Computer Intensive Methods using R

Part 1: Introduction

Prof. Dr. Ziv Shkedy

Master of Statistics
Hasselt University

General Information

Overview of the course

- Introduction.
  - The accuracy of the sample mean.
  - Random sampling.
  - The empirical distribution function and the plug-in principle.
  - Standard errors and estimated standard error.

Overview of the course (part 1)

The Bootstrap algorithm

Introduction:
- Sampling from a population.
- The empirical distribution.
- Plug in principle.

Estimation:
- Accuracy of statistics.
- Confidence intervals.

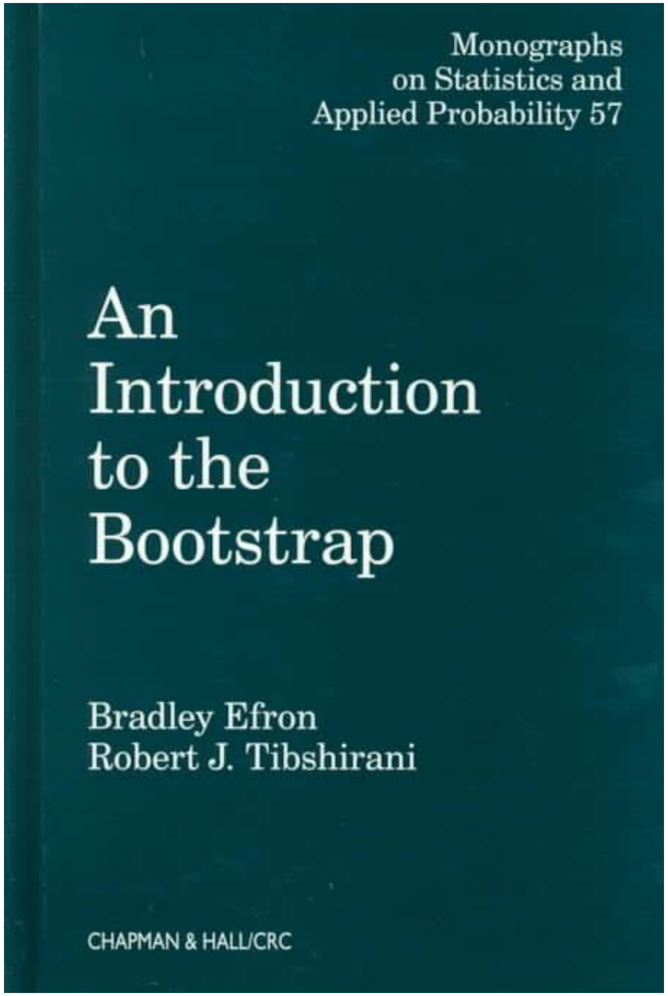
Inference:
- One sample tests.
- Two-samples tests.
- Bootstrap and permutation tests.

Modeling:
- Linear regression models.
- Non parametric regression.
- GLMs.

Reference

- Bradley Efron and Robert J. Tibshirani (1994): An introduction to bootstrap.

- Davison A.C. and Hinkley D.V: Bootstrap Methods and Their Application.



Course materials

- Slides.
- R program.
- R datasets & External datasets.
- YouTube tutorials.
- Videos for the classes (highlights of each class in the course).

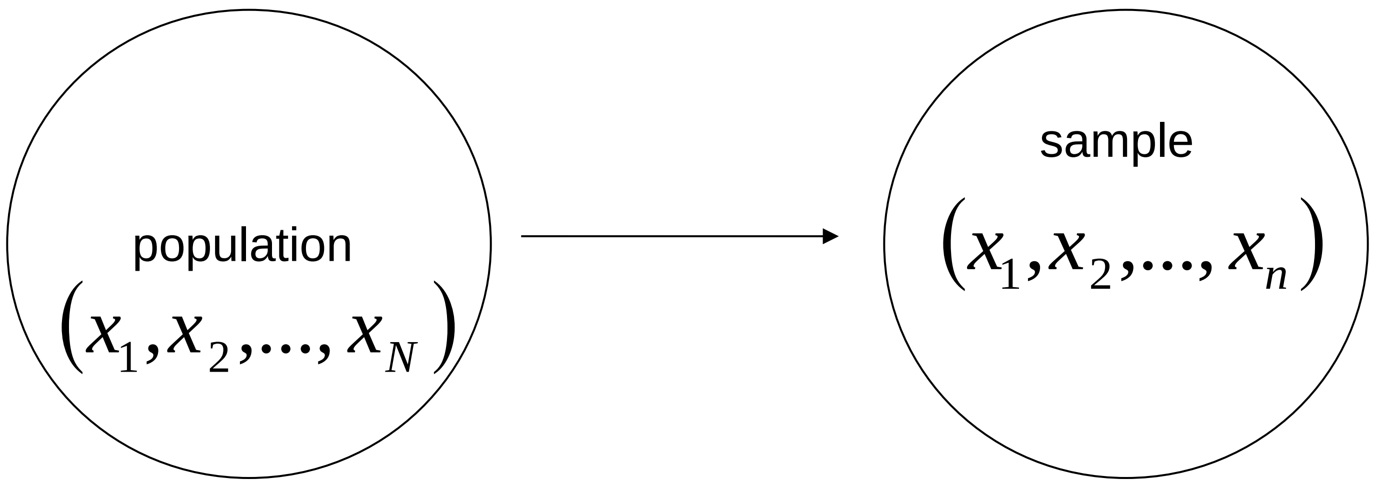
YouTube tutorials

- YouTube tutorials about bootstrap using R:
  1. One-sample bootstrap CI for the mean (host: LawrenceStats): https://www.youtube.com/watch?v=ZkCDYAC2iFg.
  2. Using the non-parametric bootstrap for regression models in R (host:Ian Dworkin):https://www.youtube.com/watch?v=ydtOTctg5So.
  3. Performing the Non-parametric Bootstrap for statistical inference using R (host: Ian Dworkin): https://www.youtube.com/watch?v=TP6r5CTd9yM
  4. Using the sample function in R for resampling of data - absolute basics (host: Ian Dworkin):https://www.youtube.com/watch?v=xE3KGVT6VLE
  5. Permutation tests in R - the basics (host:Ian Dworkin):https://www.youtube.com/watch?v=ZiQdzwB12Pk.
  6. Bootstrap Sample Technique in R software (host: Sarveshwar Inani):https://www.youtube.com/watch?v=tb6wb9ZdPH0
  7. Bootstrap confidence intervals for a single proportion (host: LawrenceStats):https://www.youtube.com/watch?v=ubX4QEPqx5o
  8. Bootstrapped prediction intervals (host:James Scott):https://www.youtube.com/watch?v=c3gD_PwsCGM.

