Uncertainty

Ciencia de Datos para la toma de decisiones en Economía

Ignacio Sarmiento-Barbieri

Agenda

1. Uncertainty: Motivation
2. Statistical/Frequentist Uncertainty
3. The Bootstrap
    - Example: Elasticity of Demand for Gasoline
4. Review
5. Further Readings

1. Uncertainty: Motivation

2. Statistical/Frequentist Uncertainty

3. The Bootstrap
   - Example: Elasticity of Demand for Gasoline

4. Review

5. Further Readings

Motivation

- The real world is messy.
- Recognizing this mess will differentiate a sophisticated and useful analysis from one that is hopelessly naive.
- This is especially true for highly complicated models, where it becomes tempting to confuse signal with noise and hence "overfit."
- The ability to deal with this mess and noise is the most important skill you need.

Motivation

- Here we will introduce the concept of uncertainty, the framework we use to characterize what you know in terms of probabilities.
- In much of econometrics and machine learning, we seek to design models that perform well in the presence of uncertainty and do so via regularization and other model stabilization tools.
- In other cases, it is necessary to do full accounting of uncertainty and assign probability distributions to important parameters.
- Real-world applications require a mix of both stability and uncertainty quantification, and to understand any of these techniques we need a clear understanding of the basics of uncertainty.

1 Uncertainty: Motivation

2 Statistical/Frequentist Uncertainty

3 The Bootstrap

- Example: Elasticity of Demand for Gasoline

4 Review

5 Further Readings

Statistical/Frequentist Uncertainty

- It is characterized by a thought experiment:

If I were able to see a new sample of data,
generated by the same processes and scenarios as my current data,
how would my estimates change?

Statistical/Frequentist Uncertainty

- In statistical inference, we're "certain" of our answer if, when we ask the same question over and over again, we get the same answer each time.
- If our answer fluctuates each time, we're uncertain—and the amount by which those answers fluctuate provides us a quantitative measure of our statistical uncertainty.
- Let's illustrate this with an example

Statistical/Frequentist Uncertainty

Suppose it were actually true that Boca Junios is ”La mitad más uno”, i.e., 51% of all Argentineans would prefer Boca and 49% would prefer River Plate.

Statistical/Frequentist Uncertainty

```r
library(tidyverse)
library(mosaic)
```

```r
set.seed(10101)
sample1<-rbinom(n=100,size=1,p=.51)
sum(sample1)
```

Statistical/Frequentist Uncertainty

```r
library(tidyverse)
library(mosaic)
```

```r
set.seed(10101)
sample1<-rbinom(n=100,size=1,p=.51)
sum(sample1)
```

```
## [1] 46
```

Statistical/Frequentist Uncertainty

```
sample2<-rbinom(n=100,size=1,p=.51)
sum(sample2)
```

```
## [1] 58
```

```
sample3<-rbinom(n=100,size=1,p=.51)
sum(sample3)
```

```
## [1] 59
```

```
sample4<-rbinom(n=100,size=1,p=.51)
sum(sample4)
```

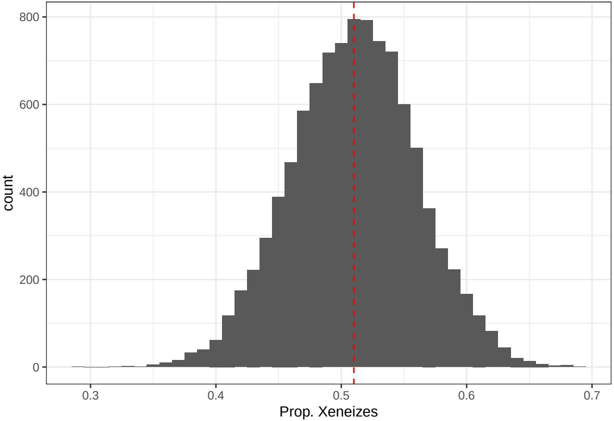
```
## [1] 57
```

```
sample5<-rbinom(n=100,size=1,p=.51)
sum(sample5)
```

```
## [1] 49
```

