Bootstrap

spatobotp

ttaprspbr

Permutation

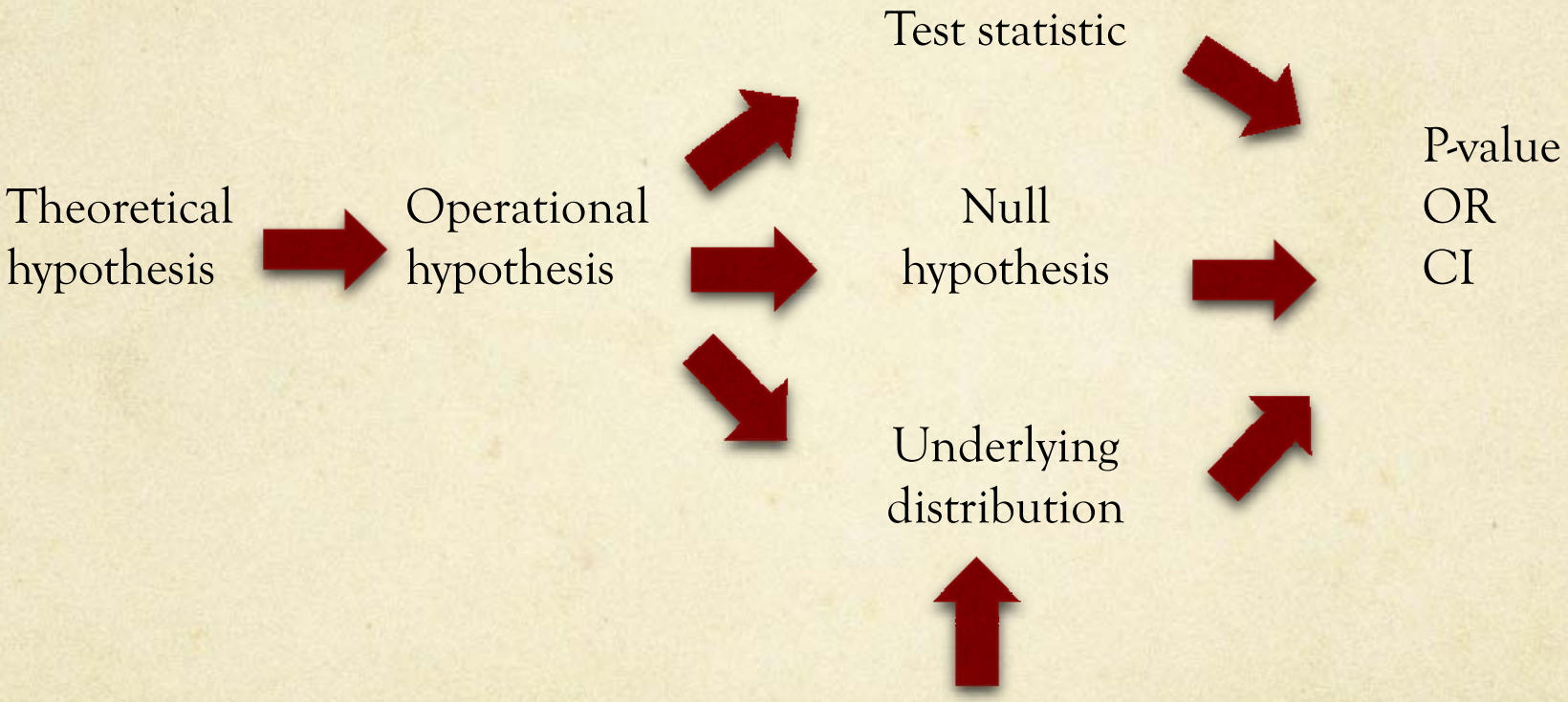
natumeprtoi

toinaumrpet

Bootstrap and permutation

Delphine Courvoisier, PhD

The process of frequentist statistics



If one of those is unusual or unknown, bootstrap or permutation is useful

Outline

- When to use bootstrap or permutation? Advantages and limitations
- Definition of bootstrap and permutation
- How to implement bootstrap and permutation
- Example: is the effect of race on BMI mediated by breastfeeding practice: the Sobel mediation test

When to use bootstrap or permutation?