- https://www.youtube.com/watch?v=gcPIyeqymOU

Introduction

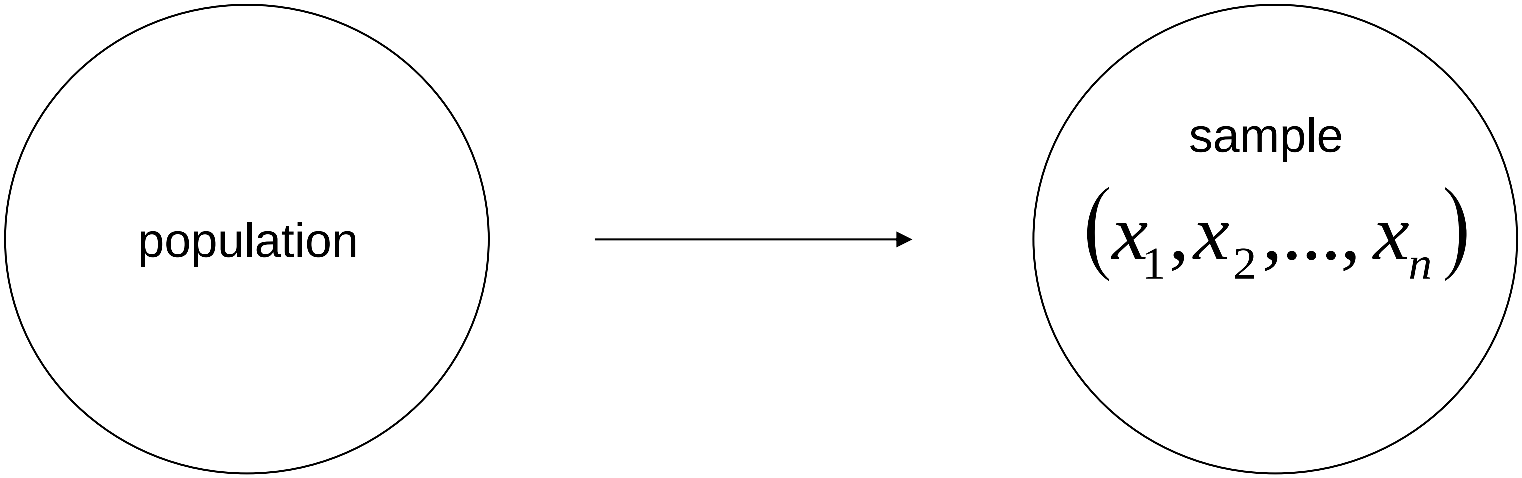
A random Sample



$$F \to (x_1, x_2, \ldots, x_n)$$

Independent and identically distributed sample from F

The probability distribution in the population



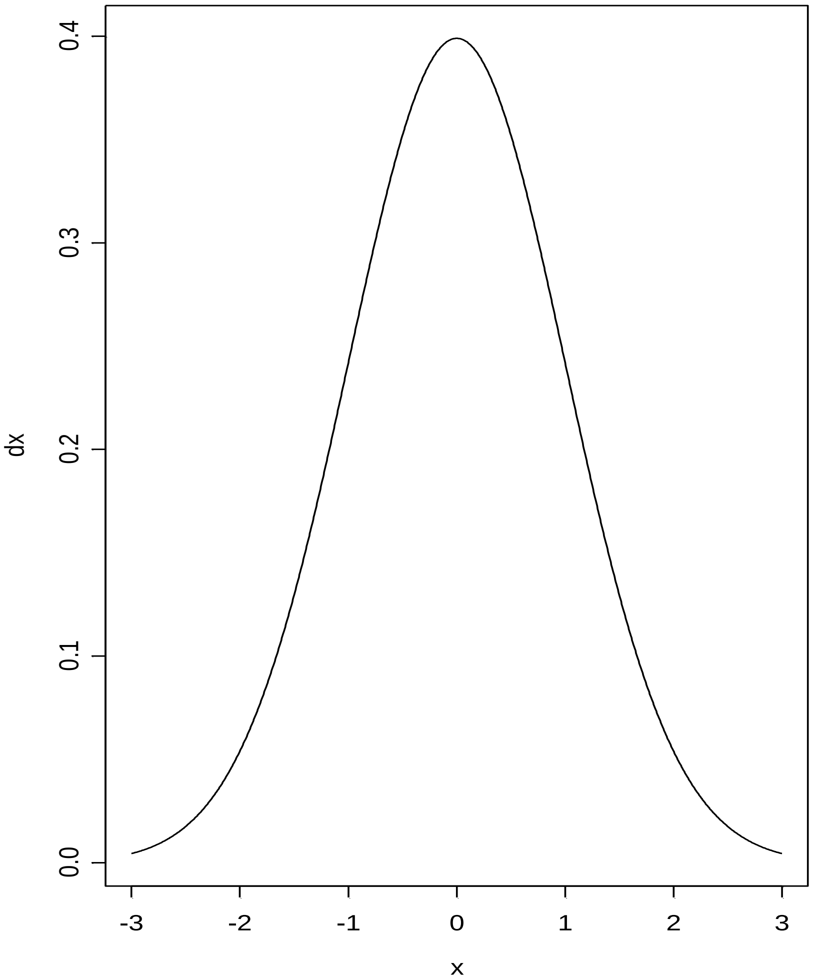
X: a random variable

$$X \sim F(\theta)$$

θ : a parameter

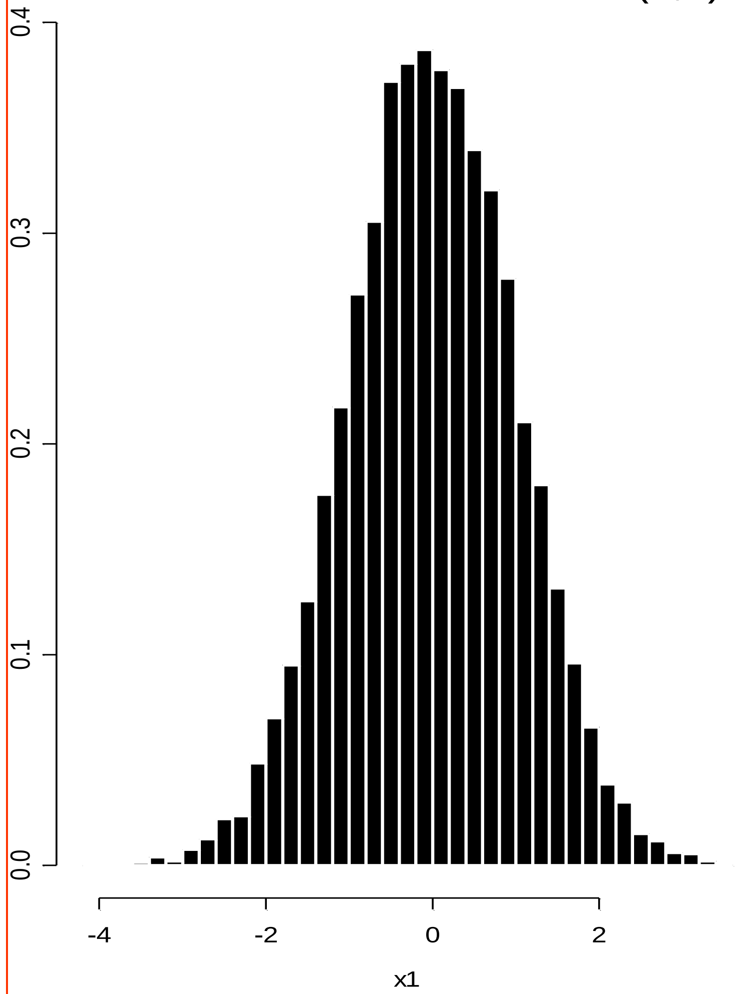
Example: a random sample from N(0,1)

N(0,1)



$$F(\theta) = N(0,1)$$

A sample of 10000 observations from N(0,1)

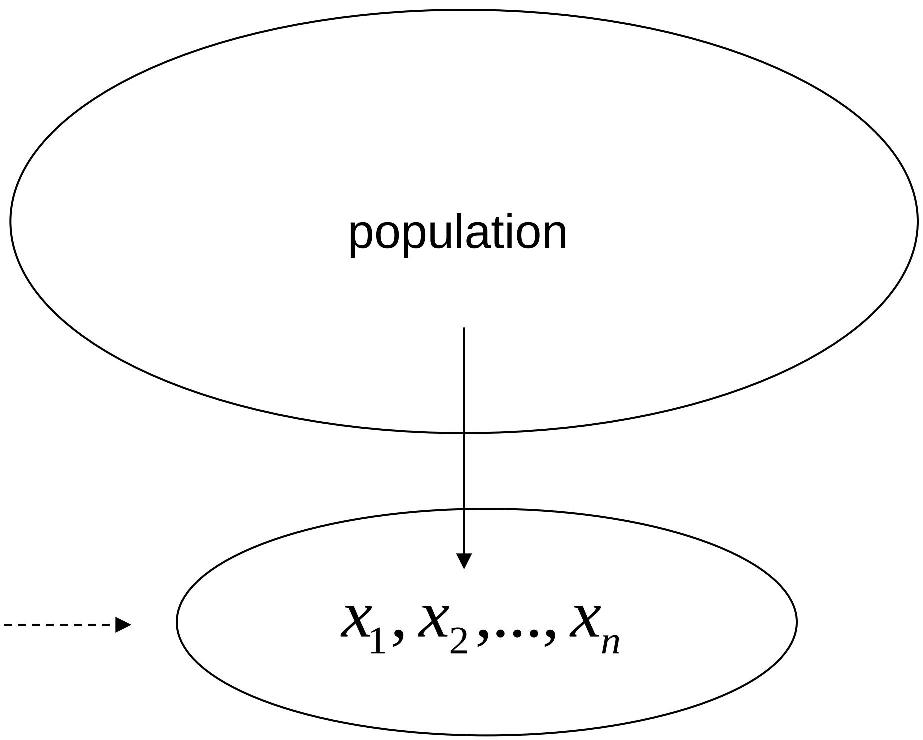


The population and the sample

$$F = N(\boxed{\mu,}\sigma^2)$$

$$\boxed{\mu} = E_F(x)$$

$$\boxed{\hat{\mu}} = \frac{1}{n}\sum_{i=1}^{n} x_i$$



Classical data analysis methods

parameter estimates

true parameter

$$\boxed{\hat{\mu}} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \longleftrightarrow \quad \boxed{\mu} = E_F(x)$$

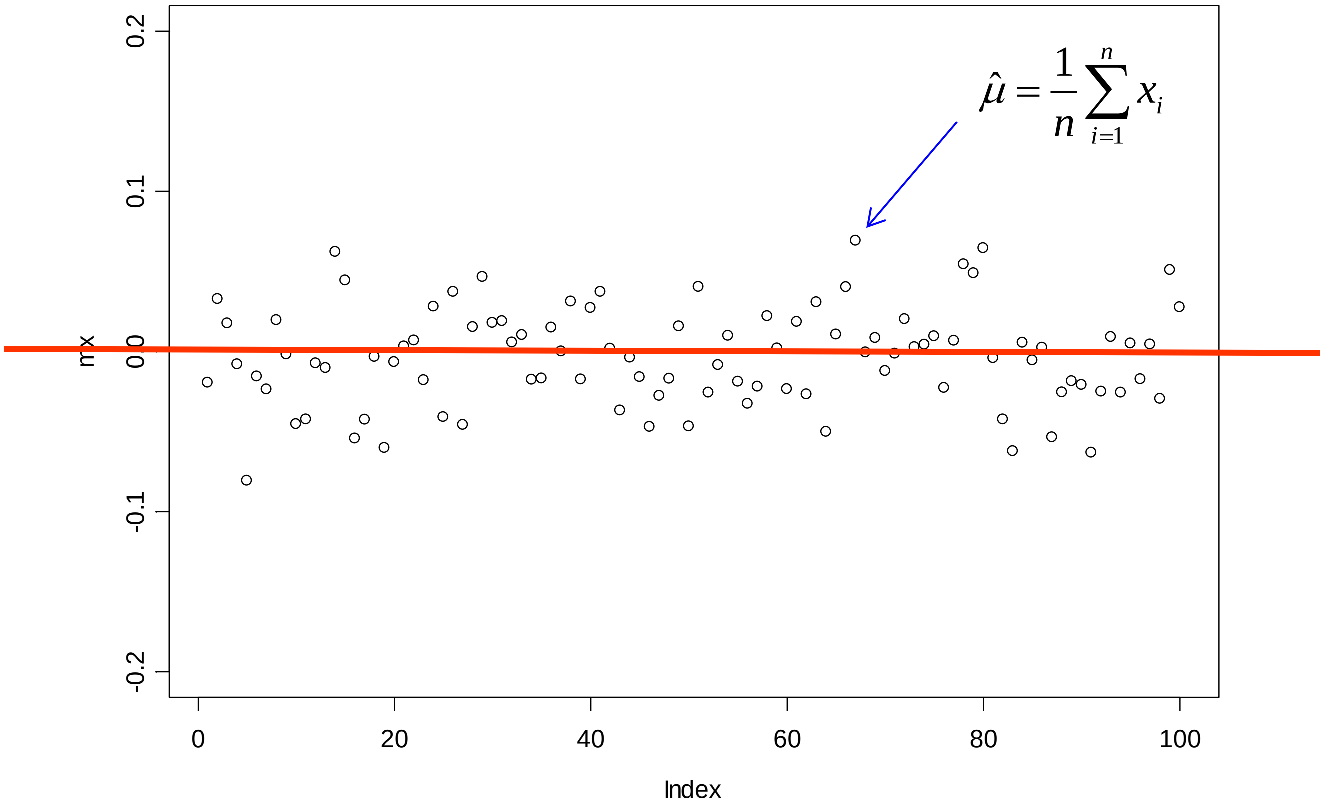
Use theoretical properties about <u>the distribution</u> of the parameter estimates

What is we are wrong ?