Statistical/Frequentist Uncertainty

```
samples<-do(10000)*sum(rbinom(n=100,size=1,p=.51))
samples<- samples %>% mutate(prop=sum/100)
```



Statistical/Frequentist Uncertainty

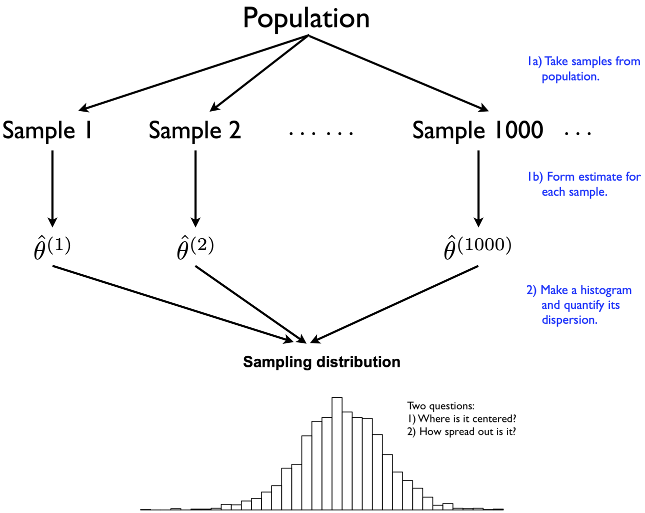
- The central limit theorem (CLT) states that the average of independent random variables becomes normally distributed (i.e., Gaussian or as a “bell curve”) if your sample size is “large enough.”
- What is this distribution? It is our best guess at the sampling distribution.
- It captures the uncertainty described by the thought experiment: “If I was able to get a new sample of observations, from the same data-generating process, what is the probability distribution on the new sample average?”

Statistical/Frequentist Uncertainty

- There are at least two important questions to ask about a sampling distribution, one about its center and one about its spread.
    - First, where is the sampling distribution centered?
        - *Definition:* The expected value of a statistical summary is the average of that summary's sampling distribution. That is, the average value of that summary under repeated sampling from the same random process that generated our data.
    - Second, how spread out is the sampling distribution? The answer to this question provides a quantitative measure measure of repeatability, and therefore statistical uncertainty.
        - *Definition:* The standard deviation of a sampling distribution is called the standard error. This reflects the typical statistical fluctuation of our summary statistic. That is, the typical magnitude of error of our summary, compared with the expected or average value of that summary under repeated sampling.

Statistical/Frequentist Uncertainty

Sampling Distribution



Population

1a) Take samples from population.

Sample 1 Sample 2 ...... Sample 1000 ...

1b) Form estimate for each sample.

$\hat{\theta}^{(1)}$ $\hat{\theta}^{(2)}$ $\hat{\theta}^{(1000)}$

2) Make a histogram and quantify its dispersion.

Sampling distribution

Two questions:
1) Where is it centered?
2) How spread out is it?

Statistical/Frequentist Uncertainty

- So the sampling distribution formalizes the concept of “statistical uncertainty”.

> *El jardín de los senderos que se bifurcan es una imagen incompleta, pero no falsa, del universo tal como lo concebía Ts'ui Pên. A diferencia de Newton y de Schopenhauer, su antepasado no creía en un tiempo uniforme, absoluto. Creía en infinitas series de tiempos, en una red creciente y vertiginosa de tiempos divergentes, convergentes y paralelos. Esa trama de tiempos que se aproximan, se bifurcan, se cortan o que secularmente se ignoran, abarca todas la posibilidades. No existimos en la mayoría de esos tiempos; en algunos existe usted y no yo; en otros, yo, no usted; en otros, los dos. En éste, que un favorable azar me depara, usted ha llegado a mi casa; en otro, usted, al atravesar el jardín, me ha encontrado muerto; en otro, yo digo estas mismas palabras, pero soy un error, un fantasma.*

El jardín de senderos que se bifurcan. Jorge Luis Borges

Statistical/Frequentist Uncertainty

- Quantifying our uncertainty would seem to require knowing all the roads not taken—an impossible task.
- In reality, however, we're stuck with one sample. We therefore cannot ever know the actual sampling distribution of an estimator, for the same reason that we cannot peer into all those other paths
- So in light of that you might ask, rather fairly: what the hell have we been doing this whole time?