- Each research question generates one or several test statistics (TS).
  - These TS can be new and thus not implemented in any software.
  - The null hypothesis can be complex (not mean1=mean2)
  - These TS can follow an unknown distribution

Historically

- 1979: seminal paper of Efron. But there were some predecessors
- Popularized in the 80's due to availability of computer.
- Strong mathematical background (even before 1979)

In practice

- Almost always based on simulations
- It is often used to:
  - Estimate standard errors,
  - Estimate bias
  - Construct confidence intervals

Advantages

- Minimal assumption: the sample is a good representation of the unknown population
- Works for almost any test statistic you can think of

Types of bootstrap

- Parametric: we assume that the TS follows a specific distribution. Bootstrap is used to obtain the estimates of the parameters
- Non-parametric: we do not know the distribution of the TS and obtain the empirical distribution

Bootstrap algorithm

- Draw a sample $\mathbf{x}^*$ with replacement from your sample x. Both samples have the same size n.
- Compute TS* for this bootstrap sample
- Repeat steps 1 and 2, B times.

- We obtain $\mathbf{TS}^* = (TS^*_1, TS^*_2, \ldots, TS^*_B)$
- $\mathbf{TS}^*$ is a sample from the unknown distribution of TS.

Bootstrap Standard Errors

- The standard deviation of TS* over the Bootstrap samples is an estimation of the standard error for a single sample

Bootstrap estimate of bias

- Bias(TS)=mean(TS^*_B)-TS

- Because it is an estimate, it will not be exactly zero. Thus, the distance from zero must be estimated relative to the bootstrap standard deviation.

Bootstrap estimate of covariance

- Let TS_1 be a test statistic and TS_2 be another test statistic (e.g., TS_1 is mean and TS_2 is variance):

- Because you estimate each TS in each replication, you can then obtain the covariance(TS_1,TS_2)

Confidence interval: parametric

- Estimate mean and standard error of the supposed distribution by using bootstrap
- Use these estimates to compute a parametric (because you assume a distribution) confidence interval
  - Example: assume normality of TS, bootstrap TS to obtain mean and standard deviation over bootstrap (remember this is similar to standard error in a single sample)

Confidence interval: percentile

- Obtain a few thousands replications
- Compute the TS
- Order the TS from the smallest to the largest
- The 95% percentile confidence interval lower (resp. higher value) is the value of TS such that 2.5% of the replications have a lower (resp. higher) TS than this value.
- This interval is not symmetric

Bias Corrected CI (Bca)

- The TS may be biased.
- The percentile confidence interval then is centered on the central tendency bootstrap value and not on the real sample estimate.

Which CI to choose?

- Percentile CI are easily applicable and intuitive.
- In order to estimate consistently extreme percentiles, we need a very large number of replications
- Bca is equal to percentile CI if TS is not biased.

Bootstrap: How to analyze the data?

- Analysis with R: www.r-project.org
- Package boot

- You can look at an example script: boot.R

Functions with R

- Functions follow this grammar (in italics what you should change):
- *nameOfTheFunction* <- function(*argument1*, *argument 2*,...)
  {*TS*<-*body of the function*
  return(*TS*)}
- The return() function indicates what your own function should give after being evaluated (eg. if the function is mean then the value being returned should be the mean).

Example of a function: my.mean()

- The following function computes the mean

- my.mean <- function(data)
  {meanDC<-sum(data,na.rm=T)/length(!is.na(data))
  return(meanDC)
  }