100 samples of size 50 from N(0,1)

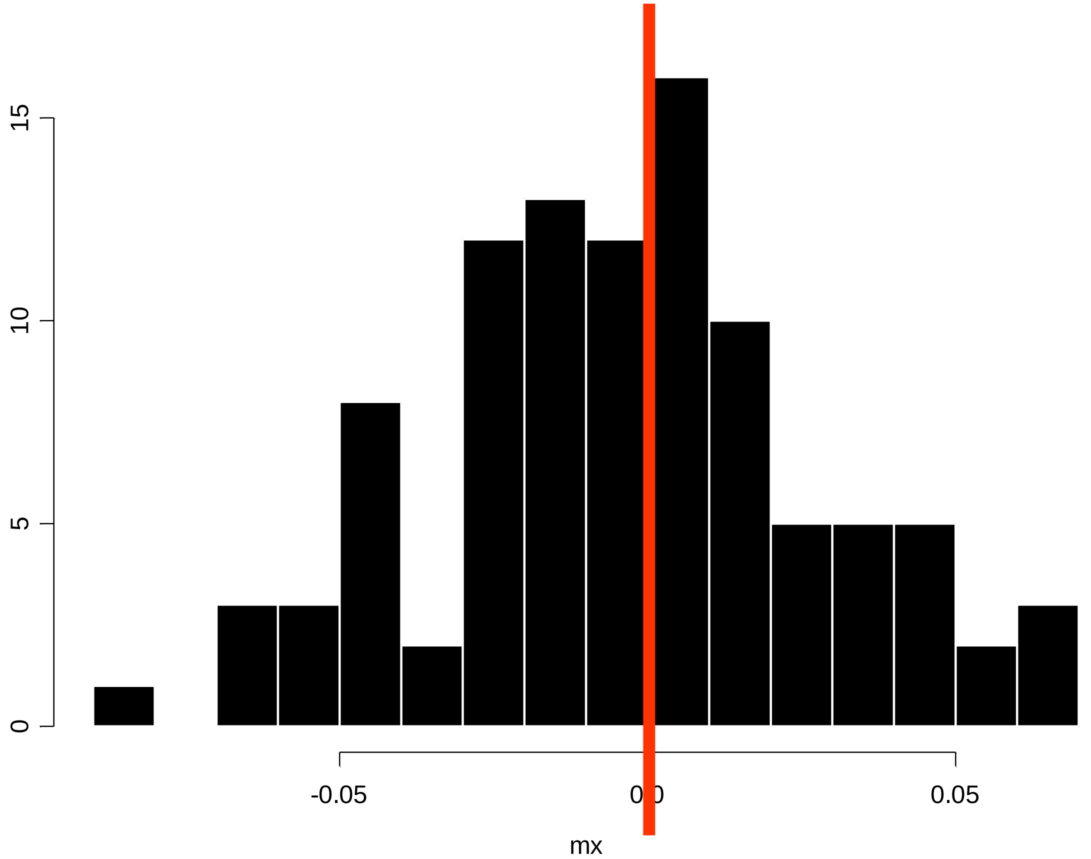
Each point in the figure represent the sample mean.

$$\mu = E_F(x)$$



100 samples of size 50 from N(0,1)

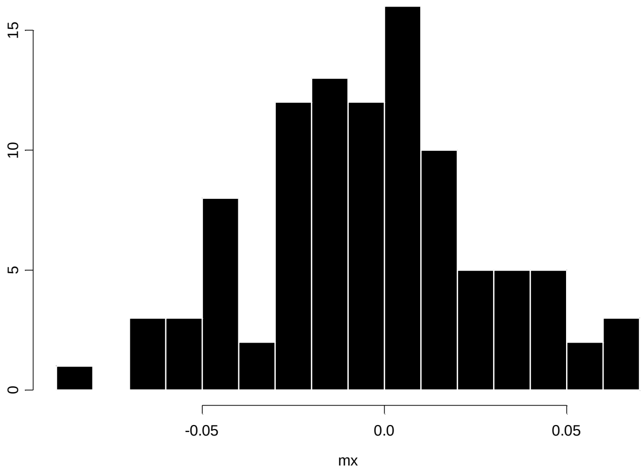
$$\mu = E_F(x)$$



Distribution of the sample mean.

The main concept of the course

100 samples



Reality : one sample

?

How can we approximate the distribution of the sample mean?

BOOTSTRAP

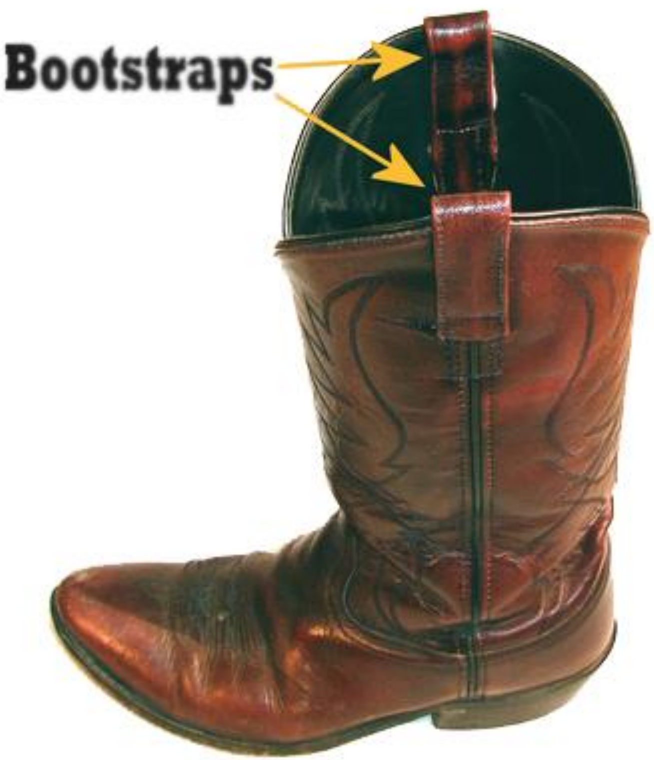
The main concept of the course

Reality : one sample

BOOTSTRAP

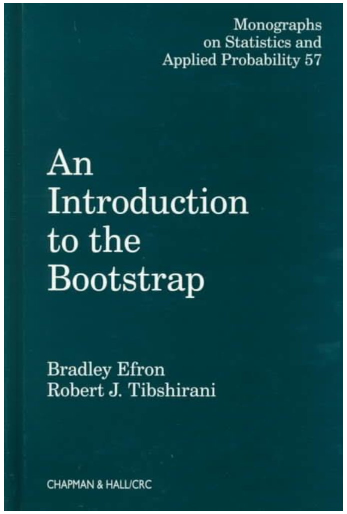
- Application for:
  - Estimation and C.I.
  - Inference.
  - Modeling.

The name



if you are stuck in the mud, you can use the bootstrap to pool yourself out of the mud.

The accuracy of the sample mean



Chapter 2

The mouse data

- A small randomized experiment were done with 16 mouse, 7 to treatment group and 9 to control group.
- Treatment was intended to prolong survival after a test surgery.
- In R:

```
> help(mouse.c)
```

Install first the R package bootstrap:

```
> library(bootstrap)
```

```
> mouse.c
[1]  52 104 146  10  50  31  40  27  46
> mouse.t
[1]  94 197  16  38  99 141  23
```

The mouse data

```
> mean(mouse.c)
[1] 56.22222
> sqrt(var(mouse.c)/9)
[1] 14.13897


> mean(mouse.t)
[1] 86.85714
> sqrt(var(mouse.t)/7)
[1] 25.23549
```

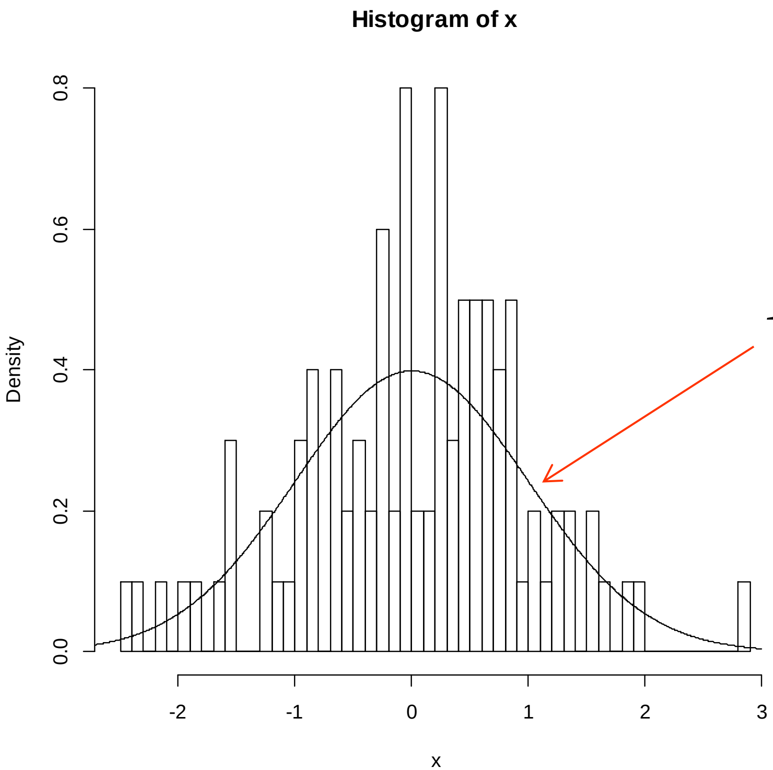
Sample mean and estimated standard error for the mean:

$$\bar{x}$$

$$SE(\bar{x}) = \sqrt{\frac{s^2}{n}}$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

A random sample from N(0,1), n=100



$N(0,1)$

$$F = N(\mu = 0, \sigma^2 = 1)$$

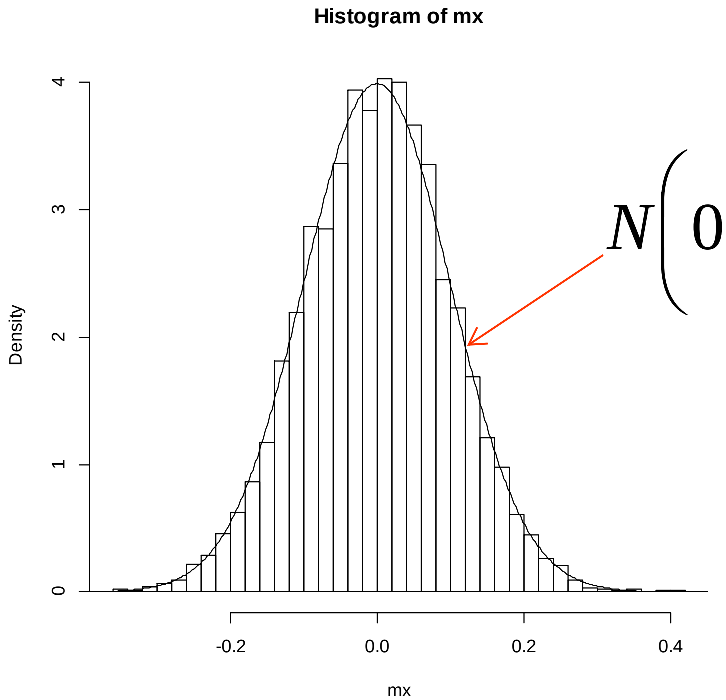
$$F \rightarrow (x_1, x_2, \ldots, x_{100})$$

$$Var(\bar{x}) = \frac{\sigma^1}{n} = \frac{1}{100}$$

```
> x<-rnorm(100,0,1)
> hist(x,nclass=50,probability=TRUE)
> mean(x)
[1] 0.03995677
```

