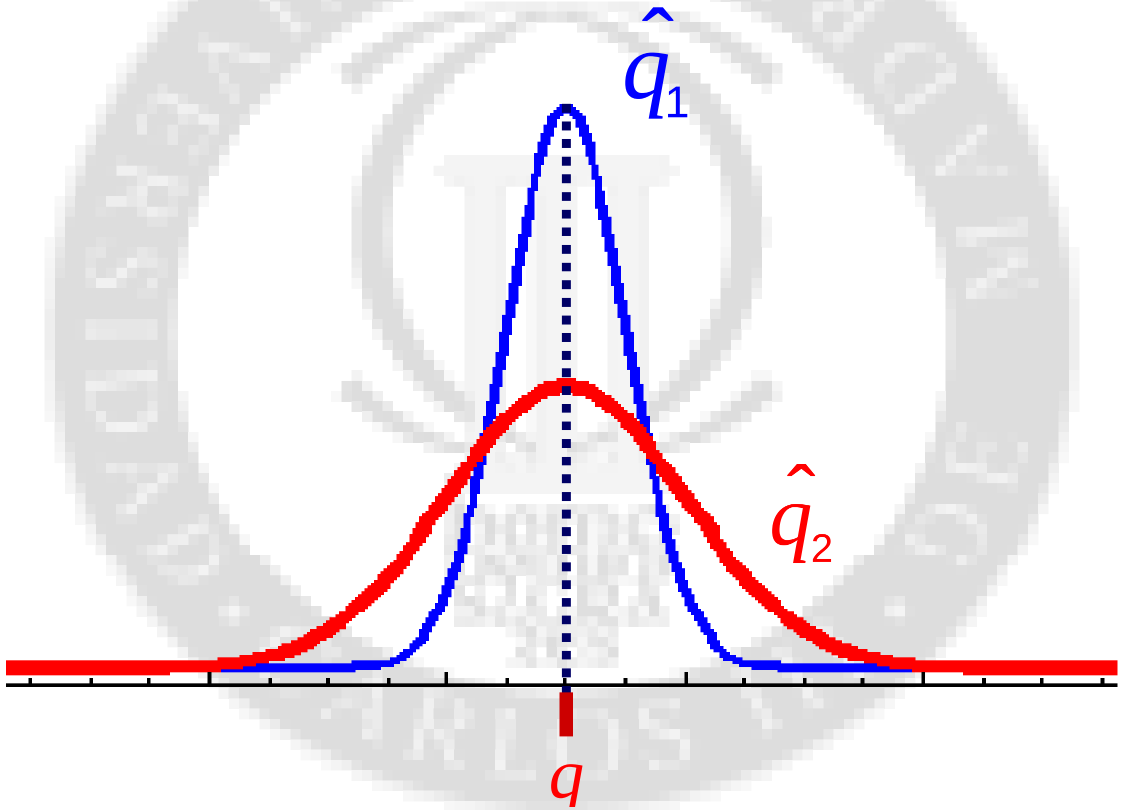
VI. Introduction to statistical inference



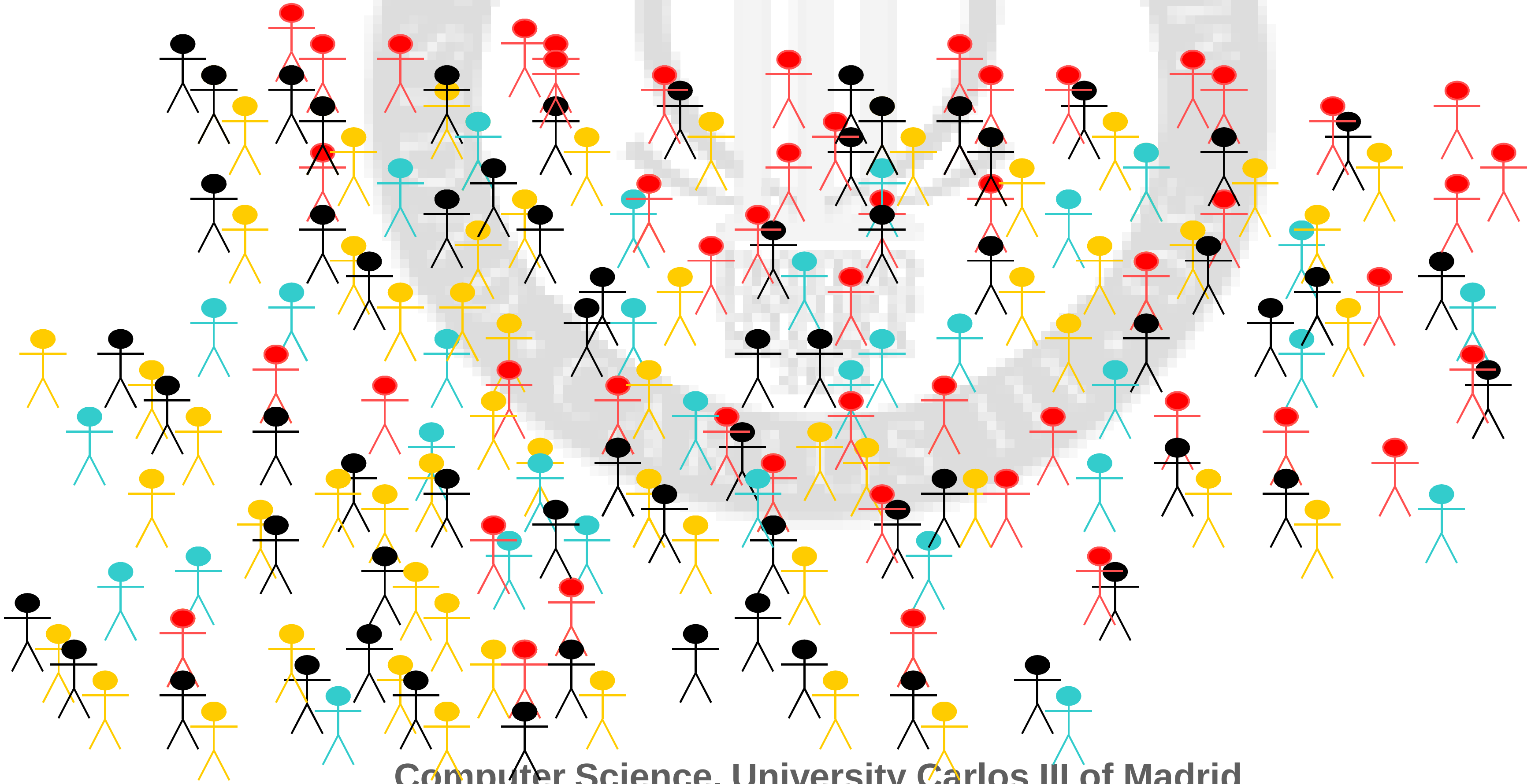
Chapter 5: Introduction to statistical inference

1. **Statistical inference. Population and sample**
2. **Sampling distribution of a statistic**
3. **The sample mean distribution**
4. **Estimation and estimators**
5. **Method of moments**
6. **Diagnosis of the model**
7. **Transformations that improve normality**

1. Statistical inference. Population and sample

Objective of the statistics:

