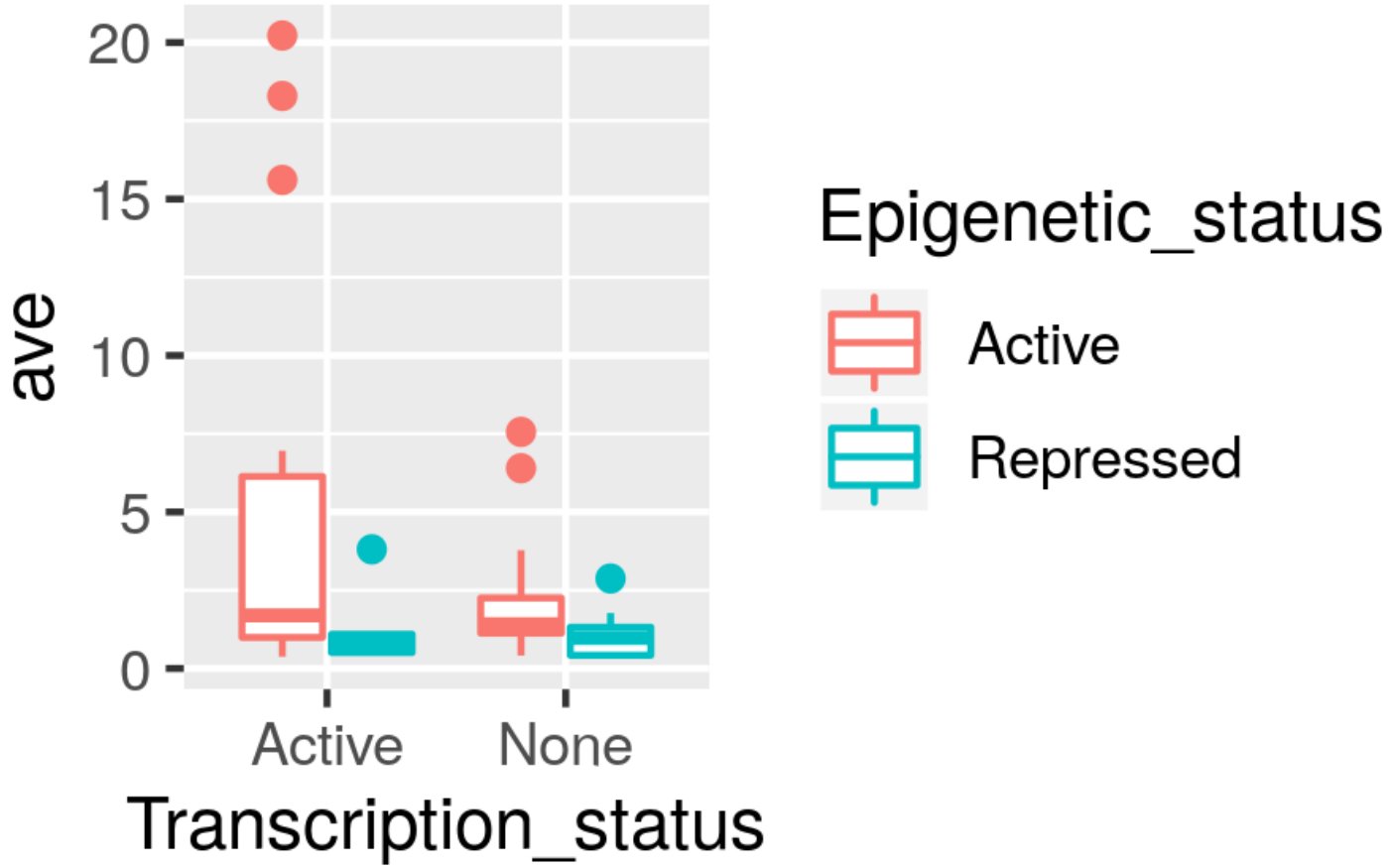
ADS 2, Lecture 21

Melanie Stefan - melanie.stefan@ed.ac.uk

Semester 2, 2019/20

Last week you learned about bootstrapping

E.g. data on enhancer activity



This lecture is about . . .

This lecture is about . . .

More bootstrapping! (Examples and further discussion)

Learning Objectives

After this week, you will be able to . . .