$\bar{x}$

Sample means of 10000 samples from N(01,), n=100



$$N\left(0, \frac{1}{100}\right)$$

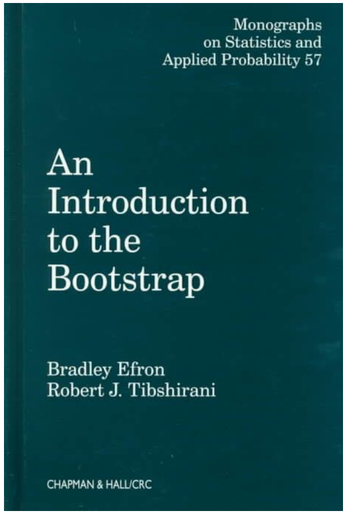
```
> mx<-c(1:10000)
> for(i in 1:10000)
+ {
+ x<-rnorm(100,0,1)
+ mx[i]<-mean(x)
+ }
>
> hist(mx,nclass=50,probability=TRUE)
> mean(mx)
[1] 0.0001821198
> var(mx)
[1] 0.009956565
```

$$\mathrm{var}(\bar{X}) = \frac{1}{100} = \frac{\sigma^2}{n}$$

The variance of the 10000 sample means.

- Histogram: 10000 values of the sample means.
- True mean: 0.

Random sample for a population



Chapter 3

The low school data

- A population of 82 USA law schools.
- Two measurements:
  - LSAT (average score on a national law test).
  - GPA (average undergraduate grade-point average).
- In R:

```
> help(law)
```

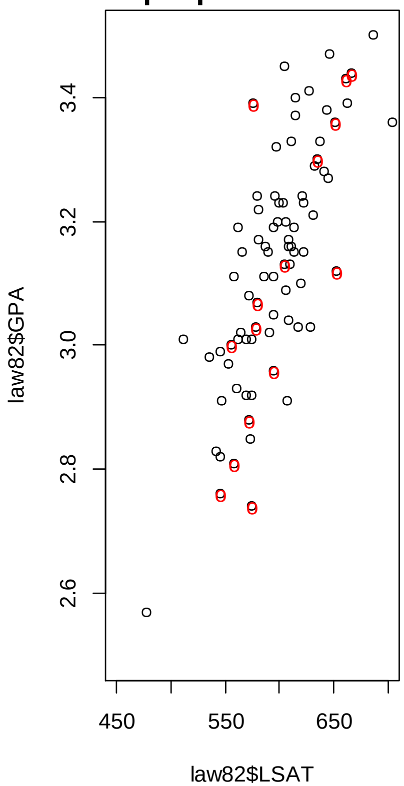
- Observation unit:

$$x_i = \left(LAST_i, GPA_i\right)$$

A random sample for the population

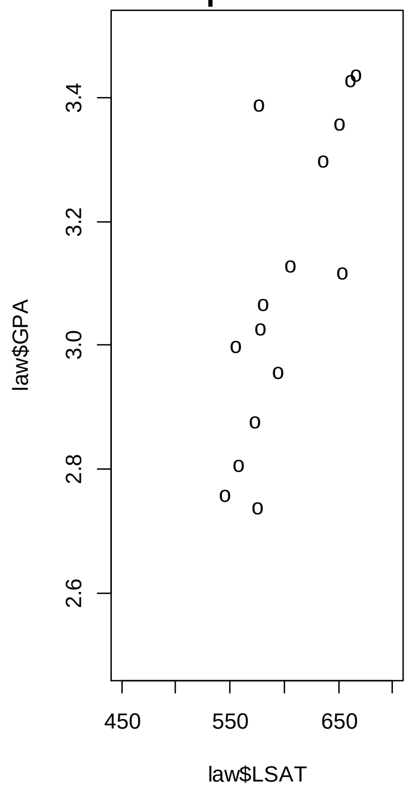
- A random sample of size *n=15* from the population of 82 USA law schools.

population



$$\rho\left(LAST_i, GPA_i\right) = 0.759$$

sample

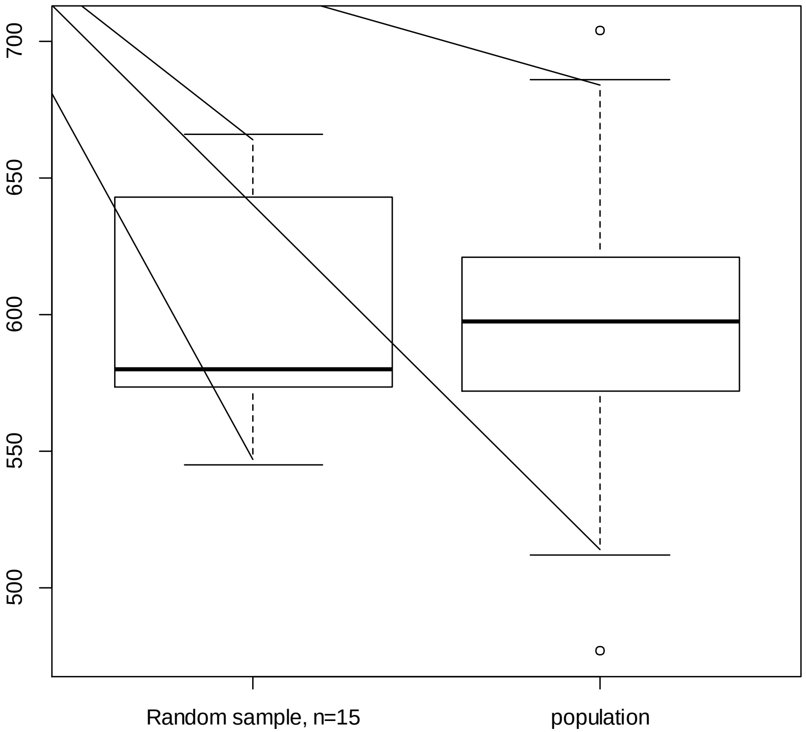


$$\hat{\rho}\left(LAST_i, GPA_i\right) = 0.776$$

```
> law
   LSAT  GPA
1   576 3.39
2   635 3.30
3   558 2.81
4   578 3.03
5   666 3.44
6   580 3.07
7   555 3.00
8   661 3.43
9   651 3.36
10  605 3.13
11  653 3.12
12  575 2.74
13  545 2.76
14  572 2.88
15  594 2.96
> cor(law82$LSAT,law82$GPA)
[1] 0.7599979
> cor(law$LSAT,law$GPA)
[1] 0.7763745
```

Sample of 15 schools

Population and sample mean of LAST



$$Population(N=82):$$

$$\mu = 597.54$$

$$\mathrm{var}(LAST) = \sigma^2 = 1481.337$$

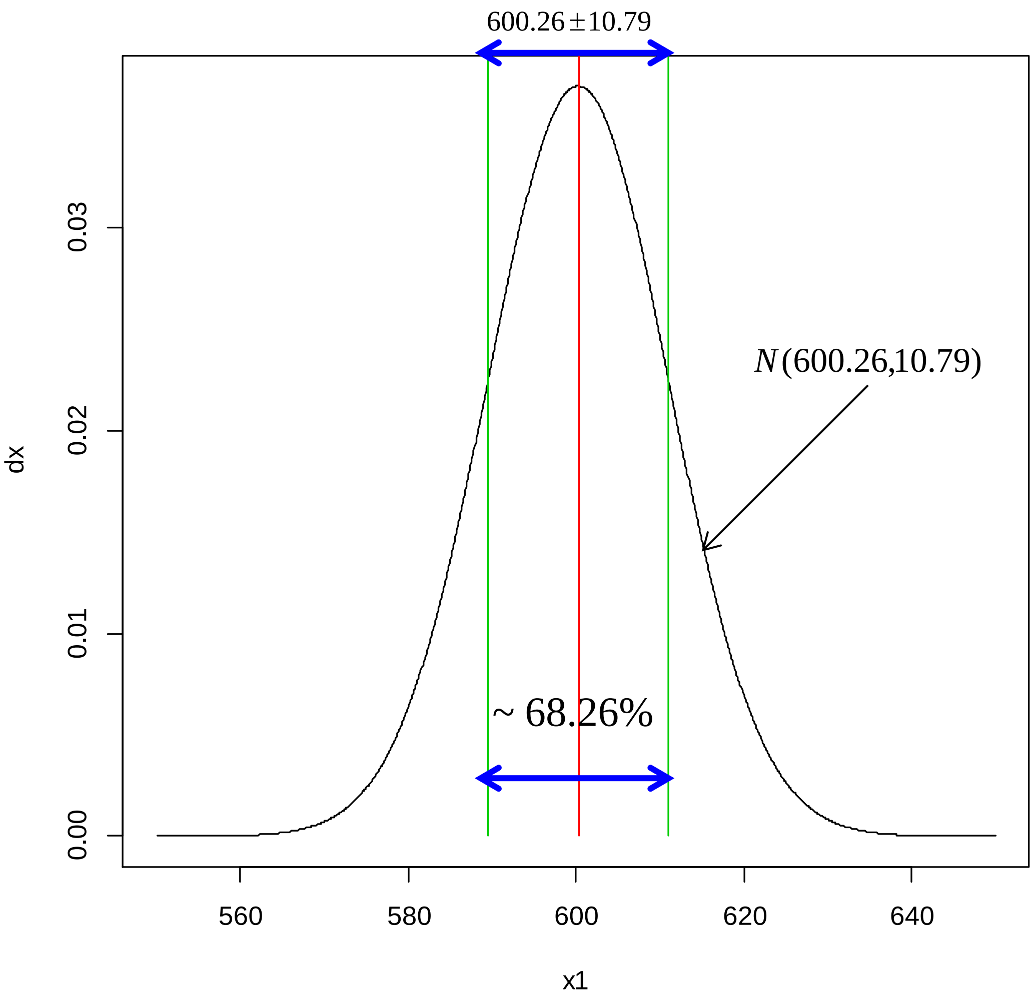
$$Sample(n=15):$$

$$\bar{x} = 600.26$$

$$SE(\bar{x}) = 10.79 = \sqrt{\frac{s^2}{n}}$$

```
> boxplot(law$LSAT,law82$LSAT,names=c("Random sample, n=15","population"))
> mean(law$LSAT)
[1] 600.2667
> sqrt(var(law$LSAT)/15)
[1] 10.7913
> mean(law82$LSAT)
[1] 597.5488
```