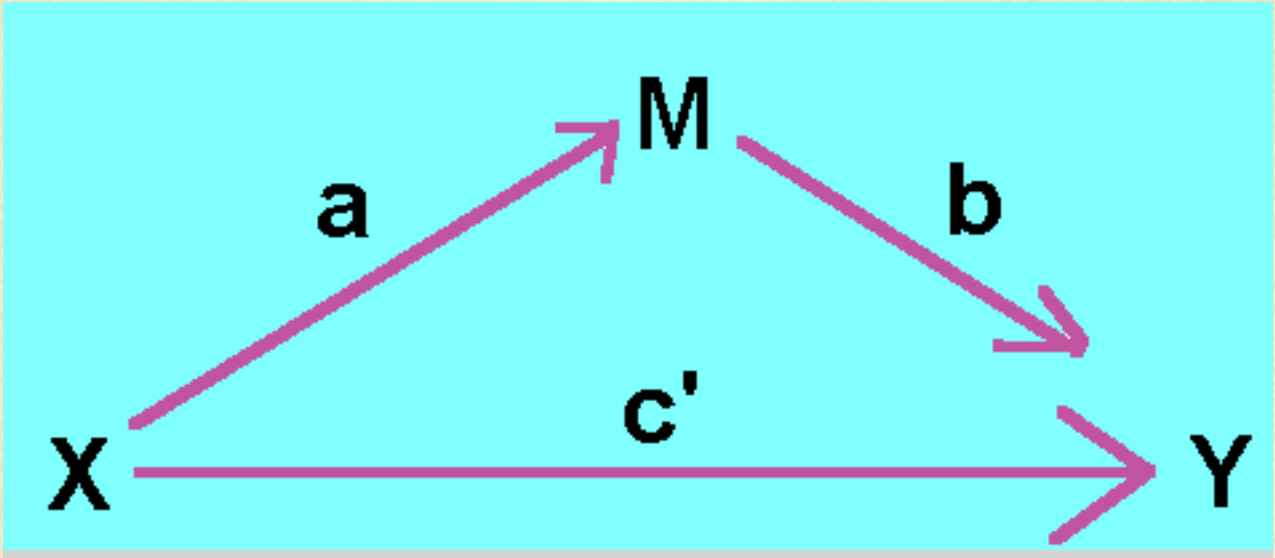
Functions with R

- The first step is to invoke the boot library
  >library(boot)
- The boot function need the following arguments:
- boot(data,statistic,R,arguments used by statistic)
- Data is the data you want to apply your statistic to.
- Statistic is a function that computes a specific TS, it must return a single value
- R is the number of bootstrap replication
- If the function computing the TS needs additional arguments, they can be provided as the last arguments

Creating the appropriate function to pass to the statistic argument

- The function boot uses the function provided in the statistic argument to sample "something" with replacement. In this course, we will only see the case when the "thing" sampled is subject.
- When your data are a single vector (see below), you sample elements
- boot.mean <- function(data,d)
  {data2<-data[d]
  bootmean<-mean(data2,na.rm=T)
  return(bootmean)
  }

Bootstrap: Sobel mediation test



- $c-c'/\sqrt{(b^2 s_a^2 + a^2 s_b^2)}$ is used as a z-test. This test assumes that the above test statistics follows a N(0;1) (but it is known to be untrue; high right skew) and that a and b are independent (true when a regression is used)
- Solution: use bootstrap to determine the distribution of the test statistic (used more and more often)
- Explanation: http://davidakenny.net/cm/mediate.htm
- Calculator: http://www.danielsoper.com/statcalc3/calc.aspx?id=31

Obtaining the c-c' of a mediation test

- Import data dependenceAnxiety.csv in R using the read.csv() function
  Don't forget to save the data under a name using <-
- Use lm() function to
  - regress B on A (and save these results)
  - regress B on A and the mediator (and save these results)
- Obtain c-c' by getting the coefficients of both lm analyses (use nameOfTheAnalysis$coefficients)
- Now, we just have to bootstrap this TS (c-c') ☺

Permutation tests

Basic concepts of permutation

- "Losing the information on the IV",
  also called
  "shuffling the observations across IV groups"

- P-value as the number of permuted samples with a TS smaller (or larger) than the real sample value
  → the p-value is usually one-tailed

In practice: an example

- You want to compare the variance of two groups (Levene's test not very good because it assumes normality).
  - Lose the information on the real group assignment.
  - Randomly assign to each subject a group.
  - Compute the variances for each "new" group.
  - Compute the difference of the variances

In practice: an example

- Do that for all possible permutations of the subjects. OR
- There are often too many possible permutations. Thus, a random sample of all possible permutations are used →Monte Carlo procedure
- The proportion of permutation samples with a TS smaller than the real sample TS represents the p-value
- The distribution of the differences of variances is literally the null hypothesis distribution.

Discussion

- Bootstrap and permutation are intuitive, direct ways to obtain frequentist test statistics, confidence intervals and p-values.
- Bootstrap is becoming increasingly
  - frequent in article: Woo 2008 for the Sobel test
  - well-implemented in statistical softwares (e.g. SPSS)
- Permutation remains more rare, because it requires to think about how to generate data under the null hypothesis.