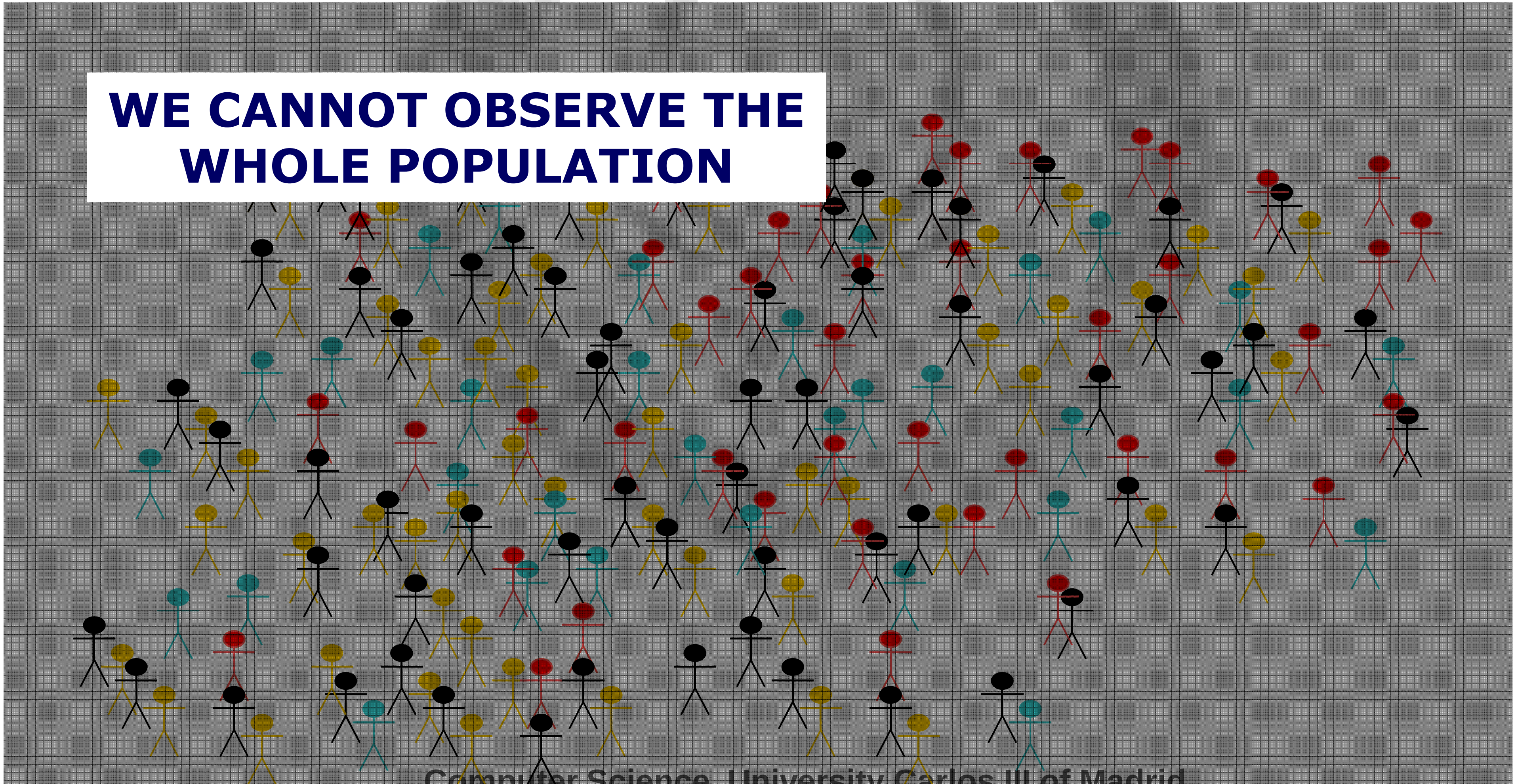
- Learning from the data
- To generalize the information <u>of a sample</u> of the variable X <u>to the population</u>



1. Statistical inference. Population and sample

Objective of the statistics:

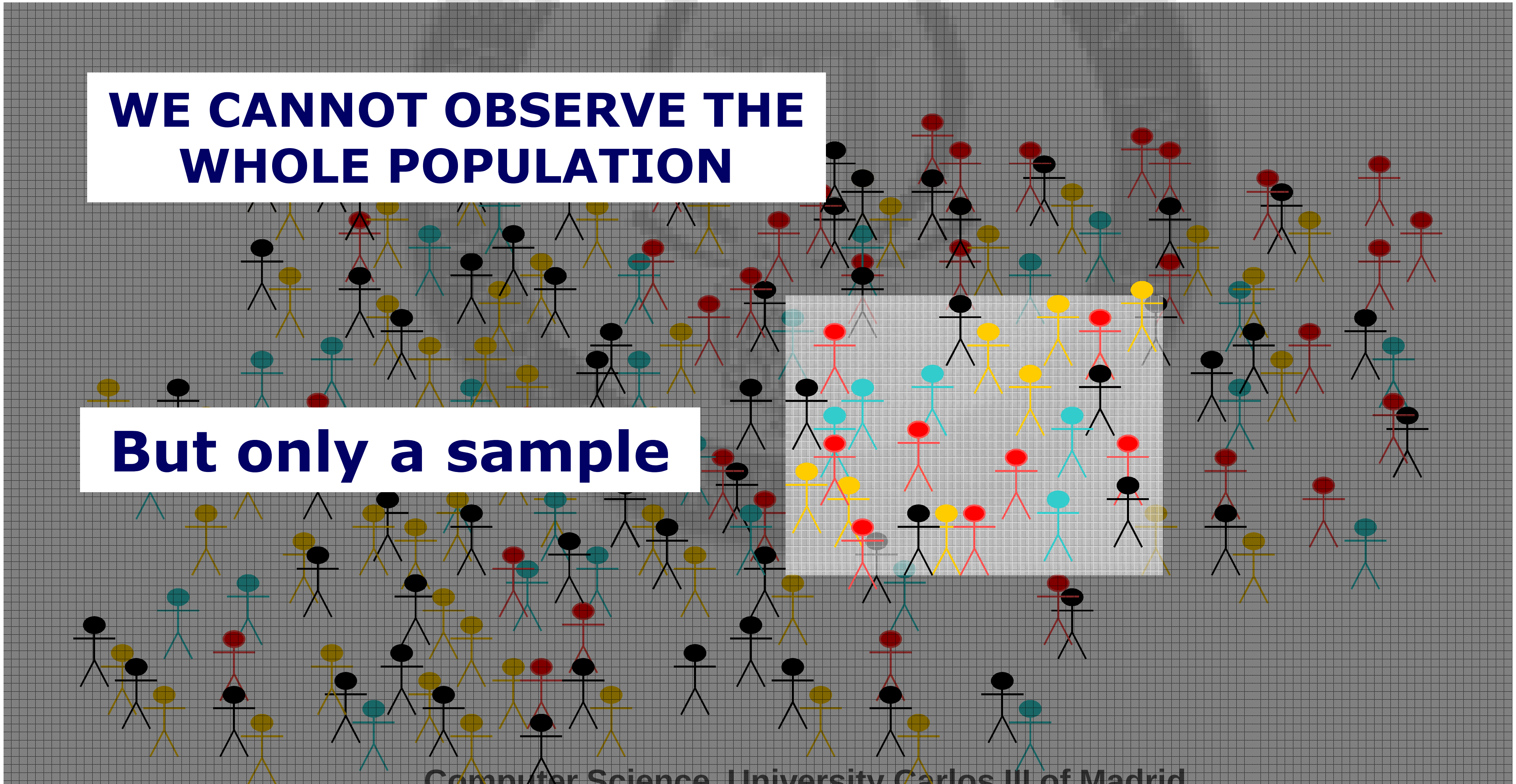
- Learning from the data
- To generalize the information of a sample of a variable X to the population



1. Statistical inference. Population and sample

Objective of the statistics:

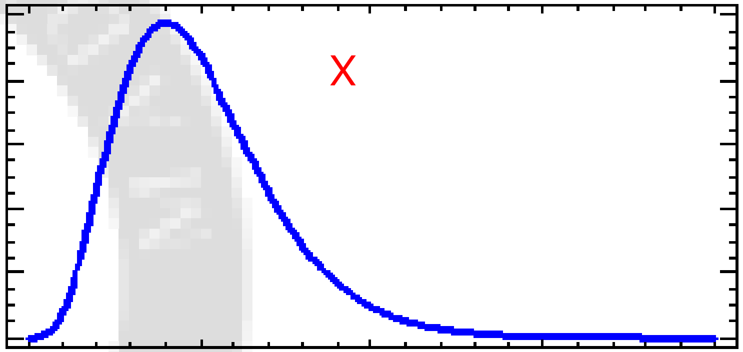
- Learning from the data
- To generalize the information of a sample of a variable X to the population



