Stat 215A - Week 7

Zoe Vernon 10/5/2018
Some slides thanks to Rebecca Barter

Updated timeline for the next few weeks

Week 8: Friday October 12 (Bin will be lecturing during regular lab time)

- ❑ Lab 2 peer reviews due
- ❑ Lab 3 released (due **Tue October 23**)
  - ❑ Shorter lab
  - ❑ There will be no peer review

Week 9: Bin out of town

- ❑ I will do the previous weeks lab during lecture time that Tuesday (10/16)

Week 10:

- ❑ Lab 3 due (Tuesday 10/23)
- ❑ Lab 4 released (Friday 10/26) - group project

Today

Any last comments or questions about lab 2?

Stability

Resampling methods

Stability

Stability: two types of questions

Computational stability

Stability: two types of questions

Computational stability

If I re-run the (possibly stochastic) algorithm again (possibly tweaking parameters) on the **same data**, do I get the same results?

Stability: two types of questions

Computational stability

If I re-run the (possibly stochastic) algorithm again (possibly tweaking parameters) on the **same data**, do I get the same results?

Generalization stability

Stability: two types of questions

Computational stability

If I re-run the (possibly stochastic) algorithm again (possibly tweaking parameters) on the **same data**, do I get the same results?

Generalization stability

If I re-run the algorithm again on a **new sample** of data points from the **same source**, do I get the same results?

Stability: two types of questions

Computational stability

If I re-run the (possibly stochastic) algorithm again (possibly tweaking parameters) on the **same data**, do I get the same results?

Asking about the randomness in the **algorithm**...

Generalization stability

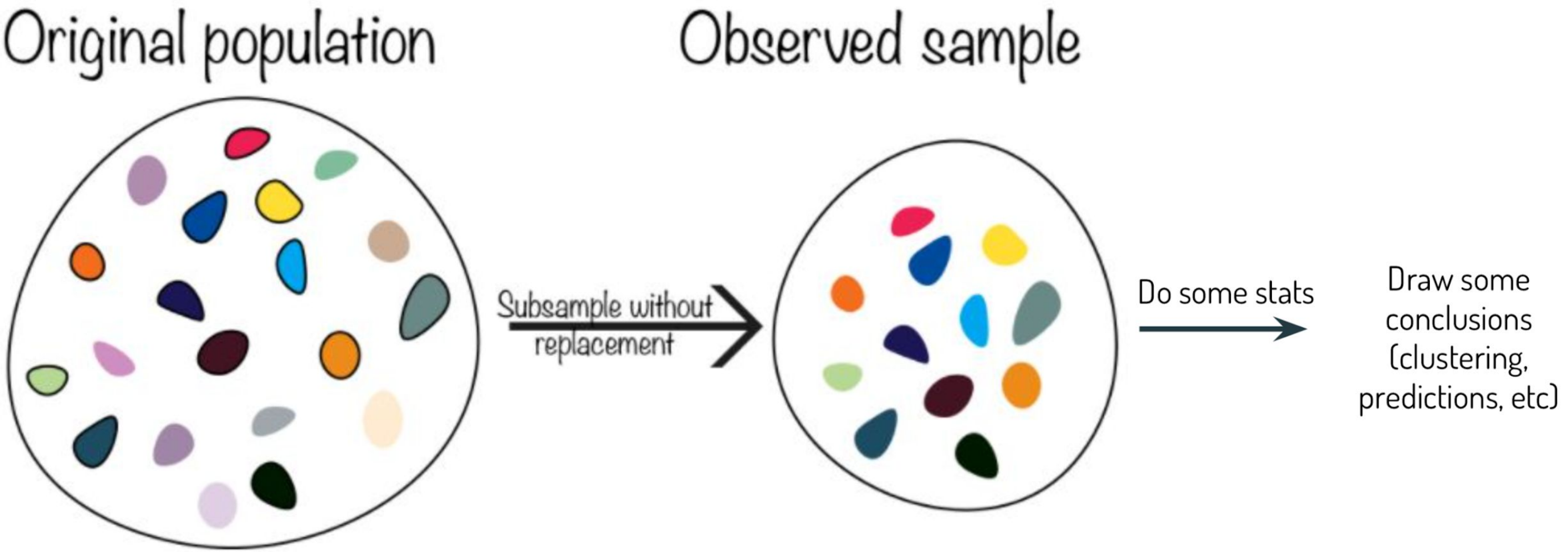
If I re-run the algorithm again on a **new sample** of data points from the **same source**, do I get the same results?