- Discuss the idea behind bootstrapping
- Name different bootstrapping methods
- Recognise situations where bootstrapping is appropriate

Outline

Why do we feel uncomfortable about bootstrapping?

Why do we feel uncomfortable about bootstrapping?

All we have is data . . .

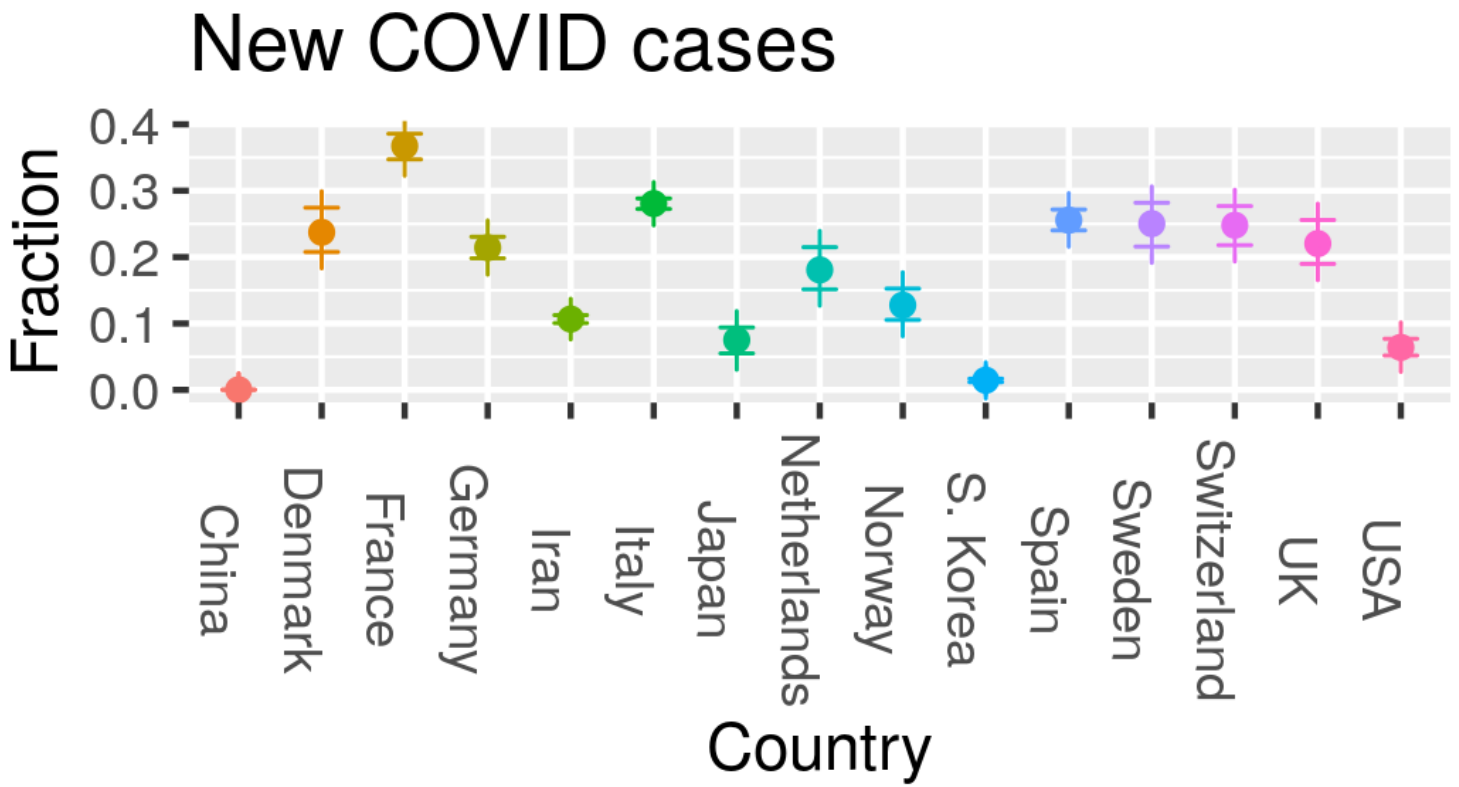
	Country	Total	New
1	China	80796	18
2	Italy	12462	3497
3	Iran	10075	1075
4	S. Korea	7869	114
5	Spain	3059	782
6	Germany	2502	536
7	France	2281	838
8	USA	1390	89

Why do we feel uncomfortable about bootstrapping?