Statistical/Frequentist Uncertainty

- Quantifying our uncertainty would seem to require knowing all the roads not taken—an impossible task.
- In reality, however, we're stuck with one sample. We therefore cannot ever know the actual sampling distribution of an estimator, for the same reason that we cannot peer into all those other paths
- So in light of that you might ask, rather fairly: what the hell have we been doing this whole time?
- Surprisingly, we actually can come quite close to performing the impossible.

Statistical/Frequentist Uncertainty

- There are two ways of feasibly constructing something like the previous histogram and approximating an estimator's sampling distribution—all without ever taking repeated samples from the true data-generating process.

  1. Mathematical approximations: that is, by recognizing that the forces of randomness obey certain mathematical regularities, and by drawing approximate conclusions about these regularities using probability theory.

Statistical/Frequentist Uncertainty

The Sampling Distribution of the OLS Estimator

- Given the model

$$y = X\beta + u$$

- Because $\hat{\beta}$ is computed from a sample, the estimators themselves are random variables with a probability distribution (the so-called sampling distribution)
- The CLT gives us

$$\sqrt{N}(\hat{\beta} - \beta) \sim_a N(0, S)$$

Statistical/Frequentist Uncertainty

- There are two ways of feasibly constructing something like the previous histogram and approximating an estimator's sampling distribution—all without ever taking repeated samples from the true data-generating process.

  1. Mathematical approximations: that is, by recognizing that the forces of randomness obey certain mathematical regularities, and by drawing approximate conclusions about these regularities using probability theory.
  2. Resampling: that is, by pretending that the sample itself represents the population, which allows one to approximate the effect of sampling variability by resampling from the sample. → The Bootstrap

1. Uncertainty: Motivation

2. Statistical/Frequentist Uncertainty

3. The Bootstrap
   - Example: Elasticity of Demand for Gasoline

4. Review

5. Further Readings

The Bootstrap

Introduction

- The bootstrap is a widely applicable and extremely powerful statistical tool that can be used to quantify the uncertainty associated with a given estimator or statistical learning method.
- In German the expression *an den eigenen Haaren aus dem Sumpf zu ziehen* nicely captures the idea of the bootstrap – *"to pull yourself out of the swamp by your own hair."*
- The sample itself is used to assess the precision of the estimate.
    - As a simple example, the bootstrap can be used to estimate the standard errors of the coefficients from a linear regression fit. In the specific case of linear regression, this is not particularly useful, since we saw in that standard statistical software such as R outputs such standard errors automatically.
    - However, the power of the bootstrap lies in the fact that it can be easily applied to a wide range of statistical learning methods, including some for which a measure of variability is otherwise difficult to obtain and is not automatically output by statistical software.

The Bootstrap

Introduction

- Why would this work?
- Remember that uncertainty arises from the randomness inherent to our data-generating process
- So if we can approximately simulate this randomness, then we can approximately quantify our uncertainty.
- That's the goal of bootstrapping: to approximate the randomness inherent to data-generating process, so that we can simulate the core thought experiment of statistical inference.

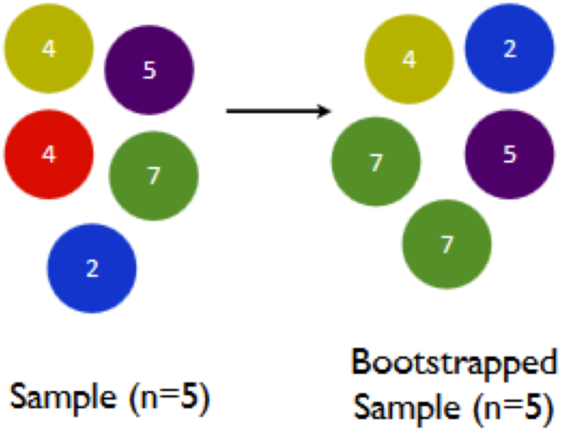
The Bootstrap

Introduction

- There are two key properties of bootstrapping that make this seemingly crazy idea actually work.