Asking about randomness in the **data**...

Generalization stability: sampling methods

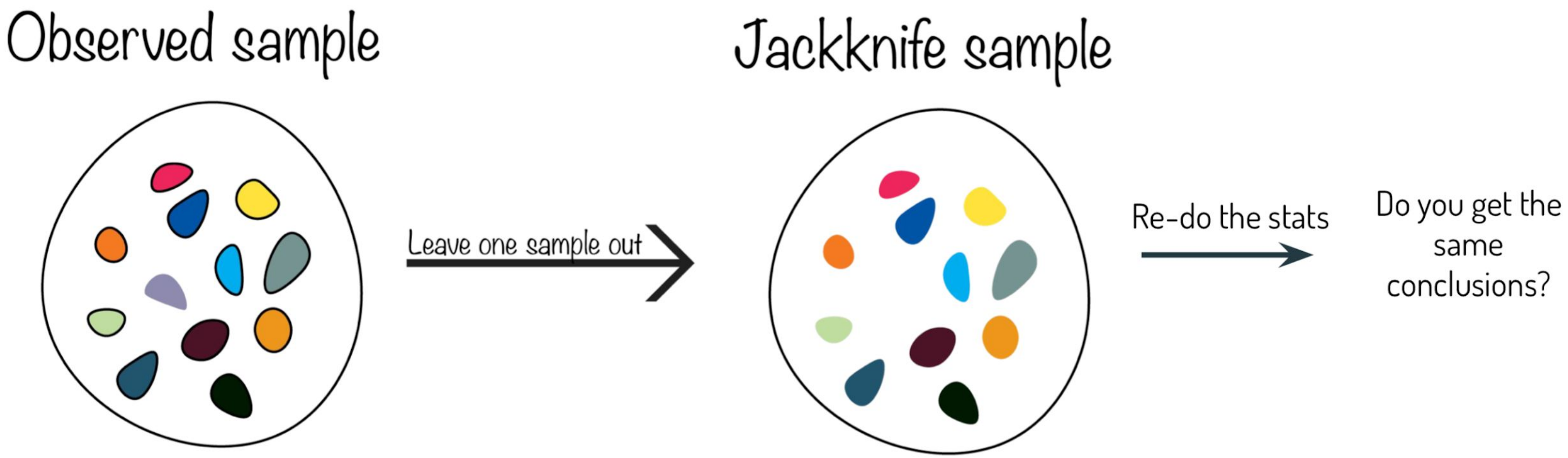
Sampling methods

The purpose of sampling methods is to simulate sampling procedure from the original population



Jackknife resampling

- Obtain a subsample containing all but one of the data points
    - Repeat for all possible excluded data points
- The subsample has one fewer data point that the observed sample
- Non-random sampling



Jackknife resampling: variance estimation

Data: $X_1, \ldots, X_n$

Define: $X_{(i)} = \{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\}$

Estimate variance of $\hat{\theta} = T(X)$ (e.g. sample mean or median)