INFERENCE

SAMPLE of **n** OBSERVATIONS → POPULATION

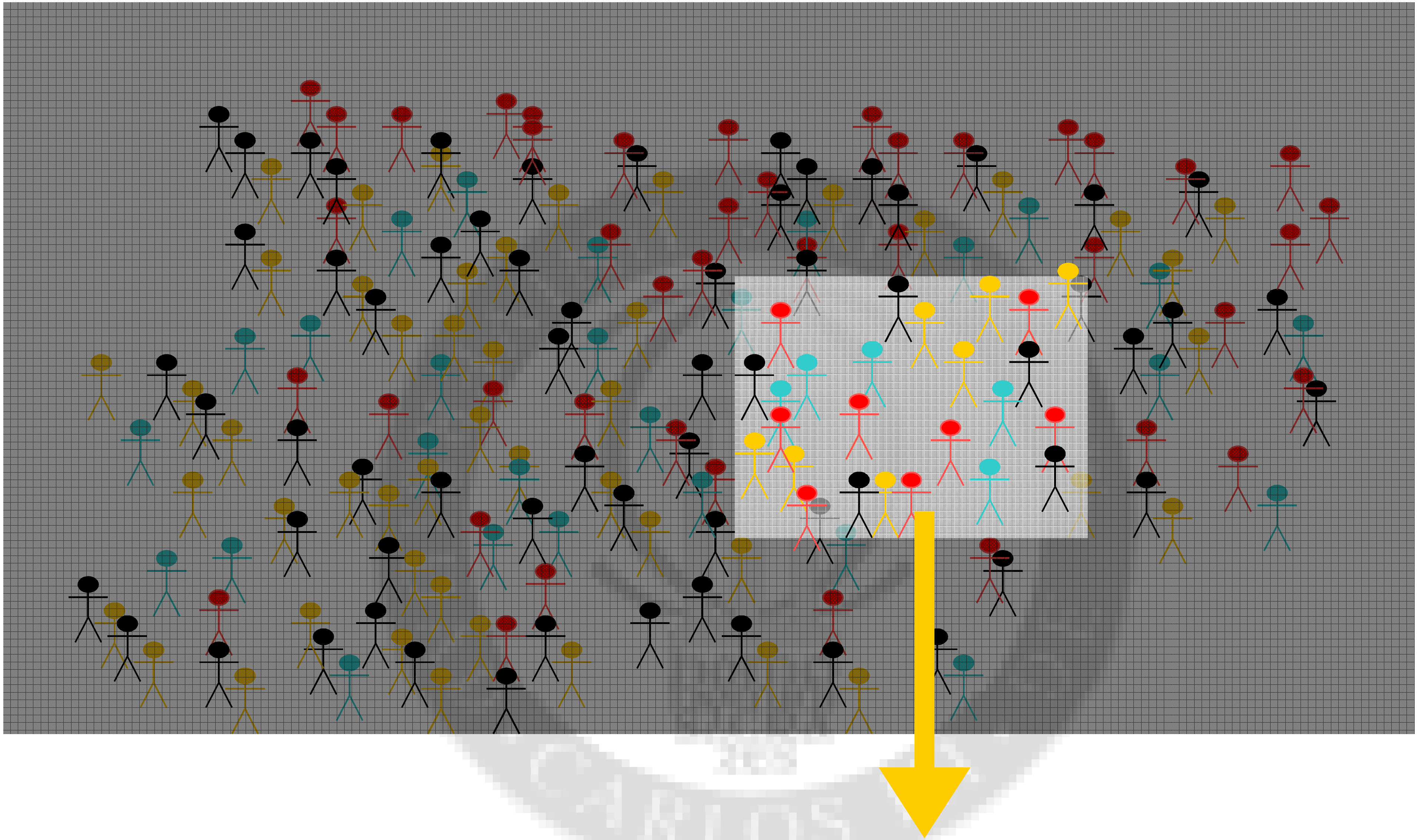
$X_1, X_2, \ldots, X_n$



Simple random sample:

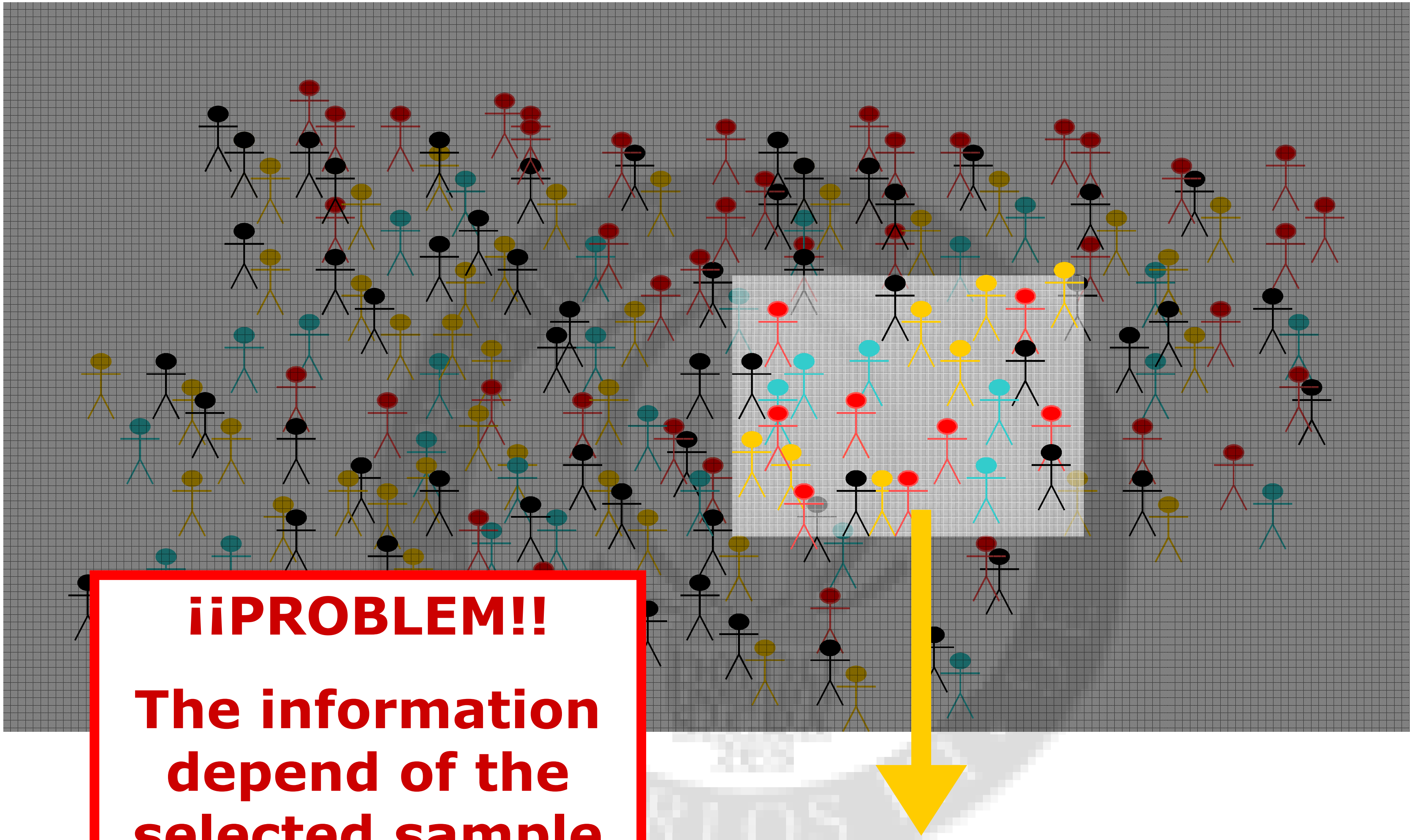
- all X_i has the same characteristics that X
- They are all independent of each other

$X_1, X_2, \ldots, X_n$ are a sequence of independent and identically-distributed random variables



We extract information from the sample by:

- Histogram
- Sample mean
- Sample variance...

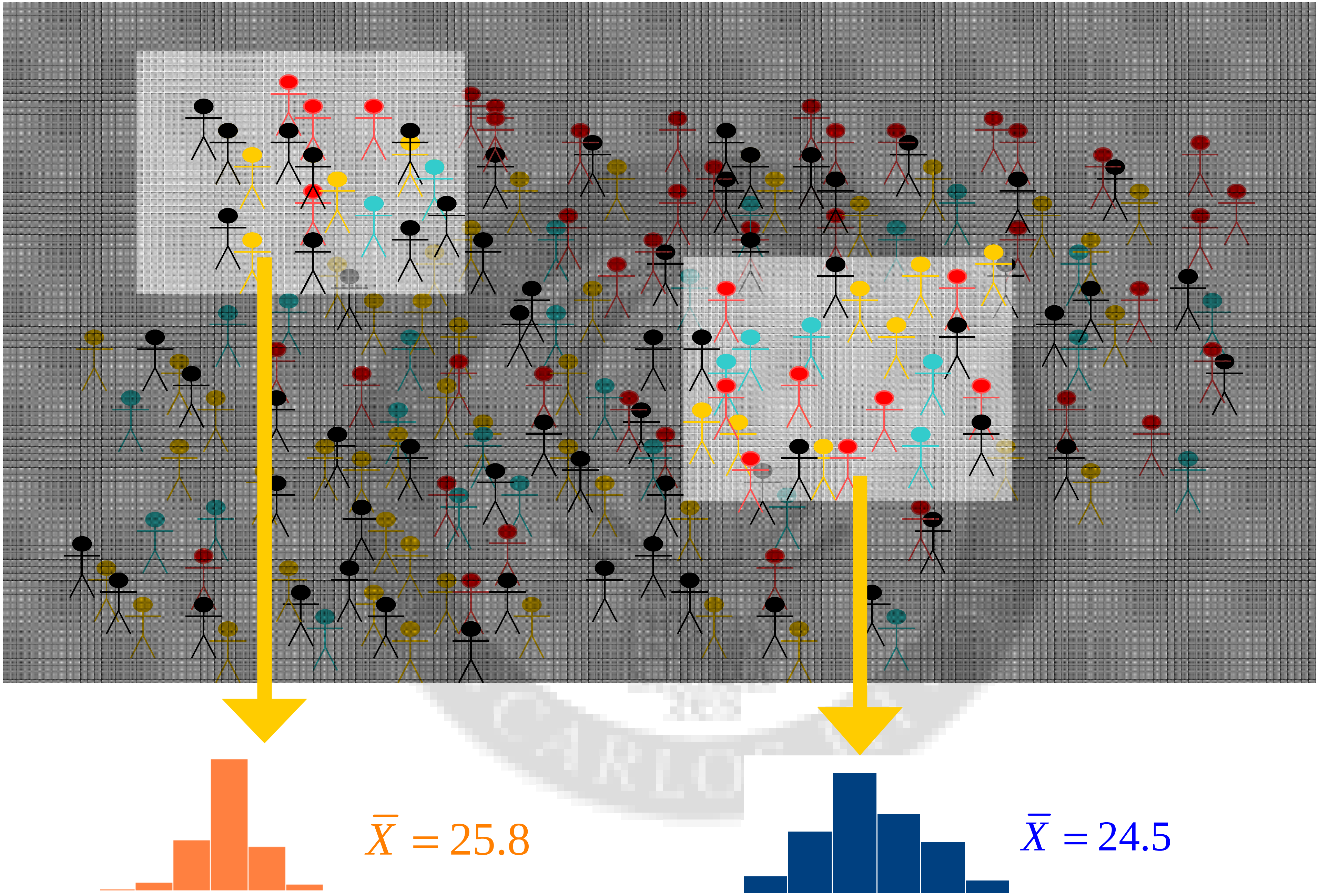


¡¡PROBLEM!!