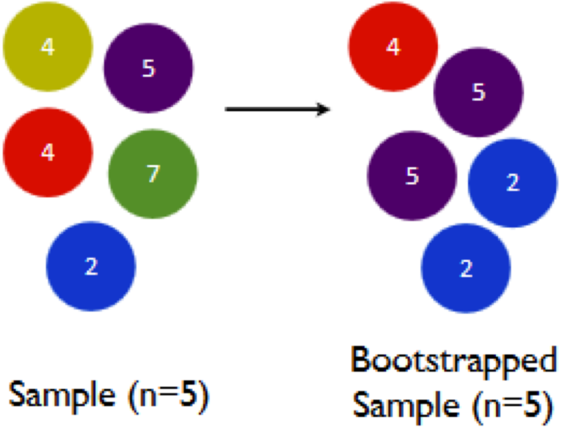
  1. Each bootstrap sample must be of the same size (N) as the original sample
  2. Each bootstrap sample must be taken with replacement from the original sample

Sampling with replacement



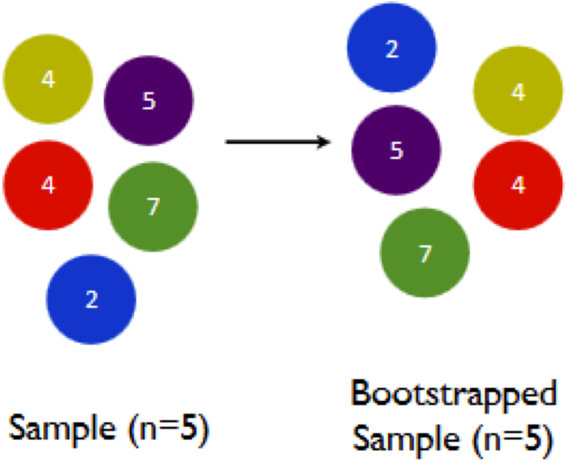
Sampling with replacement

Resampling creates synthetic variability

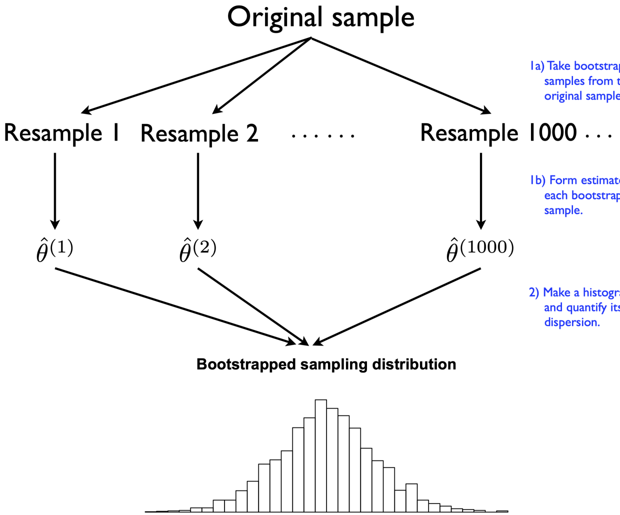


Sampling with replacement

Resampling creates synthetic variability



The Bootstrap



Original sample

Resample 1 Resample 2 . . . . . . Resample 1000 . . .

$\hat{\theta}^{(1)}$ $\hat{\theta}^{(2)}$ $\hat{\theta}^{(1000)}$

Bootstrapped sampling distribution

1a) Take bootstrapped samples from the original sample.

1b) Form estimate for each bootstrapped sample.

2) Make a histogram and quantify its dispersion.