And suddenly . . .

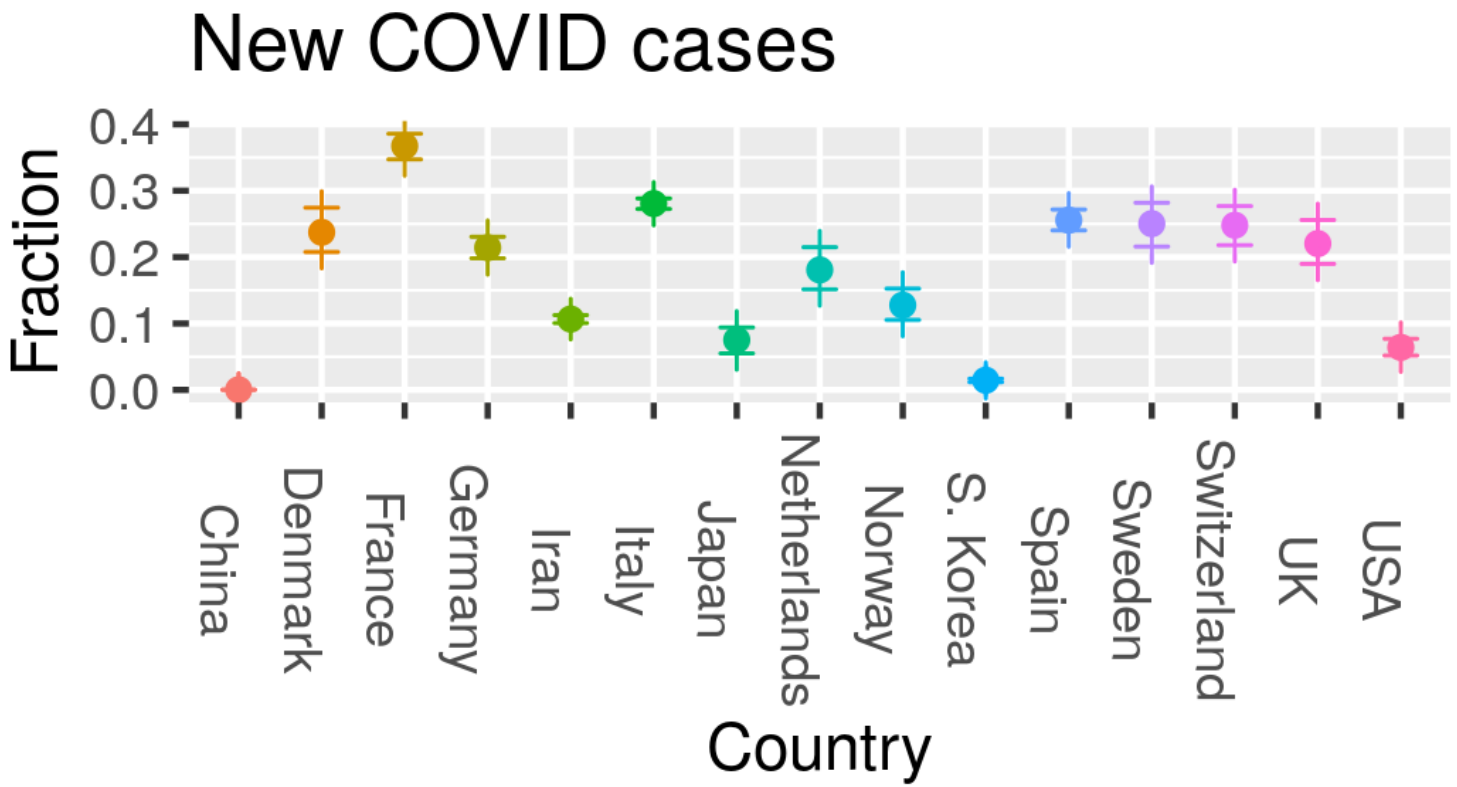
Why do we feel uncomfortable about bootstrapping?

And suddenly . . .



Why do we feel uncomfortable about bootstrapping?