Population and sample mean of LAST



$$\mu = 597.54$$

$$\bar{x} = 600.26$$

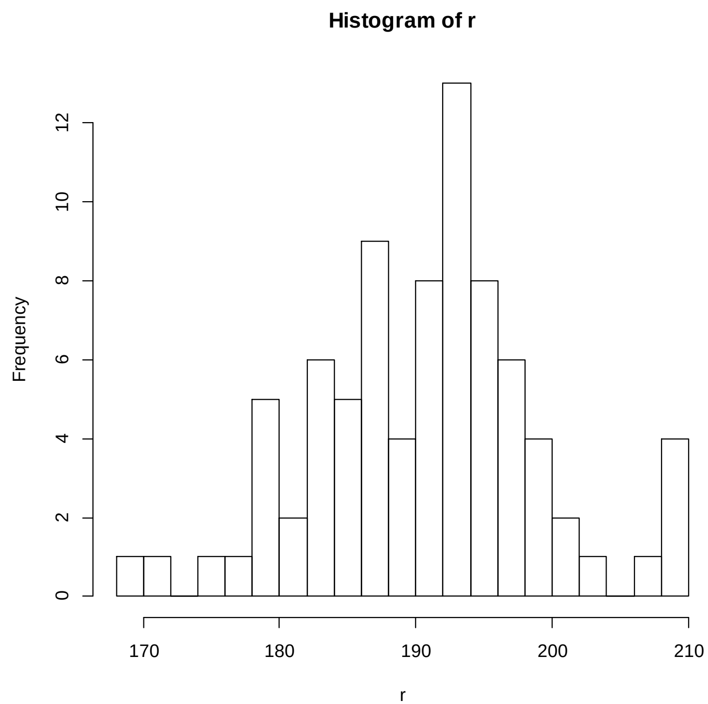
$$SE(\bar{x}) = 10.79$$

The ratio of LAST and GPA

$$x_i = (LAST_i, GPA_i) = (y_i, z_i)$$

$$r_i = \frac{y_i}{z_i}$$

$$\mu_r = E(LAST / GPA) = 190.74$$

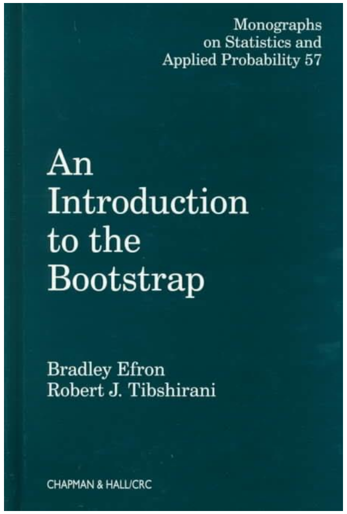


- What is the distribution of the ratio ?
- What is the standard error of the ratio ?

- See example in a later chapter.

```
> r<-law82$LSAT/law82$GPA
> hist(r,nclass=25)
> mean(r)
[1] 190.7476
```

The empirical distribution function and the plug-in principle



Chapter 4

The probability distribution

Let X be a random variable such that

$$X \sim F(\theta)$$

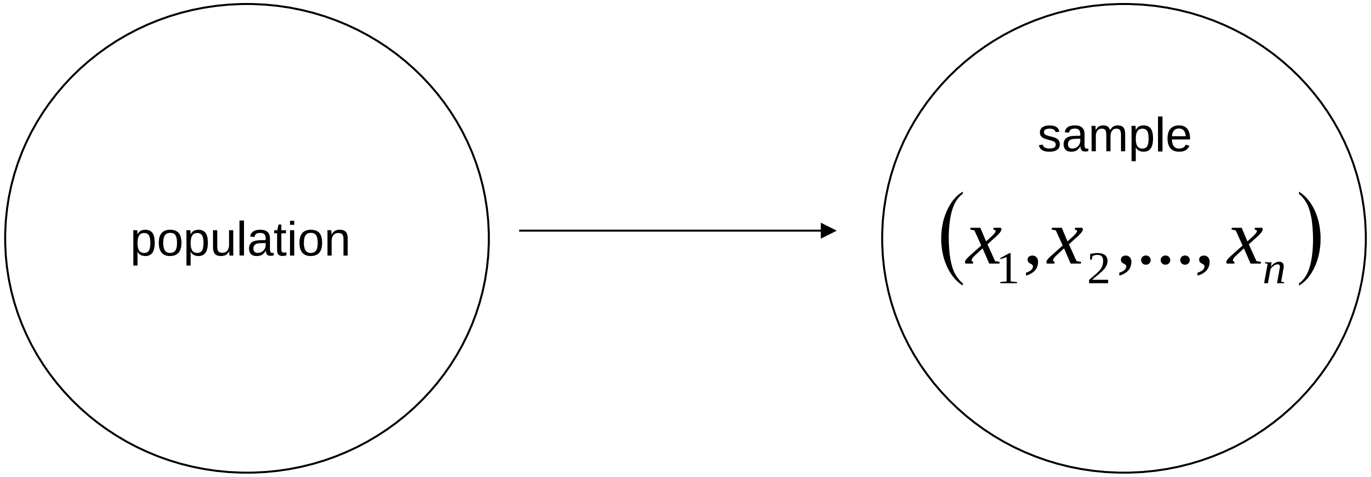
F is the probability distribution of X

θ is an unknown parameter

A random Sample from F

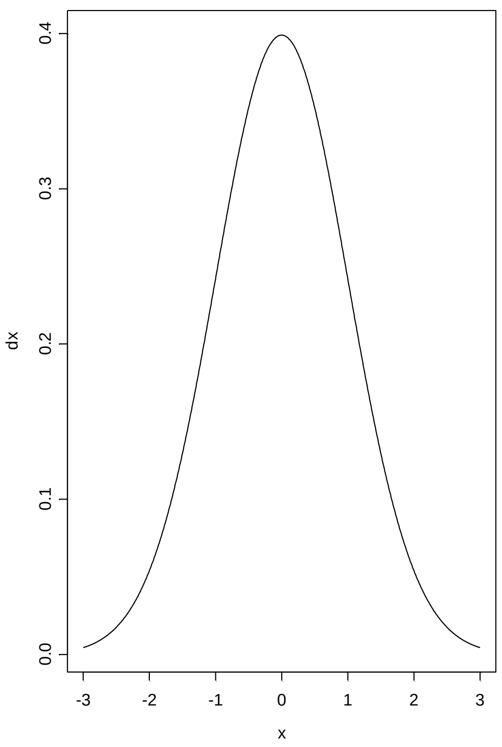
We observed a random sample from the probability distribution F

$$F \to (x_1, x_2, \ldots, x_n)$$

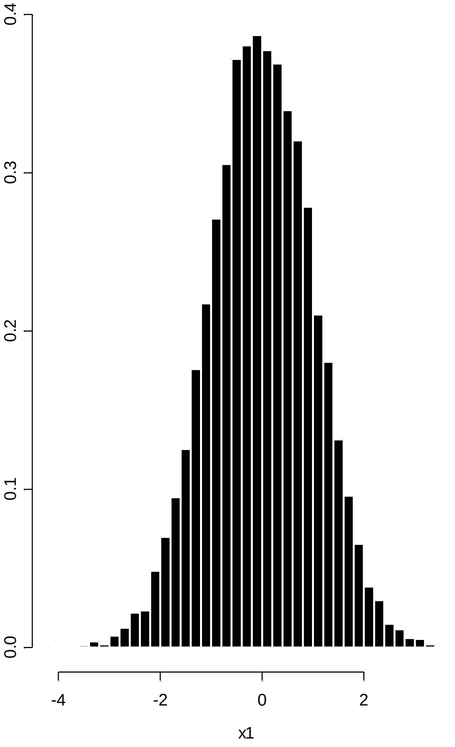


Example: a random sample from N(0,1)

N(0,1)



A sample of 10000 observations from N(0,1)



The empirical distribution

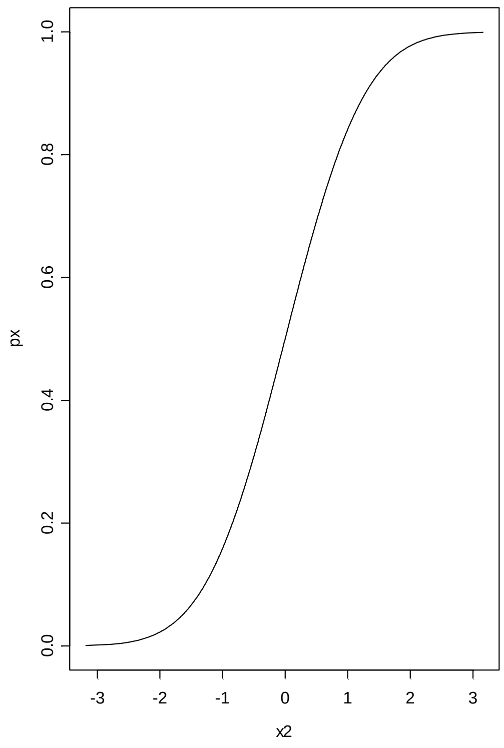
The empirical distribution function is defined to be the discrete distribution that puts probability of 1/n on each value of x_i

$$F \to \left(x_1, x_2, \ldots, x_n\right)$$

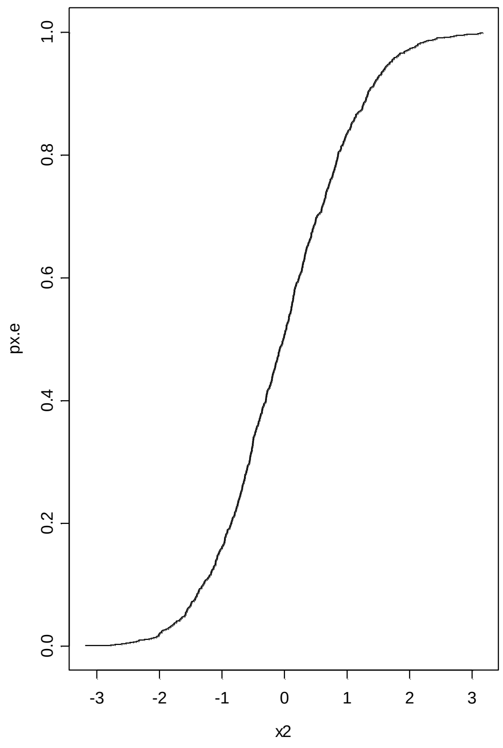
$$P(A) = \hat{F} = \frac{\#(x_i \in A)}{n}$$