1 Uncertainty: Motivation

2 Statistical/Frequentist Uncertainty

3 The Bootstrap
- Example: Elasticity of Demand for Gasoline

4 Review

5 Further Readings

Example

- Let's illustrate the bootstrap on a toy example in which we wish to determine the elasticity of demand for gasoline.

Example

- Let's illustrate the bootstrap on a toy example in which we wish to determine the elasticity of demand for gasoline.

- Supose we have the following model for the demand for gasoline:

$$\ln Quantity = \alpha + \theta_1 \ln Price + \theta_2 \ln Income + u$$

- The price elasticity, gives the percentage change in quantity demanded when there is a one percent increase in price, holding everything else constant.

$$\eta_{q,p} = \frac{\partial Q}{\partial P}\frac{P}{Q} = \beta_1$$

- It measures the responsiveness of the quantity demanded of a good to a change in its price.

Example

- Load the data

```
gas<-read.csv("gas.csv",header=T)
head(gas)
```

```
  consumption     price   income
1    5.090526 -1.602419 2.342972
2    5.092324 -1.572913 2.317068
3    5.093950 -1.547504 2.340076
4    5.067488 -1.505120 2.323478
5    5.049762 -1.456725 2.335196
6    5.049566 -1.438641 2.364967
```

Example

```r
require("stargazer")
mod1<- lm(consumption~price+income,gas)
stargazer(mod1)
```

Table 1

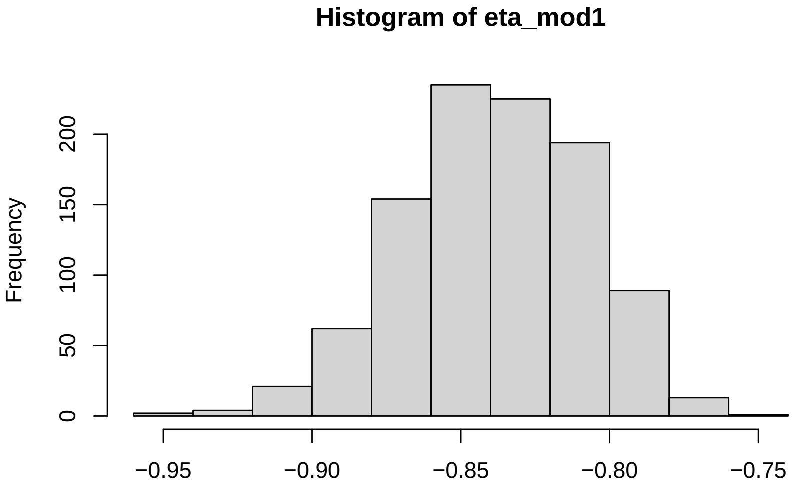
	Dependent variable:
	consumption
θ_1	-0.838^{***} (0.025)
θ_2	2.117^{***} (0.048)
α	-1.056^{***} (0.158)
Observations	246
R^2	0.917
Note:	*p<0.1; **p<0.05; ***p<0.01

Example

```
require("tidyverse")
 set.seed(112)
    n<-length(gas$consumption)

    R<-1000 # Number of Repetions

    eta_mod1<-rep(0,R)

    for (i in 1:R){

      db_sample<- sample_frac(gas,size=1,replace=TRUE)
      f<-lm(consumption~price+income,db_sample)
      coefs<-f$coefficients
      eta_mod1[i]<-coefs[2]
    }
```

Example

```
plot(hist(eta_mod1))
```



Example

```
mean(eta_mod1)
```

```
## [1] -0.838806
```

```
sqrt(var(eta_mod1))
```

```
## [1] 0.03065554
```

```
quantile(eta_mod1,c(0.025,0.975))
```

```
##       2.5%      97.5%
## -0.9005080 -0.7847117
```

Example

- Using the boot package

```
require("boot")
```

- It has the boot function

```
boot(data, statistic, R)
```

- The statistic function

```
eta.fn<-function(data,index){
  coef(lm(consumption˜price+income, data = data, subset = index))
}
```