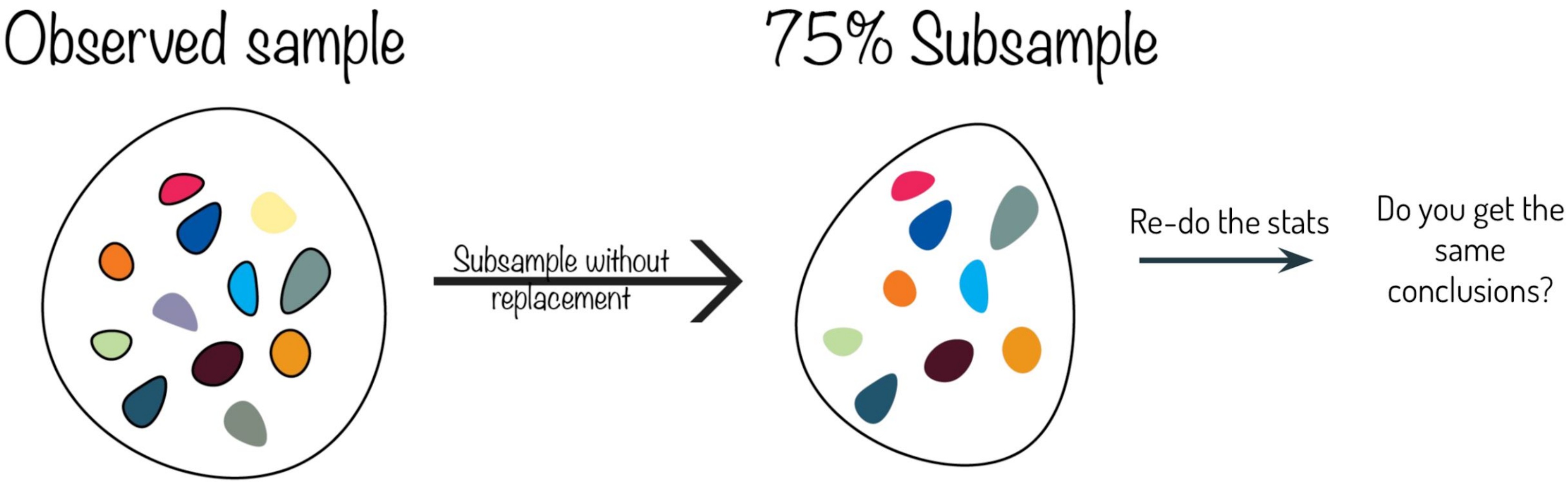
Compute Jackknife replicates: $\hat{\theta}_{(i)} = T\left(X_{(i)}\right)$ and their empirical mean $\hat{\theta}_{(\cdot)} = \frac{1}{n}\hat{\theta}_{(i)}$

Then

$$\mathrm{Var}_{jack}\left(\hat{\theta}\right) = \frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)}\right)^2$$

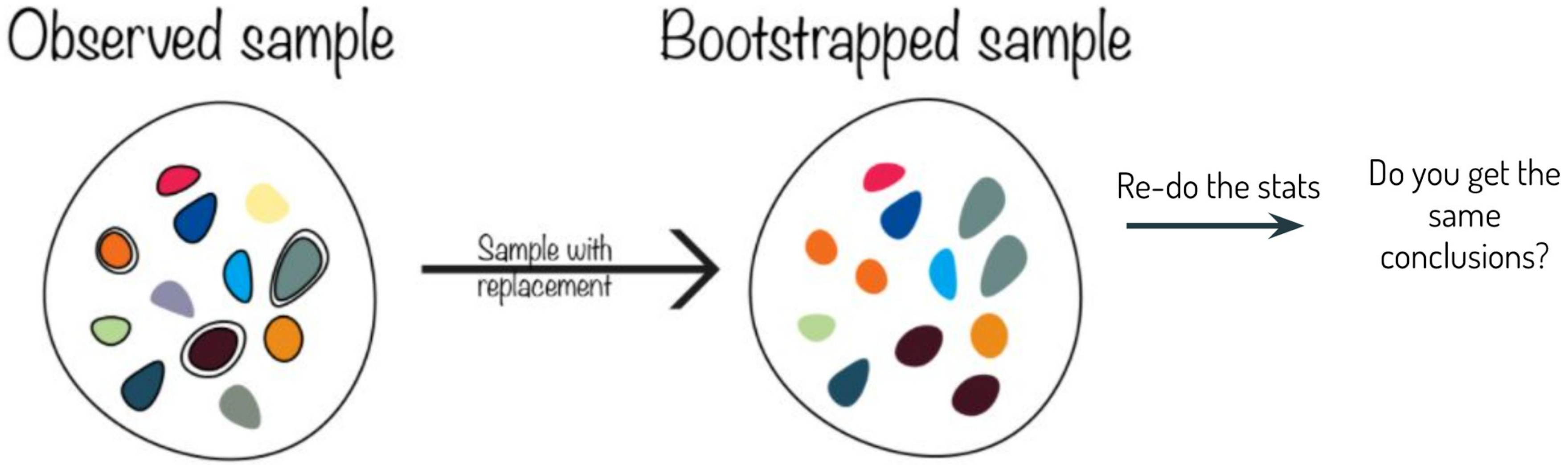
Subsampling

- Sample without replacement
  - Repeat a pre-specified number of times (e.g. 1000)
- The subsample has to be smaller than the observed sample