Example: a random sample from N(0,1)

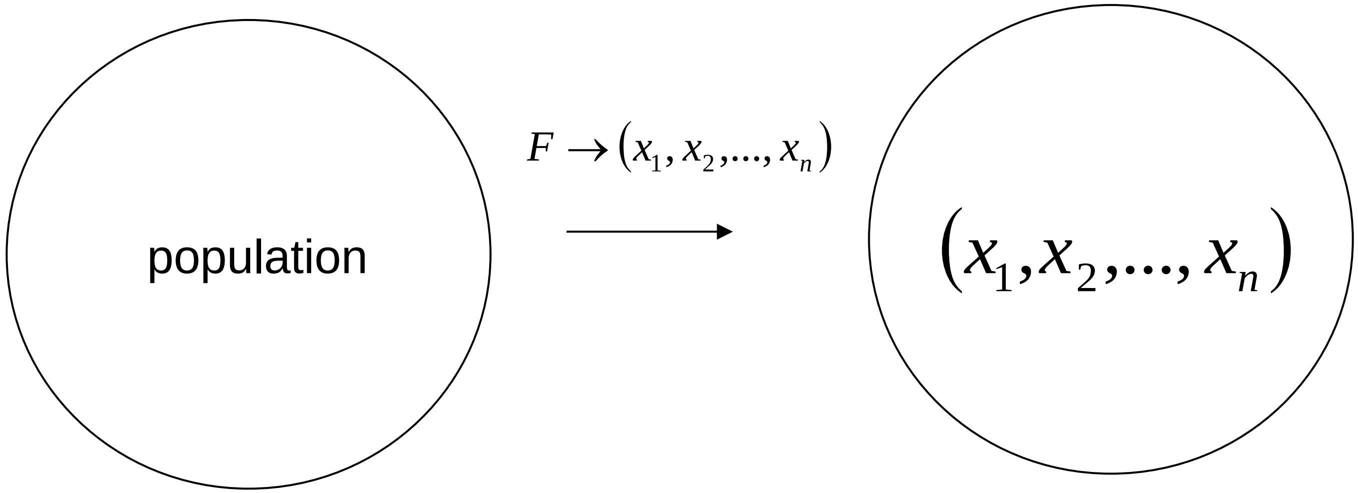
The probability distribution N(0,1)



The empirical probability distribution of a sample (n=500)



The empirical distribution



F

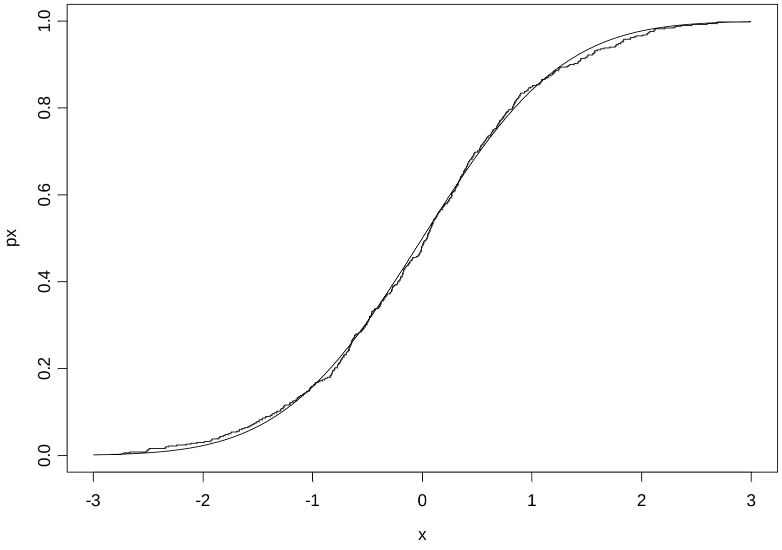
$\hat{F}$

The probability distribution

The empirical probability distribution

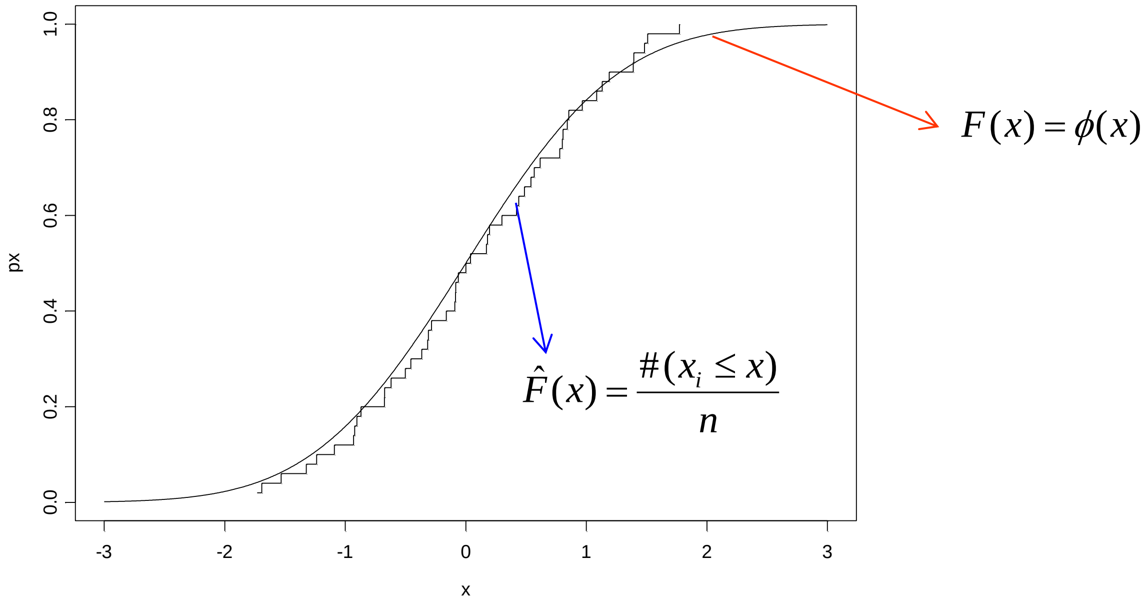
A sample of 500 observations from N(0,1)

The number of x_i in the sample that are smaller or equal to x



$$\hat{F}(x) = \frac{\#(x_i \le x)}{n}$$

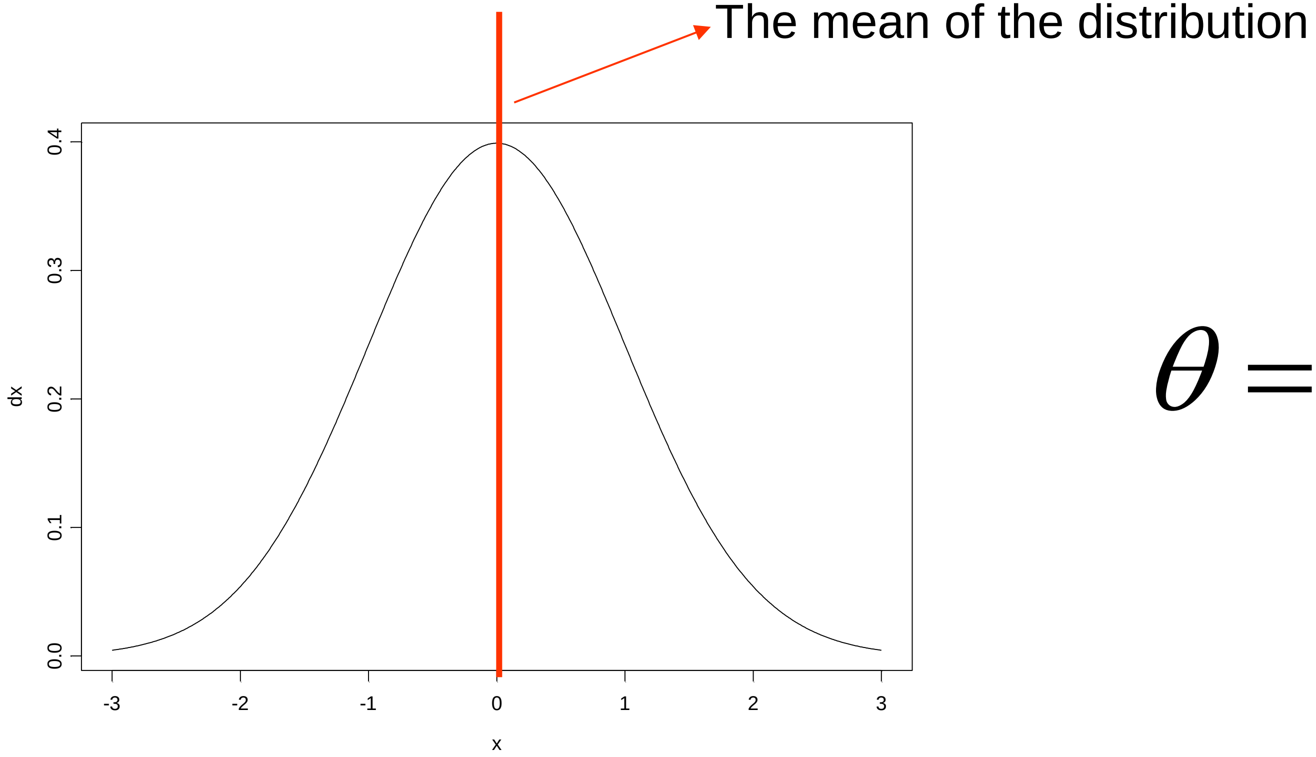
A sample of 50 observations from N(0,1)



The cumulative distribution of N(0,1)