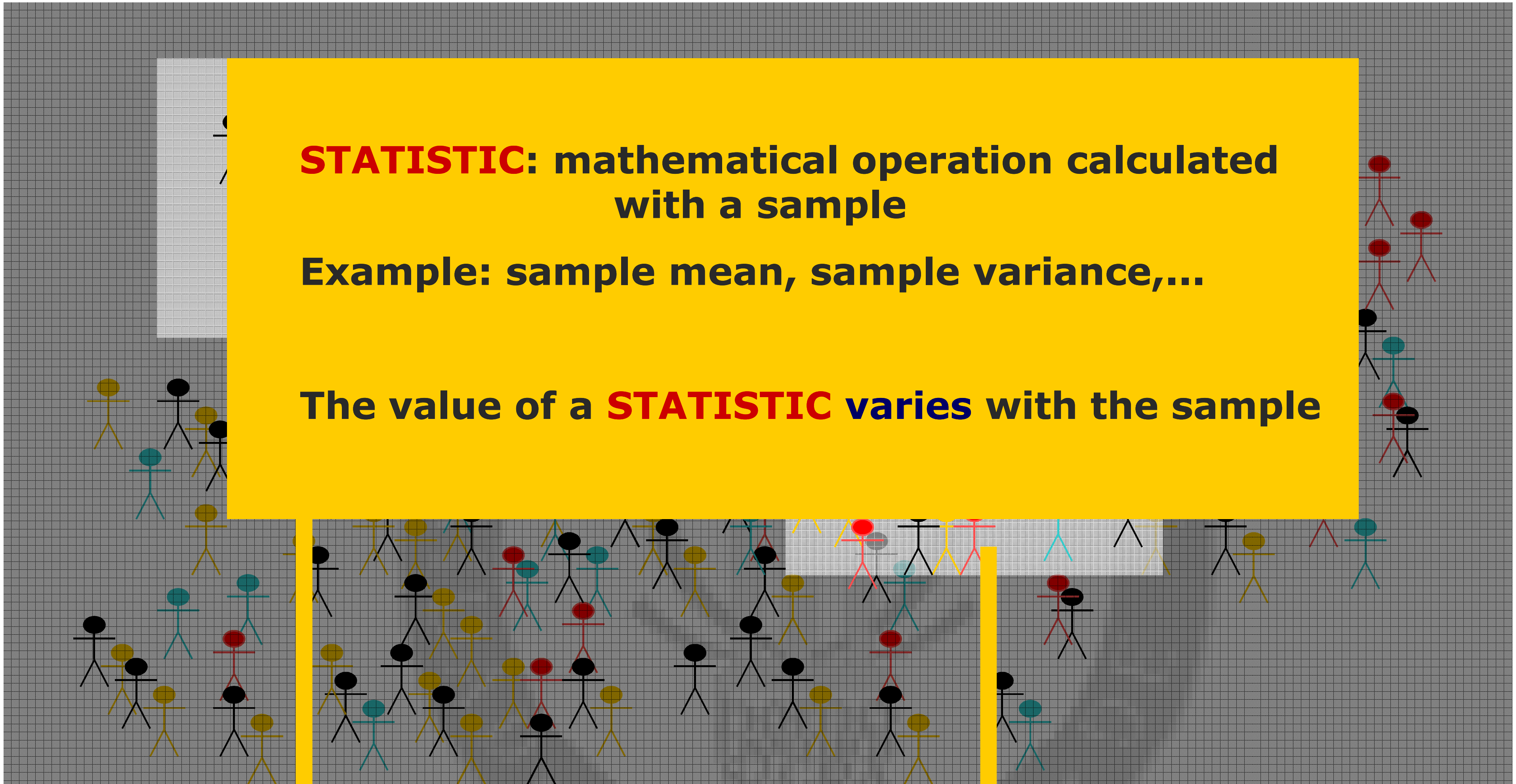
The information depend of the selected sample

Extract information of the sample:

- Histogram
- Sample mean
- Sample variance...



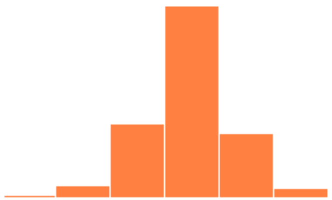
What credibility has the information of only one sample?



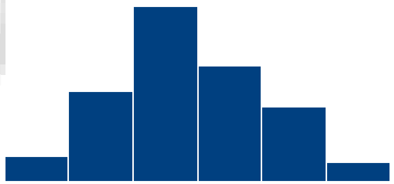
STATISTIC: mathematical operation calculated with a sample

Example: sample mean, sample variance,...

The value of a STATISTIC varies with the sample



$\bar{X} = 25.8$



$\bar{X} = 24.5$

What credibility has the information of only one sample?

Chapter 5: Introduction to statistical inference

1. Statistical inference. Population and sample
2. **Sampling distribution of a statistic**
3. The sample mean distribution
4. Estimation and estimators
5. Method of moments
6. Diagnosis of the model
7. Transformations that improve normality

2. Sampling distribution of a statistic

Statistic: Any mathematical operation calculated with a sample

Example: Sample mean

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

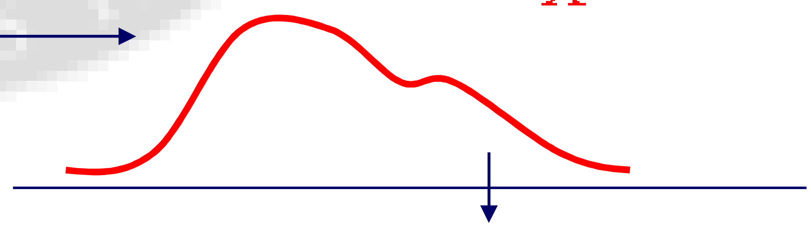
The elements of the sample are random variables all identically distributed like X. Their values depend on the sample

A statistic is always a random variable. Its value changes by changing sample

The statistic distribution is called sampling distribution or Distribution of the sample