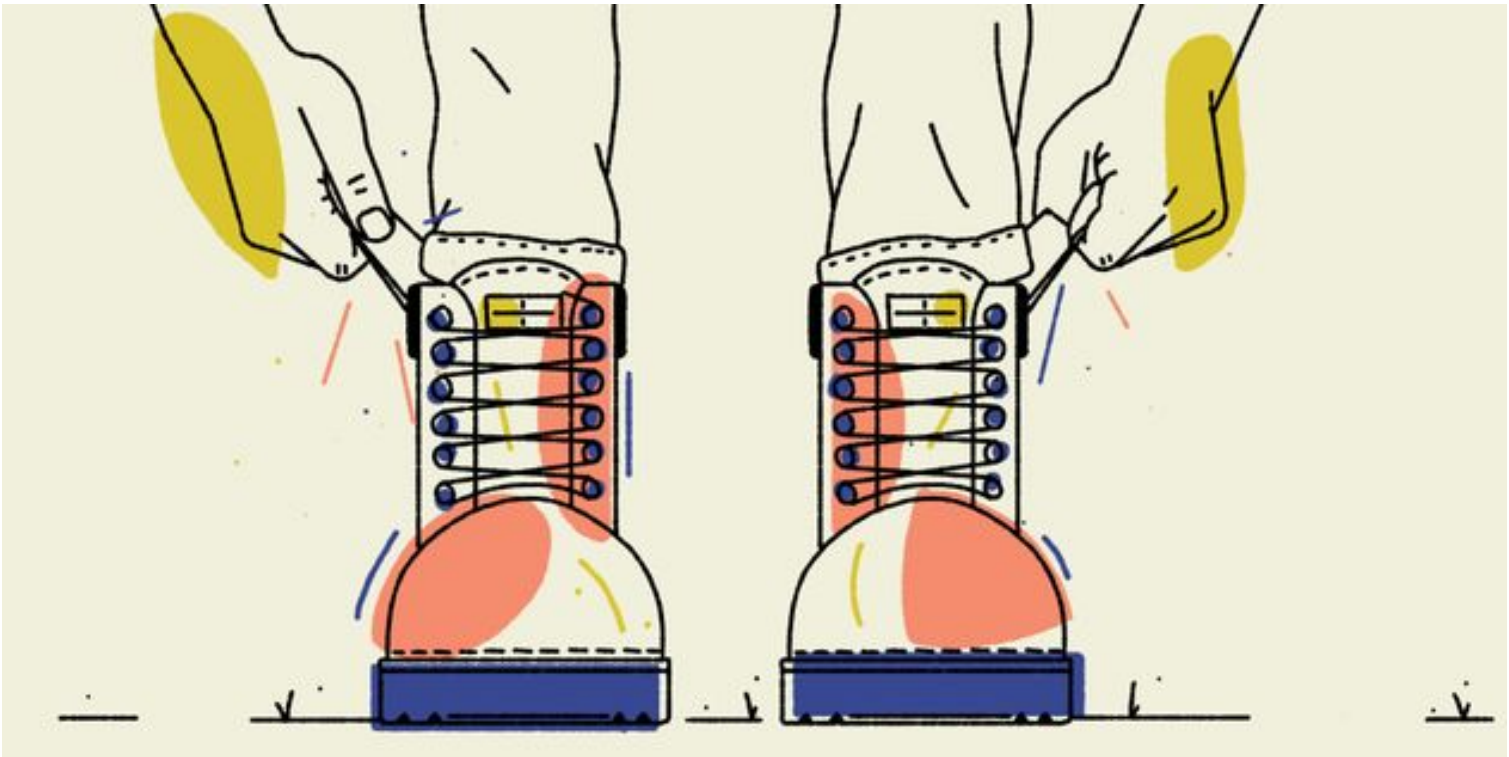
And suddenly . . .



What is this magic?!

And why is it called "Bootstrapping"?

And why is it called "Bootstrapping"?



How can we derive knowledge about the world if all we have is our dataset?

How can we derive knowledge about the world if all we have is our dataset?

But a data is all we ever have!

How can we derive knowledge about the world if all we have is our dataset?

But a data is all we ever have!

Also, data is our best guess as to what the world looks like. [Example]

How can we derive knowledge about the world if all we have is our dataset?