A parameter

- A parameter θ is a function of the probability distribution F



$$\theta = t(F)$$

A statistic

parameter

$$\theta = t(F)$$

A random sample from F

$$F \rightarrow \left(x_1, x_2, \ldots, x_n\right)$$

A statistic is a function of the observed sample **x**

$$\hat{\theta} = t(\hat{F})$$

The mean for B(n,p)

$$F = B(n, p)$$

sample

$$X_i = \begin{cases} 1 & p \\ 0 & 1-p \end{cases}$$

$$\underbrace{(X_1, X_2, \ldots, X_n) \qquad X = \sum_{i=1}^{n} X_i}_{X \sim B(n,p)}$$

$$\theta = E_F(x)$$

$$\theta = t(F) = np$$

$$\hat{\theta} = t(\hat{F}) = n\hat{p} = n \times \frac{x}{n}$$

The plug-in principle

The plug-in estimate of the parameter

$$\theta = t(F)$$

is defined as

$$\hat{\theta} = t(\hat{F})$$

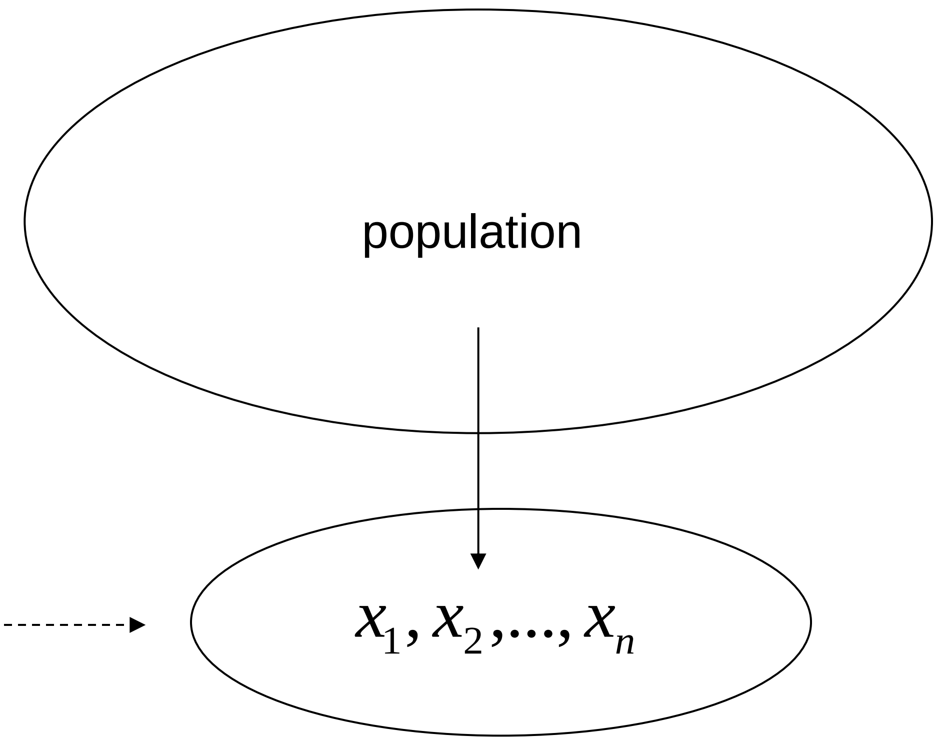
We use the same function from F, $t(F)$ on the empirical distribution

The population mean and the parameter estimate from the sample

$$F = N(\mu, \sigma^2)$$

$$\mu = E_F(x)$$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$



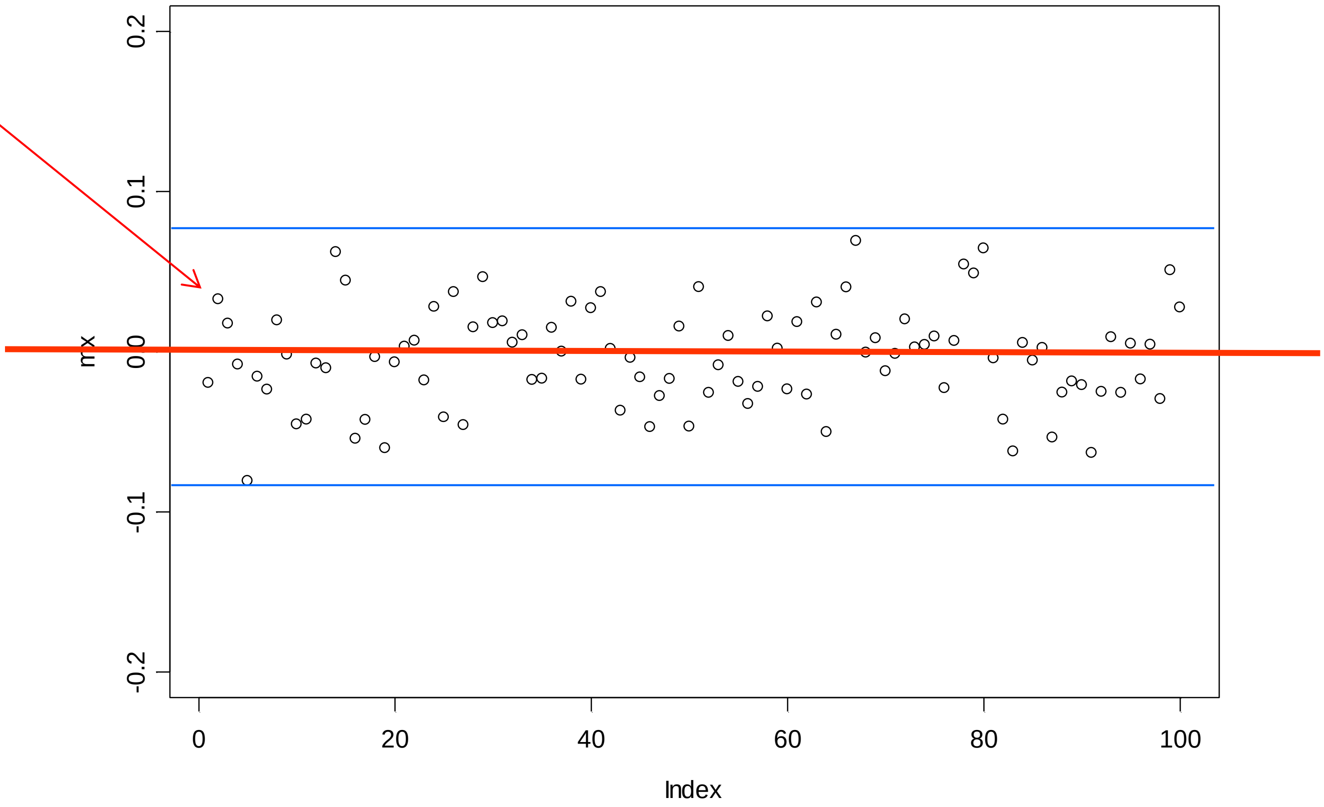
$$x_1, x_2, \ldots, x_n$$

100 samples of size 50 from N(0,1)

$$(x_1, x_2, \ldots, x_{50}) \sim N(0,1) \Leftrightarrow F_{N(0,1)} \to (x_1, x_2, \ldots, x_{50})$$

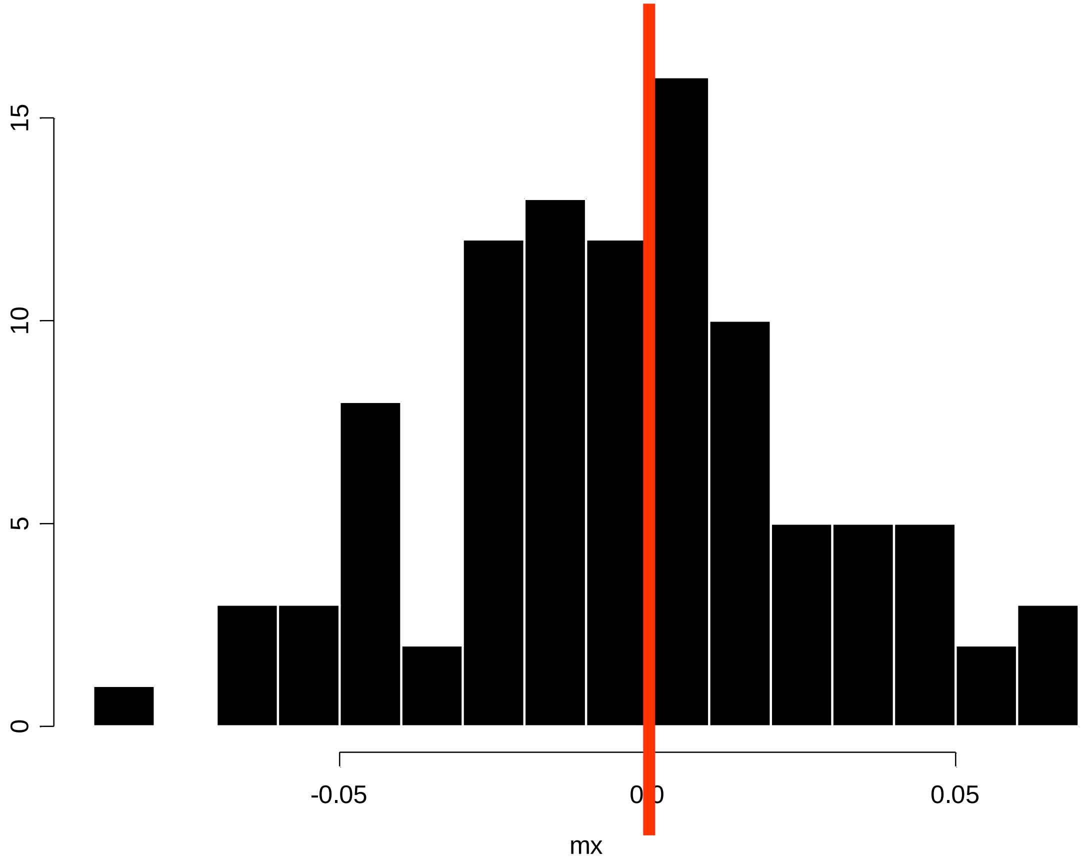
$$\bar{x} = \sum_{i=1}^{50} x_i$$

$$\mu = E_F(x)$$



100 samples of size 50 from N(0,1)

$$\mu = E_F(x)$$



The standard error of the sample mean (1)

population

$$X \sim (\mu_F, \sigma_F{}^2)$$

$$\mu_F = E_F(x)$$

$$\sigma_F^2 = Var_F = E_F[(x - \mu_F)^2]$$

sample

$$F \rightarrow \left(x_1, x_2, \ldots, x_n\right)$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$E(\bar{x}) = \frac{1}{n}E(\sum_{i=1}^{n} x_i) = \mu_F$$

$$\hat{\sigma}_F = \left\{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right\}^{\frac{1}{2}}$$