It depends on the function and on the properties of the random variable X

$\bar{X}$

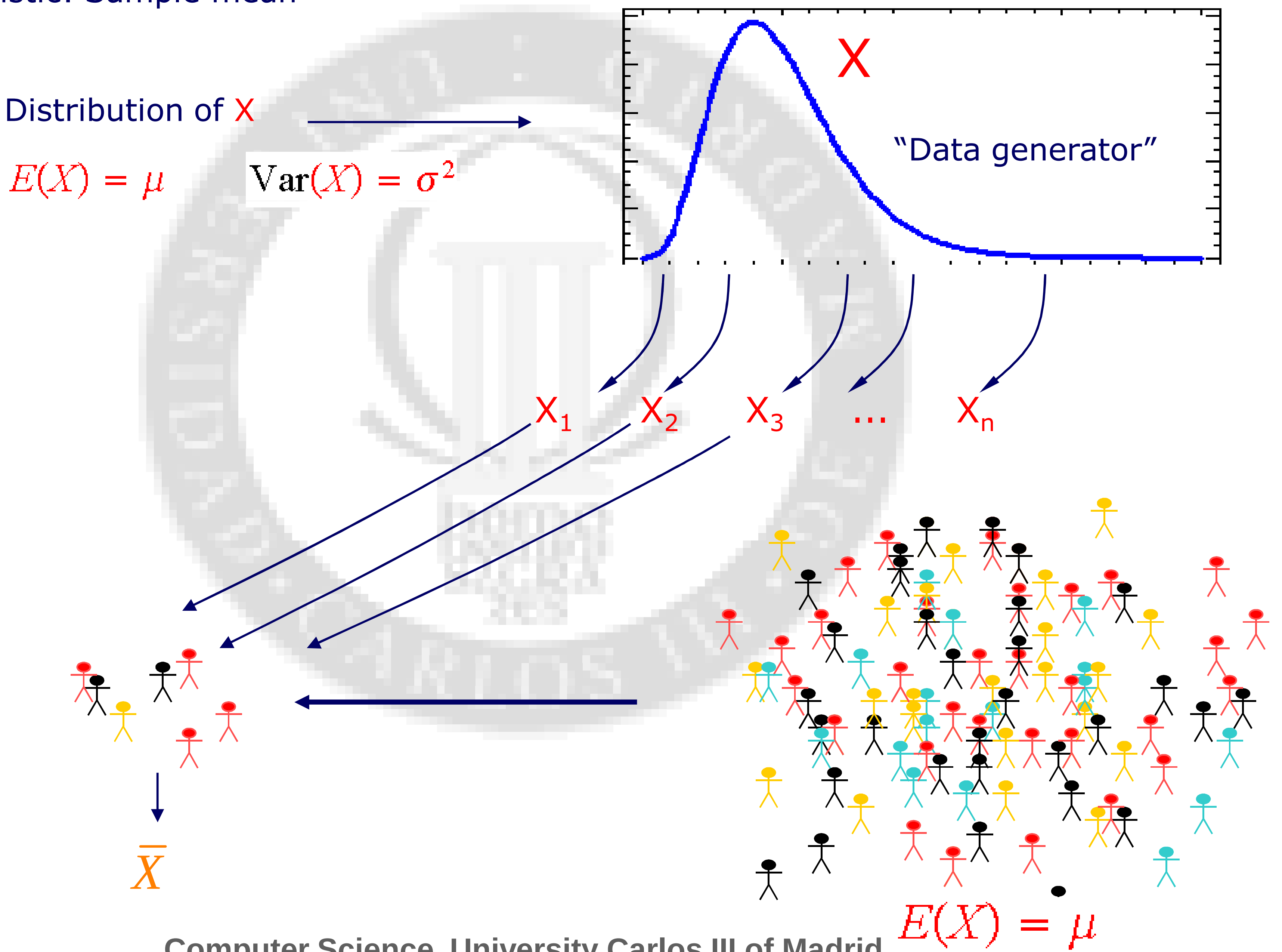


With a sample we have only one realization of $\bar{X}$

Chapter 5: Introduction to statistical inference

3. Sampling distribution of the mean

Statistic: Sample mean

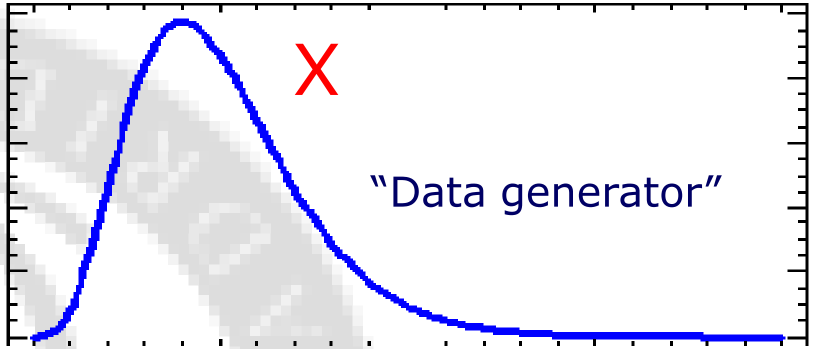


3. Sampling distribution of the mean

Statistic: Sample mean

Distribution of X

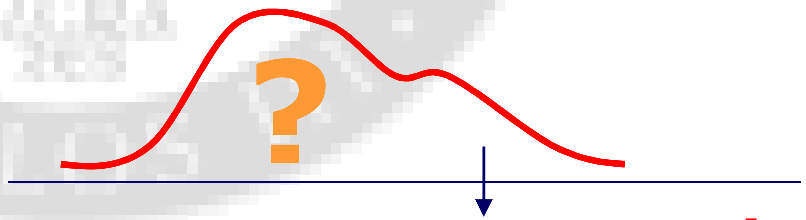
$E(X) = \mu$ $\mathrm{Var}(X) = \sigma^2$



X_1 X_2 X_3 … X_n

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

$\bar{X}$



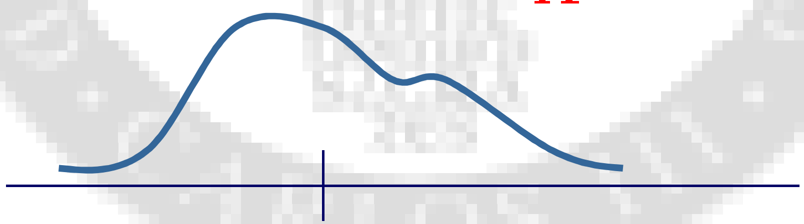
With a sample we have only one realization of $\bar{X}$

- ¿What is the shape of the sample mean distribution? (the distribution that is obtained by indefinitely changing the elements of the sample)
- ¿Is the sample mean a good approximation to the population mean μ?

3. Sampling distribution of the mean

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

$$E(\bar{X}) = E\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{E(X_1) + E(X_2) + \cdots + E(X_n)}{n} = \frac{n\mu}{n} = \mu$$



$$\bar{X}$$

$$E(\bar{X}) = \mu$$

The population mean is the centre of the different sample means that we could obtain with different samples

3. Sampling distribution of the mean

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{\mathrm{Var}\overbrace{(X_1 + X_2 + \cdots + X_n)}^{\text{independent}}}{n^2}$$

$$= \frac{\mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n)}{n^2} = \frac{n\sigma^2}{n} = \frac{\sigma^2}{n}$$

Decrease with n

- If n is sufficiently large, the sample mean changes very little from a sample to another
- It is very unlikely that the sample mean assumes a value very far from μ

3. Sampling distribution of the mean

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} \longrightarrow \bar{X} = \frac{X_1}{n} + \frac{X_2}{n} + \cdots + \frac{X_n}{n}$$

We are adding random variables.

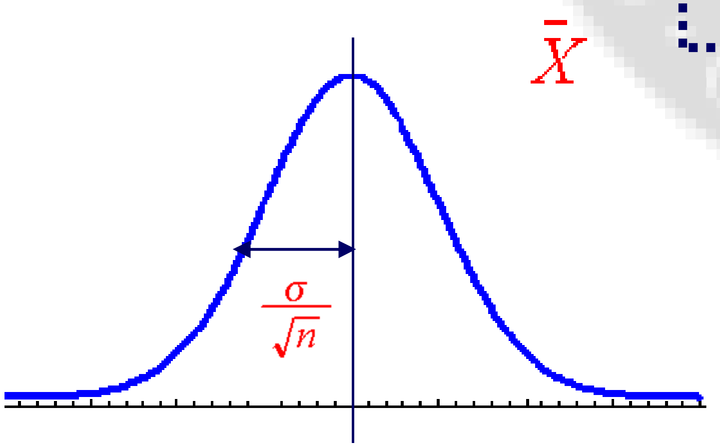
From the **Central Limit Theorem**, if n is large ($n>30$)

$\downarrow$

NORMAL

Independently of how is X!!!!!!

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$



$E(\bar{X}) = \mu$

- Symmetric
- Concentrated around of μ
- with n large it is very likely that with any sample we obtain a value close to μ

Chapter 5: Introduction to statistical inference

4. Estimation and estimators

Parameters → Numerical values summarizing characteristics of the population:

μ , σ^2 , λ , quartiles,...

If the parameter is unknown → We have to assign to it an approximated value extracting it from a sample of the data

ESTIMATION OF THE PARAMETER

ESTIMATION: It is a numeric value computed from the sample with the aim to assign the most accurate value to an unknown parameter

ESTIMATOR: It is a statistic used to estimate an unknown parameter

(being a statistic, it is a random variable)

Example: the sample mean can be used like an estimator of the population mean

4. Estimation and estimators

NOTATION: The symbol to denote an estimator will be the same of the parameter but with circumflex accent ^

Estimator for the population mean μ $\longrightarrow$ $\hat{\mu}$

Estimator for the variance σ^2 $\longrightarrow$ $\hat{\sigma}^2$

Estimator for a parameter θ $\longrightarrow$ $\hat{\theta}$

It is possible to set several estimators for the same parameter.
We want to be able to choose the best

Example:

For a simple random sample of size $n=3$ of a normal random variable of mean μ and known variance $\sigma^2=1$, we consider the following estimators for μ:

$$\hat{\mu}_1 = (1/3)X_1 + (1/3)X_2 + (1/3)X_3$$

$$\hat{\mu}_2 = (1/4)X_1 + (1/2)X_2 + (1/4)X_3$$

$$\hat{\mu}_3 = (1/8)X_1 + (3/8)X_2 + (1/2)X_3$$

How to choose the most appropriate?

How to choose the best estimator?

What properties should to have an estimator of a parameter θ?