- Bootstrapping makes use of the fact that our data is our best available representation of the world.
- Uses the data to guess at what the world looks like
- Fails if the data is not actually a good representation of the world.

How can we derive knowledge about the world if all we have is our dataset?

- Bootstrapping makes use of the fact that our data is our best available representation of the world.
- Uses the data to guess at what the world looks like
- Fails if the data is not actually a good representation of the world. But then, *any* statistical test would fail!

Outline

When to bootstrap

Good news: You can always use a bootstrap!

When to bootstrap

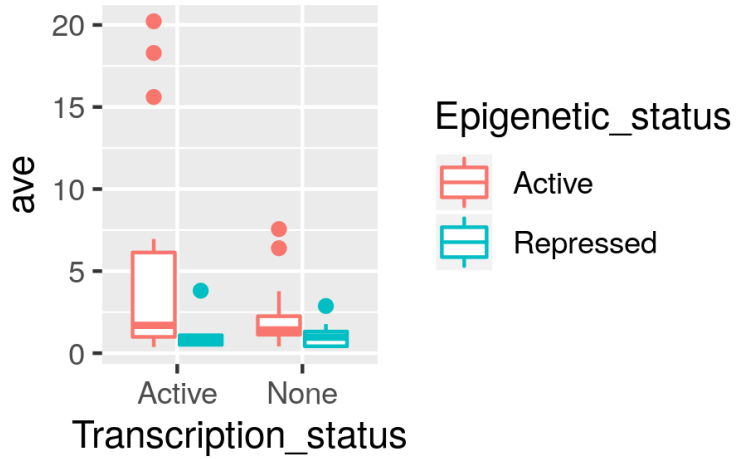
Good news: You can always use a bootstrap!

But *should* you?

When no tests exist for the thing you are testing

Examples

- Comparing medians (not means)
- Comparing any other feature of a data set (skew, range, . . .)
- Comparing bounded data sets (e.g. head movement angle)
- Power calculations for complicated statistical tests
- Adding confidence intervals to observed data



When a test exists, but assumptions are not met

- Comparing means, but the residuals are not normally distributed
- Comparing within-group and between-group differences where you would do an ANOVA if assumptions were met
- ...

When the theoretical outcome is knowable, but complicated

Have you done this before?

When the theoretical outcome is knowable, but complicated

Have you done this before?

- Event that you test for is complex, but the probabilities of individual components are known.
- Outcome distributions of such events are often not intuitive and difficult to calculate correctly.
- Does the observed data fit with the theoretical prediction?

When you have ALL THE DATA

- This does not happen often, but sometimes you can measure the entire population, not just a sample of it.
- Examples?

When you have ALL THE DATA

- This does not happen often, but sometimes you can measure the entire population, not just a sample of it.
- Examples?
  - Conservation work on threatened species
  - Analysis of data in comprehensive databases (e.g. all papers in PubMed)
  - Research on rare diseases
  - ...

Remember

You can always use a bootstrap!

You just have to:

- Know exactly what H_0 and H_A are
- Think carefully about how to bootstrap
- Document your methods