Bootstrap

- Sample with replacement
  - Repeat a pre-specified number of times (e.g. 1000)
- The bootstrap sample has the same sample size as the observed sample



Bootstrap: variance estimation

Data: $X_1, \ldots, X_n \sim F$ and statistic $\hat{\theta} = T(X)$

Compute empirical distribution:: $F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{X_i \leq t\}$

Sample with replacement: $X_1^*, \ldots, X_n^* \sim F_n$

For $b = 1, \ldots, B$ compute: $\hat{\theta}^{(b)} = T(X_1^*, \ldots, X_n^*)$

Then

$$\widehat{\mathrm{Var}}\left(\hat{\theta}\right) = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}^{(b)} - \bar{\theta}^{(\cdot)}\right)$$

where $\bar{\theta}^{(\cdot)} = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^{(b)}$

Resampling techniques

At the end of the day, no matter what resampling approach you use, you will have many versions of particular estimator.

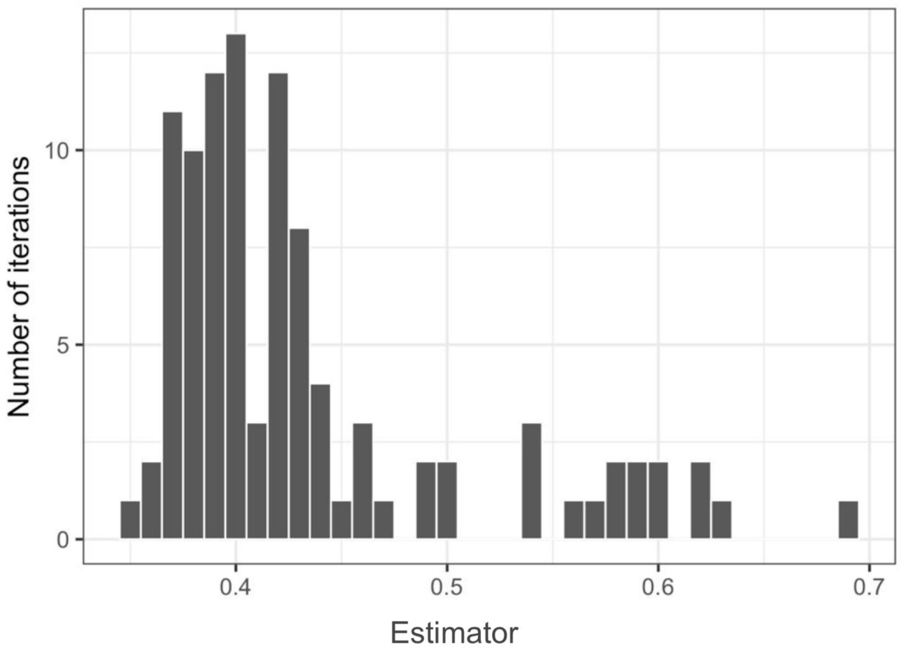
You can use these different versions of the estimate to approximate its distribution as if you had re-drawn samples from the original population.

This allows you to estimate the precision of your estimator non-parametrically.

Resampling techniques: example

The estimator is a random variable

This is an empirical estimate of its distribution drawn from 100 bootstrapped samples



Comparison of techniques

Bootstrap

Different runs give **different** estimates

Estimates the **distribution** of the point estimator

Subsampling

Different runs give **different** estimates

Estimates the **distribution** of the point estimator

Valid under weaker conditions than bootstrap

But have to choose size of subsample

Jackknife

Different runs give **same** estimate

Estimates **properties** of the point estimator (e.g. bias, variance)

Not good for median estimation

Question:

How are these methods related to cross-validation?

Question:

How are these methods related to cross-validation?

Answer:

In **cross-validation**, you build a model using the sampled data and evaluate the model using the left-out data.

In **subsampling/bootstrapping/etc.**, you recalculate statistics on the sampled data and ignore the left-out data entirely

Stability for clustering: wines_stability.Rmd

Wine cluster example from a couple of weeks ago

Let's evaluate the stability of the clusters using these techniques!

1. Test algorithmic stability: re-generate the clusters using the same dataset
   a. Compare the clusters (how?)
2. Test generalization stability: re-generate the clusters using different datasets (bootstrap, subsample, jackknife)
   a. Compare the clusters (how?)