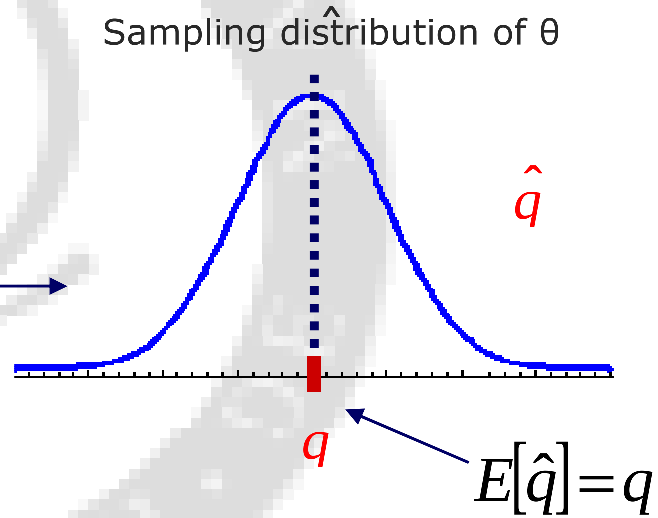
Although it is a random variable whose value depends on the selected sample, we want that it assumes values close to the true parameter with high probability

How to evaluate it?

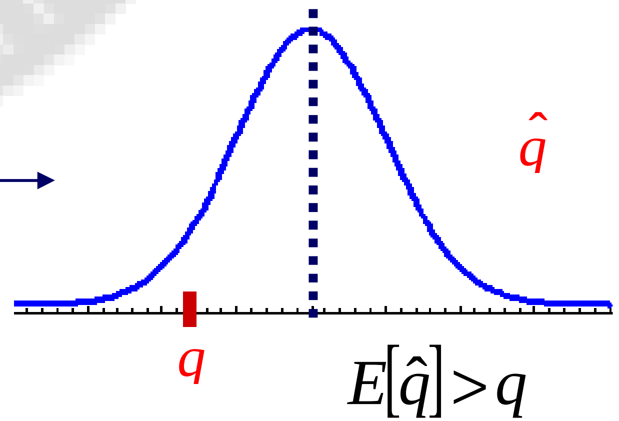
Different criterions:

1 That E($\hat{\theta}$) is close to θ

GOOD estimator of the parameter θ



BAD estimator of the parameter θ. Tendency to overestimate the value



What properties should to have an estimator of a parameter θ?

Although it is a random variable whose value depends on the selected sample, we want that it assumes values close to the true parameter with high probability

How to evaluate it?

Different criterions:

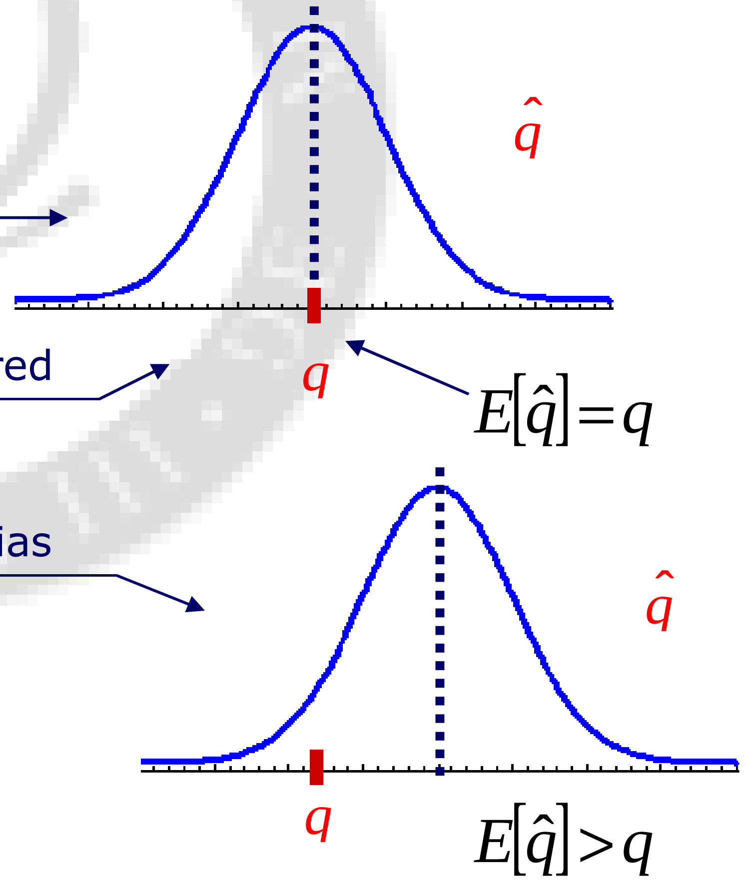
1 That $E(\hat{\theta})$ is close to θ

$$\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

Sampling distribution of $\hat{\theta}$

unbiased o centered

biased. Positive bias



Less is the bias
better will be the estimator

Example:

For a simple random sample of size $n=3$ of a normal random variable of mean μ and known variance $\sigma^2=1$, is consider the following estimators for μ:

$$\hat{\mu}_1 = (1/3)X_1+(1/3)X_2+(1/3)X_3$$

$$\hat{\mu}_2 = (1/4)X_1+(1/2)X_2+(1/4)X_3$$

$$\hat{\mu}_3 = (1/8)X_1+(3/8)X_2+(1/2)X_3$$

Let's see their biases:

$$E(\hat{\mu}_1) = \frac{1}{3}E(X_1) + \frac{1}{3}E(X_2) + \frac{1}{3}E(X_3) = \frac{1}{3}3\mu = \mu$$

$$E(\hat{\mu}_2) = \frac{1}{4}E(X_1) + \frac{1}{2}E(X_2) + \frac{1}{4}E(X_3) = \frac{1}{4}\mu + \frac{1}{2}\mu + \frac{1}{4}\mu = \mu$$

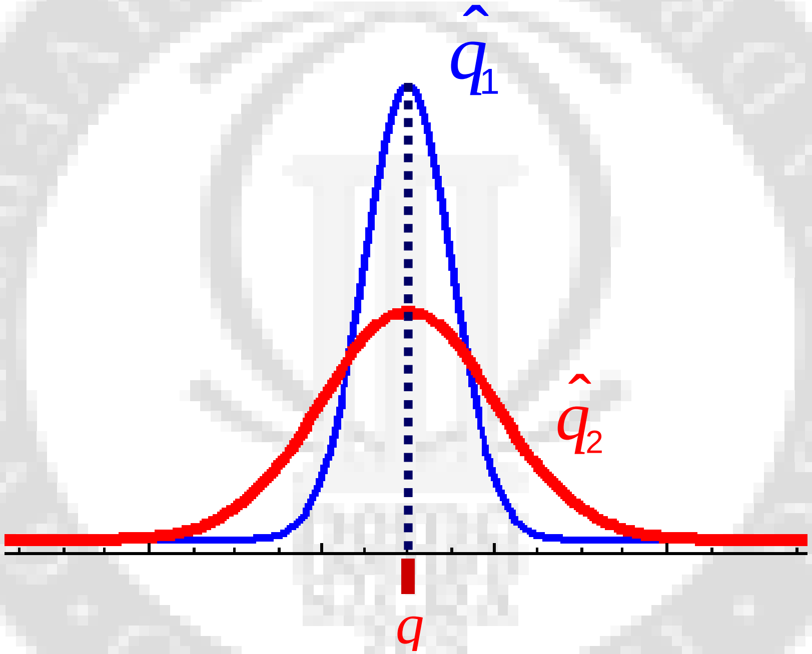
$$E(\hat{\mu}_3) = \frac{1}{8}E(X_1) + \frac{3}{8}E(X_2) + \frac{1}{2}E(X_3) = \frac{1}{8}\mu + \frac{3}{8}\mu + \frac{1}{2}\mu = \mu$$

The three estimator are all unbiased

How to choose the best estimator?

1 That $E(\hat{\theta})$ is close to θ

2 That $\hat{\theta}$ has lower variance



Although both estimators are unbiased, the estimator $\hat{\boldsymbol{\theta}}_2$ is worse than $\hat{\boldsymbol{\theta}}_1$, as it has lower variance, i.e. it is less accurate.