Outline

Different bootstrap methods exist

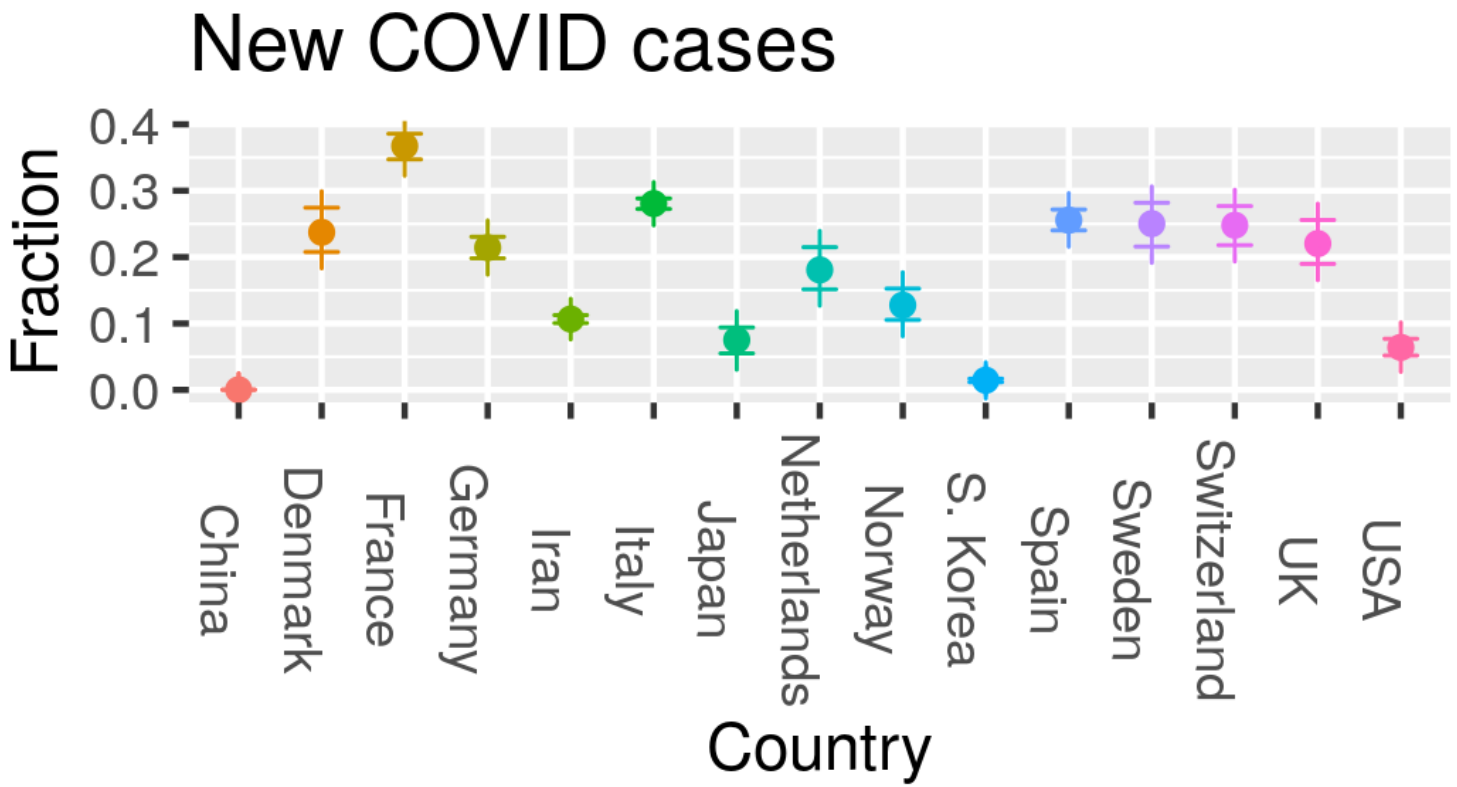
Different bootstrap methods exist for different purposes, and the difference between them is often quite subtle. We will not go into much detail and only look at a few example methods.

Case resampling

- Sample (with replacement) from the data.
- Sample size is the size of the original dataset.
- Compute sample parameter of interest.
- Repeat for many samples.

Have you done this before?