The standard error of the sample mean (2)

population

$$X \sim (\mu_F, \sigma_F{}^2)$$

$$\sigma_F^2 = E_F[(x - \mu_F)^2]$$

sample

$$F \rightarrow (x_1, x_2, \ldots, x_n)$$

$$Var(\bar{x}) = \frac{1}{n^2} Var(\sum_{i=1}^{n} x_i) = \frac{\sigma_F{}^2}{n}$$

$$SE(\bar{x}) = \frac{\sigma_F}{\sqrt{n}}$$

$$\hat{S}E(\bar{x}) = \frac{\hat{\sigma}_F}{\sqrt{n}} = \frac{s}{\sqrt{n}}$$

Distribution of the sample mean

population

$$X \sim (\mu_F, {\sigma_F}^2)$$

sample

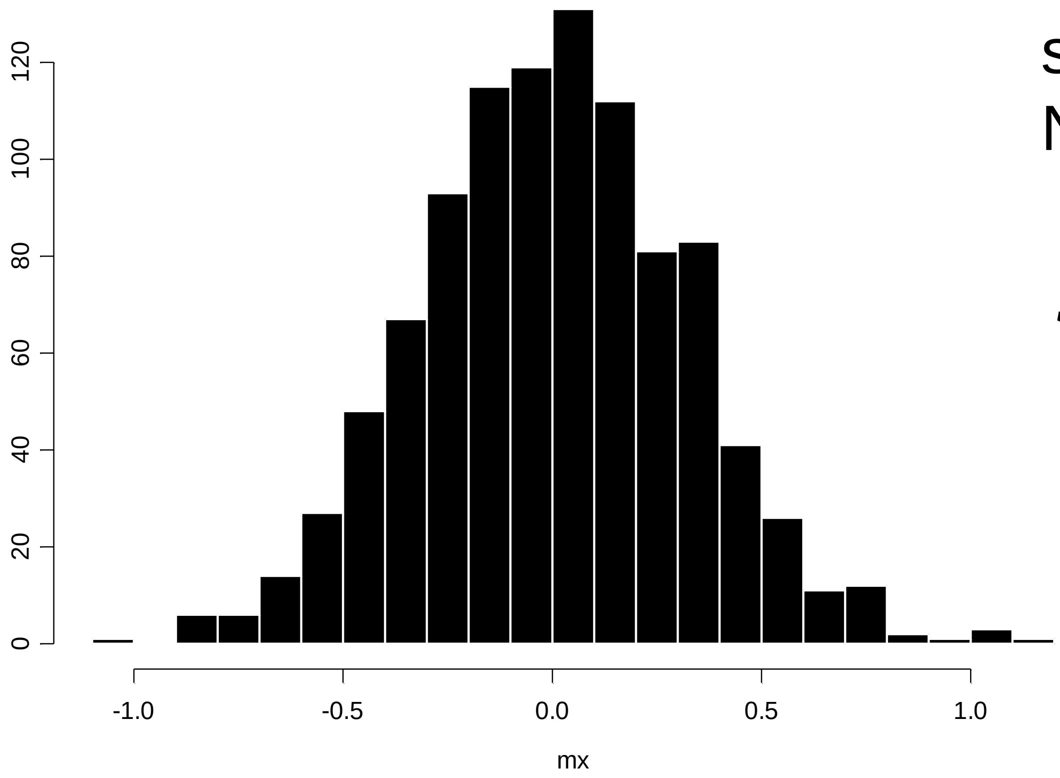
$$F \rightarrow (x_1, x_2, \ldots, x_n)$$

For large n:

$$\bar{x} \sim N(\mu_F, \frac{\sigma_F^2}{n})$$

Is it always the case ?
What does it mean “large n” ?

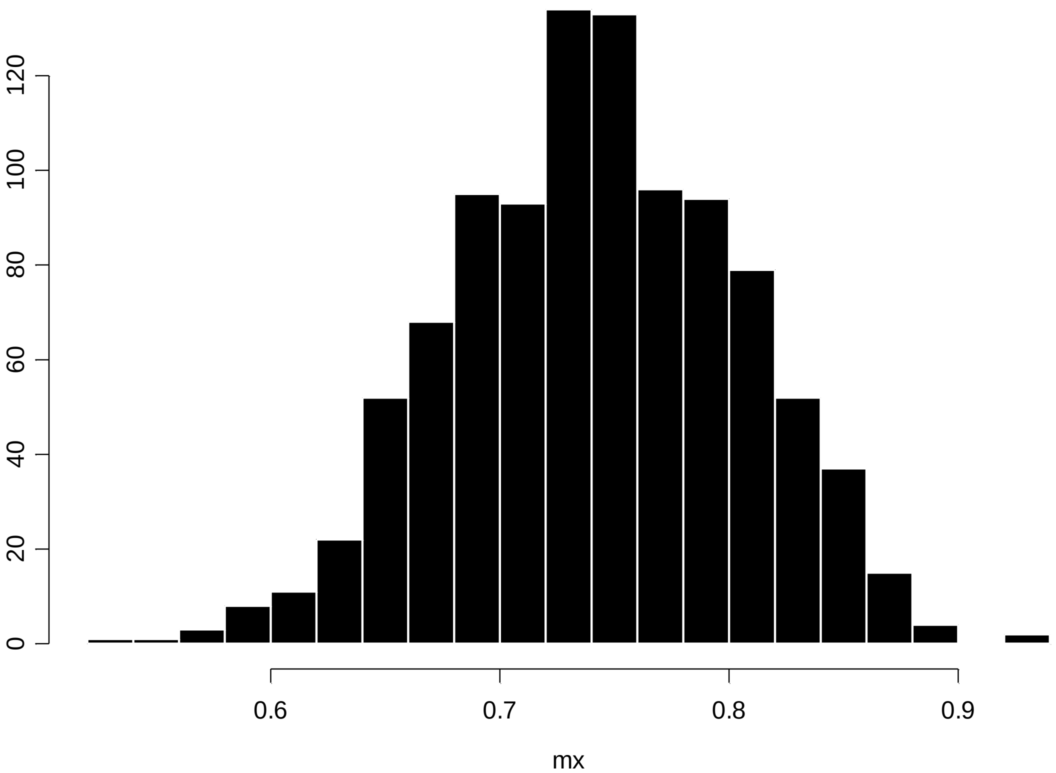
Distribution of the sample mean



1000 samples of size 50 from N(0,1)

$$SE(\bar{x}) = \frac{\sigma_F}{\sqrt{n}} = \frac{1}{\sqrt{50}}$$

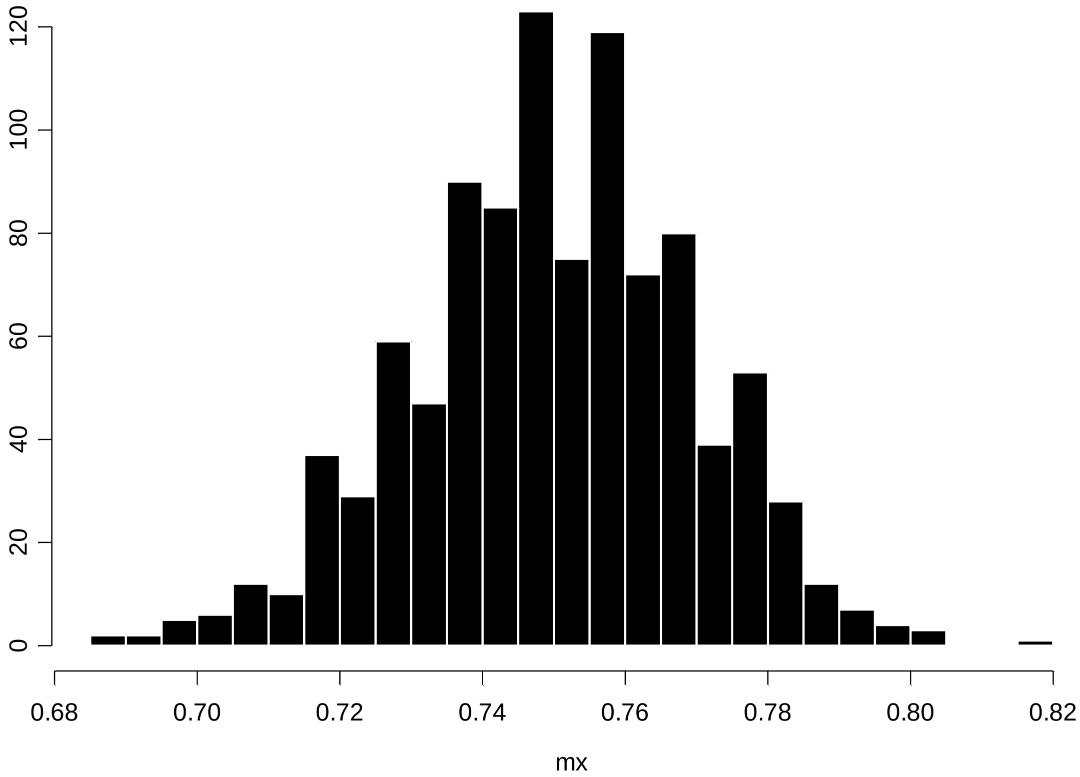
Distribution of the sample mean



1000 samples of size 50 from B(1,0.75)

$$SE(\bar{x}) = \frac{\sigma_F}{\sqrt{n}} = \frac{0.75 \times 0.25}{\sqrt{50}}$$

Distribution of the sample mean



1000 samples
of size 500
from B(1,0.75)

$$SE(\bar{x}) = \frac{\sigma_F}{\sqrt{n}} = \frac{0.75 \times 0.25}{\sqrt{500}}$$

R code: the empirical distribution

```
par(mfrow=c(1,1))
x<-seq(from=-3,to=3,length=1000)
dx<-dnorm(x,0,1)
plot(x,dx,type="l")

x1<-rnorm(10000,0,1)
hist(x1,nclass=50,col=1,probability=T)

par(mfrow=c(1,2))
x2<-rnorm(1000,0,1)
x2<-sort(x2)
px<-pnorm(x2,0,1)
plot(x2,px,type="l")
n<-length(x2)
px.e<-c(1:length(x2))/n
plot(x2,px.e,type="s")
```

R code: the sample mean from N(0,1)

```
par(mfrow=c(1,1))
x2<-rnorm(50,0,1)
x2<-sort(x2)
x<-seq(from=-3,to=3,length=1000)
px<-pnorm(x,0,1)
plot(x,px,type="l")
n<-length(x2)
px.e<-c(1:length(x2))/n
lines(x2,px.e,type="s")
```

R code for 1000 samples of size 10 from N(0,1)

```
nsim<-1000
mx<-c(1:nsim)
for(i in 1:nsim)
{
x<-rnorm(10,0,1)
mx[i]<-mean(x)
}
```

1000 samples out of the population

population

$$F = N(\mu_F = 0, \sigma_F{}^2 = 1)$$

Sample of 10 obs. out of the population

$$F_{N(0,1)} \rightarrow (x_1, x_2, ,,,, x_n)$$

```
hist(mx,nclass=20,col=1)
```

R code: sample mean from B(1,0.75)

```
nsim<-1000
mx<-c(1:nsim)
for(i in 1:nsim)
{
x<-rbinom(500,1,0.75)
mx[i]<-mean(x)
}

hist(mx,nclass=20,col=1)
```

1000 samples

$$F = B(n = 500, p = 0.75)$$

$$F_{B(500,0.75)} \rightarrow \left(x_1, x_2 .,,,.x_{500}\right)$$

$\hat{p}$