Example:

For a simple random sample of size $n=3$ of a normal random variable of mean μ and known variance $\sigma^2=1$, is consider the following estimators for μ:

$$\hat{\mu}_1 = (1/3)X_1+(1/3)X_2+(1/3)X_3$$

$$\hat{\mu}_2 = (1/4)X_1+(1/2)X_2+(1/4)X_3$$

$$\hat{\mu}_3 = (1/8)X_1+(3/8)X_2+(1/2)X_3$$

Let's compute their variances:

$$Var(\hat{\mu}_1) = Var\left(\frac{1}{3}X_1+\frac{1}{3}X_2+\frac{1}{3}X_3\right) = \frac{1}{9}Var(X_1)+\frac{1}{9}Var(X_2)+\frac{1}{9}Var(X_3) = \frac{1}{3}\sigma^2 = \frac{1}{3} = 0.333$$

$$Var(\hat{\mu}_2) = Var\left(\frac{1}{4}X_1+\frac{1}{2}X_2+\frac{1}{4}X_3\right) = \frac{1}{16}Var(X_1)+\frac{1}{4}Var(X_2)+\frac{1}{16}Var(X_3) = \frac{3}{8}\sigma^2 = \frac{3}{8} = 0.375$$

$$Var(\hat{\mu}_3) = Var\left(\frac{1}{8}X_1+\frac{3}{8}X_2+\frac{1}{2}X_3\right) = \frac{1}{64}Var(X_1)+\frac{9}{64}Var(X_2)+\frac{1}{4}Var(X_3) = \frac{26}{64}\sigma^2 = \frac{13}{32} = 0.406$$

The first estimator has the lowest variance, therefore it is the best estimator for μ

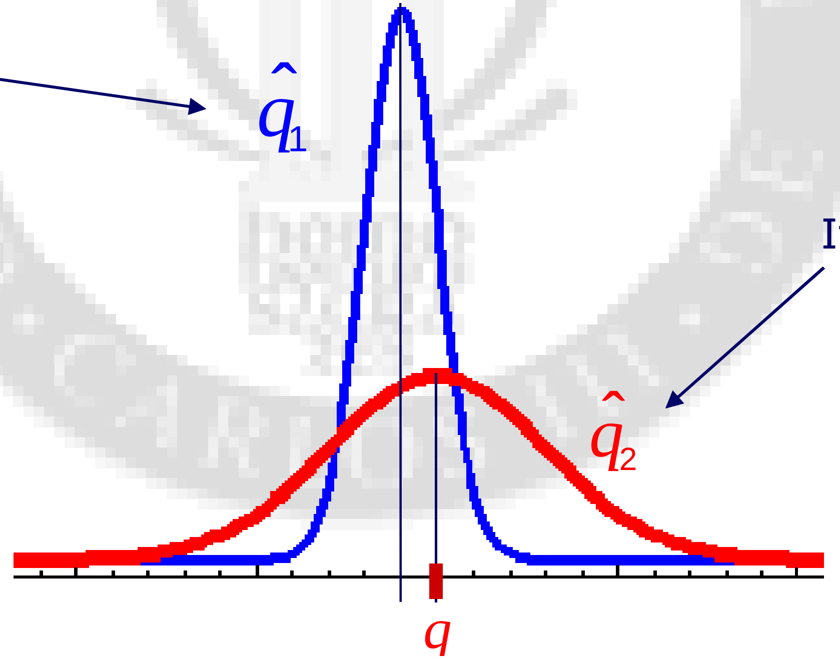
How to choose the best estimator?

1. That $E(\hat{\theta})$ is close to θ
2. That $\hat{\theta}$ has lower variance
3. If there are several estimators, with different biases and variances we choose the estimator with lower MEAN SQUARED ERROR (MSE)

$$\mathrm{MSE}\left(\hat{\theta}\right) = \mathrm{E}\left[\left(\hat{\theta}-\theta\right)^2\right] = \mathrm{Var}\left[\hat{\theta}\right] + \mathrm{Bias}\left[\hat{\theta}\right]^2$$

It is biased, but with less variance

It is unbiased but with larger variance



We have to calculate the MSE for each one and choose the estimator with less MSE

EFFICIENT ESTIMATOR

It assumes values close to the true value of the population parameter with the highest probability

Chapter 5: Introduction to statistical inference

1. Statistical inference. Population and sample
2. Sampling distribution of a statistic
3. The sample mean distribution
4. Estimation and estimators
5. **Method of moments**
6. Diagnosis of the model
7. Transformations that improve normality

5. Methods of the moments

- It is a method to find estimators which have good properties
- It is a method straightforward to compute, but that not always is the best

IDEA: *To estimate a given characteristic of the population we use the respective characteristic of the sample*

To estimate the population mean μ → Sample mean
To estimate the population variance σ^2 → Sample variance
To estimate the population proportion p → Sample proportion

Example:

The number of customers who arrive to a service station in an hour is a Poisson random variable. During 5 hours we count the customers who arrive in each hour and we get the followings numbers: 5, 0, 3, 3, 4. Estimate the parameter λ of the Poisson.

Since $E(X)=\lambda$ we have

$$\hat{\lambda} = \bar{X} = \frac{\sum_{i=1}^{5} X_i}{5} = \frac{5+0+0+3+4}{5} = 3$$

6. Diagnosis of the model

Assuming that the histogram shows that the observations could be distributed as certain probability model, how could we compare the observations with the model predictions?

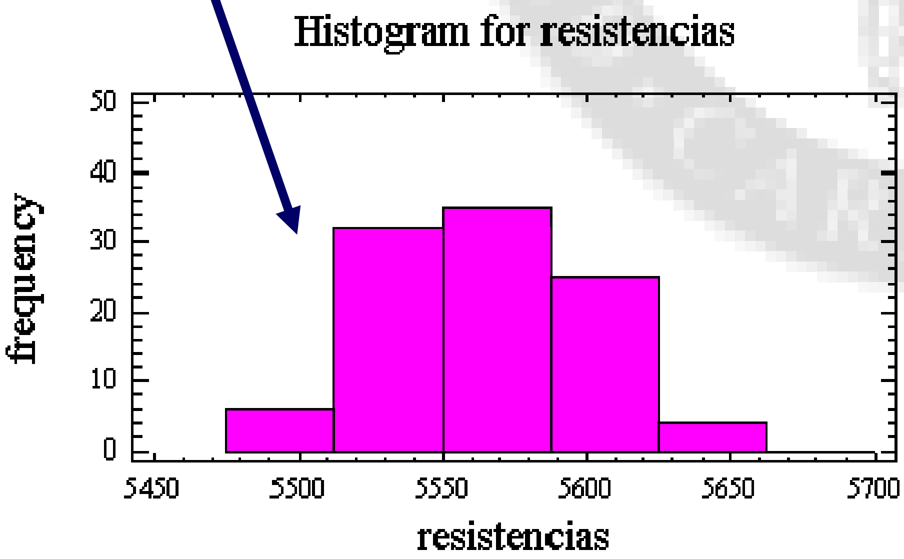
Chi-square test

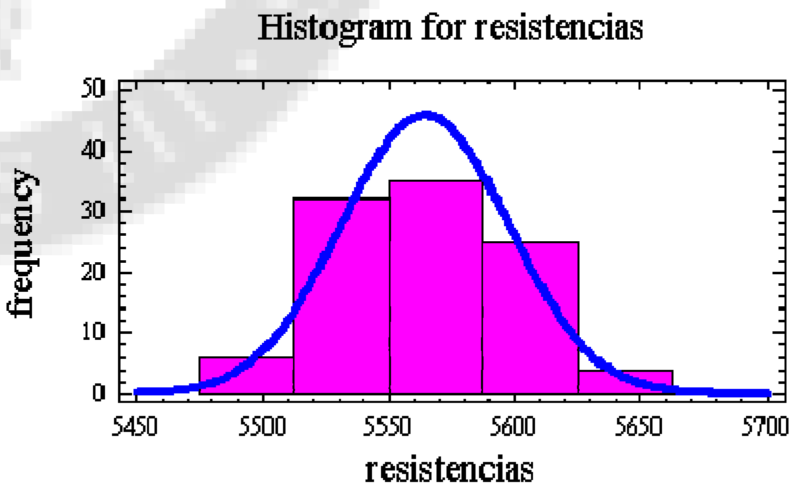
- It is a method for assess the goodness of fit of a model.
- It is a method for continuous and discrete models.