Case resampling

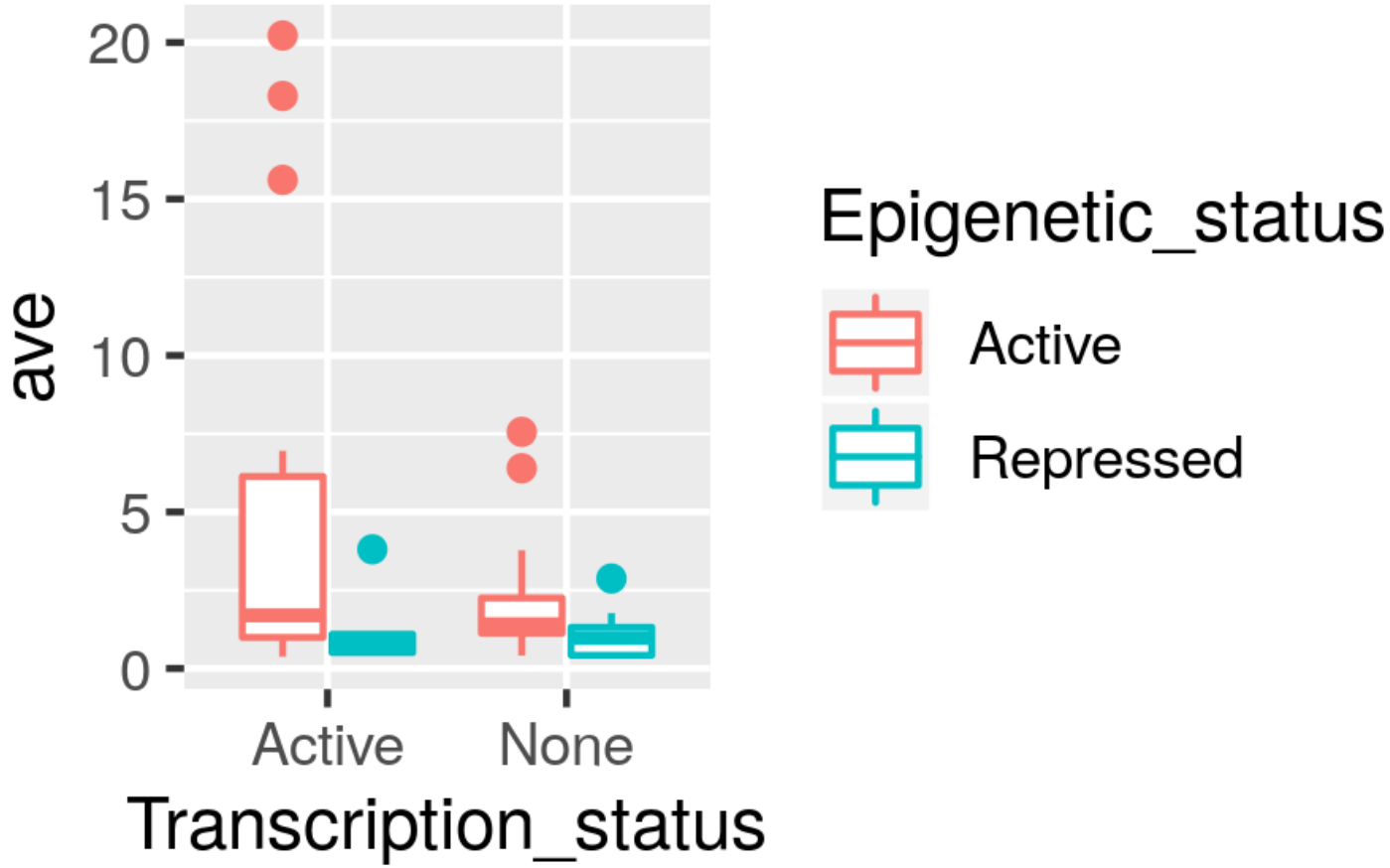


Permutation test

- Sample (without replacement)
- Sample size is size of original dataset
- This essentially amounts to re-distributing labels
- Compare between new samples with different labels
- Repeat many times

Have you done this before?

Permutation test



Parametric bootstrap

- If the real-world distribution is known, we can just simulate the drawing of samples from such a distribution
- E.g. knowing population is normally distributed with given μ and σ allows us to sample from that distribution
- (More realistic scenario: Overall outcome is a combination of events with known probabilities)
- Simulated sample is compared to actually measured sample
- Repeat many times

Review

Now, you should be (more) able to ...

- Discuss the idea behind bootstrapping
- Name different bootstrapping methods
- Recognise situations where bootstrapping is appropriate

What questions do you have?

Image Credits

- Bootstraps. By Isabella Carapella/HuffPost, 2018.
- Covid cases in different countries. Plot my own work (in R), using data provided by Robert Young, 2020.
- Enhancer activity by transcription status and epigenetic status. Plot my own work (in R), using data provided by Robert Young, 2020.
- Man celebrating his birthday. By Silverije - Own work, CC BY-SA 4.0, `https://commons.wikimedia.org/w/index.php?curid=75530328`
- Northern white rhino. By Sheep81 - Own work, Public Domain, `https://commons.wikimedia.org/w/index.php?curid=4745820`
- ZJE students having a good day. Screenshot from: Grace Qin, Hannah Huang, Baizhi Chen, Sherry Chen (writers). Good day of ZJEers, 2017.