Example

```
boot(gas, eta.fn, R = 1000)
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = gas, statistic = eta.fn, R = 1000)
##
##
## Bootstrap Statistics :
##       original        bias    std. error
## t1* -1.0561479  0.0036985327  0.16093112
## t2* -0.8383634  0.0007732122  0.03025319
## t3*  2.1168926 -0.0010322045  0.04821903
```

Example

- What happens if the true model is something like this:

$$\ln Quantity = \alpha + \beta_1 \ln Price + \beta_2 \ln Price^2 + \beta_3 \ln Income + \beta_4 \ln Price \times \ln Income + u$$

- The price elasticity then takes the form

$$\eta_{q,p} = \frac{\partial Q}{\partial P}\frac{P}{Q} = \beta_1 + \beta_2 \times 2 \times \ln Price + \beta_4 \ln Income$$

Example

- How do we get this elasticity from the data?
- Begin by constructin the variables

```
  gas<- gas %>% mutate(price2=price^2,
                        price_income=price*income )
```

- then regress:

```
mod2<-lm(consumption~income+price+price2+price_income,gas)
```

Example

- Obtain the coefficients of regression:

```
coefs<-mod2$coef
coefs
```

```
(Intercept)        price       price2       income price_income
 -0.5081995   -1.7825417   -0.2986801    1.9734354    0.2788964
```

Example

- Extract the coefficients to scalars:

```
b0<-coefs[1]
b1<-coefs[2]
b2<-coefs[3]
b3<-coefs[4]
b4<-coefs[5]
```

Example

- Calculate the elasticity (I’m going to do it at the sample mean):

```
price_bar<-mean(gas$price)
income_bar<-mean(gas$income)



elastpt<-b2+2*b3*price_bar+b4*income_bar

elastpt
```

```
    price
-0.671777
```

Example

How do we calculate the standard errors?

```
eta_mod2.fn<-function(data,index,
                        price_bar=mean(gas$price),
                        income_bar=mean(gas$income)){
      f<-lm(consumption~income+price+price2+price_income,data, subset = index)
      coefs<-f$coefficients
      b2<-coefs[3]
      b3<-coefs[4]
      b4<-coefs[5]
      elastpt<-b2+2*b3*price_bar+b4*income_bar
      return(elastpt)
    }

eta_mod2.fn(gas,1:n)
```

```
    price
-0.671777
```

Example

```
results <- boot(data=gas, eta_mod2.fn,R=1000)
results
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = gas, statistic = eta_mod2.fn, R = 1000)
##
##
## Bootstrap Statistics :
##      original        bias    std. error
## t1* -0.671777 0.0004630156  0.01541431
```

1. Uncertainty: Motivation
2. Statistical/Frequentist Uncertainty
3. The Bootstrap
   - Example: Elasticity of Demand for Gasoline
4. **Review**
5. Further Readings

Review and Caveats

- The bootstrap is a widely applicable and extremely powerful statistical tool that can be used to quantify the uncertainty associated with a given estimator or statistical learning method.
- The power of the bootstrap, and resampling in general, lies in the fact that it can be easily applied to a wide range of statistical learning methods.
- In particular, it does not assume that the regression errors are iid so it can accommodate heteroscedasticity for example.
- Of course it does still assume that the observations are independent.
- Resampling dependent observations is an inherently more difficult task which has generated its own rather large literature. (more on this latter)

Further Readings

- Davidson, R., & MacKinnon, J. G. (2004). Econometric theory and methods (Vol. 5). New York: Oxford University Press.

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

- Koenker, R. (2013) Economics 508: Lecture 5. The δ-Method and the Bootstrap. Introduction to Nonlinear Inference. Mimeo

- Scott, J (2022). Data Science in R: A Gentle Introduction. Mimeo.

- Taddy, M. (2019). Business data science: Combining machine learning and economics to optimize, automate, and accelerate business decisions. McGraw Hill Professional.