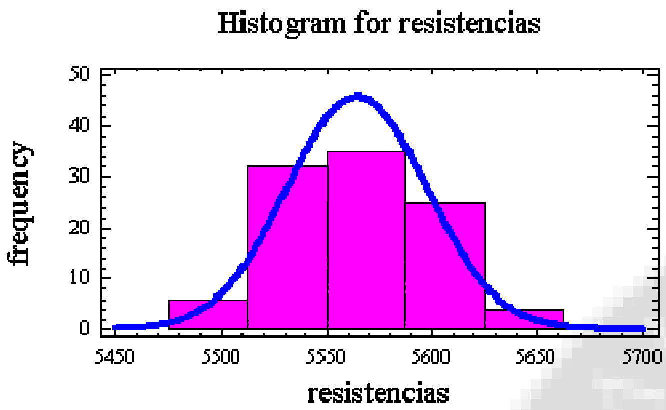
These data suggest a Normal population.
With the observations estimate μ and σ^2

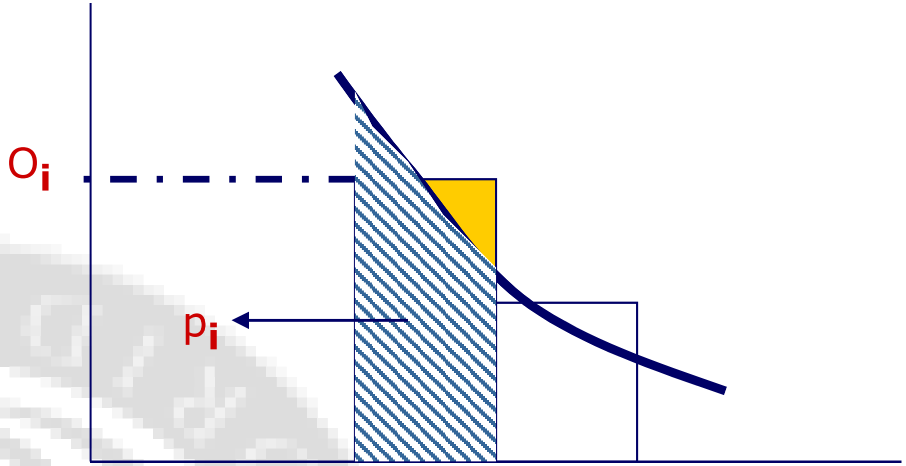
$$¿N\left(\hat{\mu},\hat{\sigma}^2\right)?$$

We compare the histogram with the curve of the Normal estimated









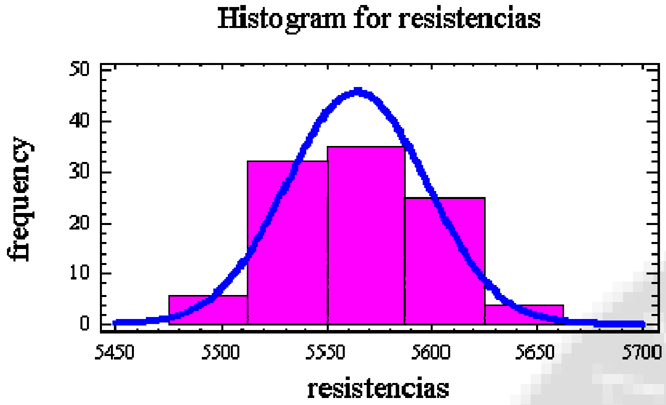
For each bin of the histogram:

- We count the number of observed observations in each bin: O_i
- We calculate, depending on the model, the probability that the observation is in this bin: p_i
- We calculate, the expected number of observations: $E_i = np_i$

$$X_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

If X_0^2 **is** very large: there is a lot of discrepancy among data and the model.

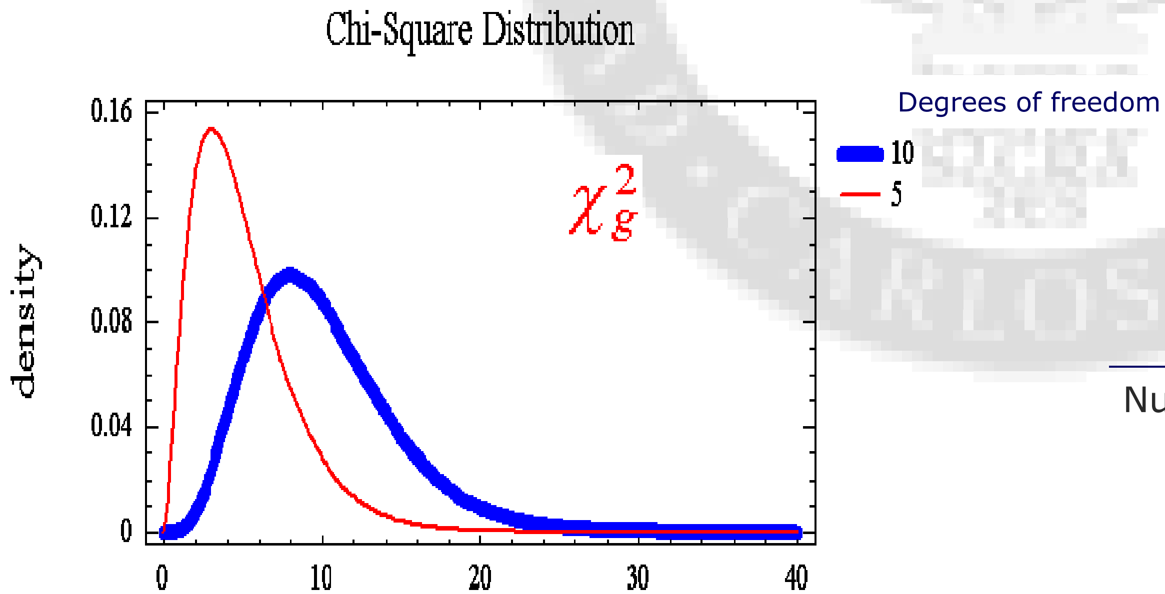
Reject the model



$$X_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

It is proved that:

If the model is not adequate, X_0^2 is a large value that will tend to be at right of a Chi-square distribution

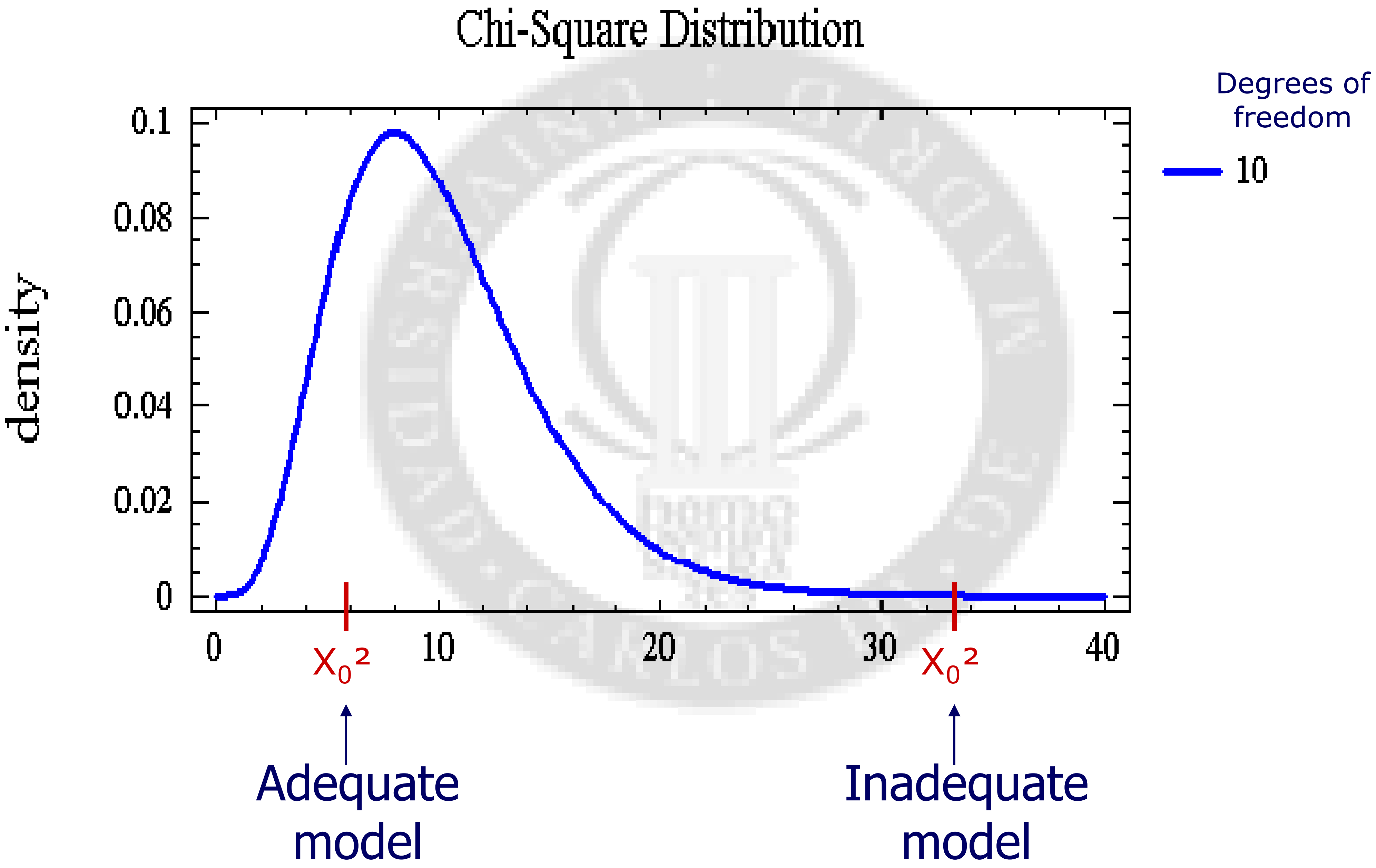


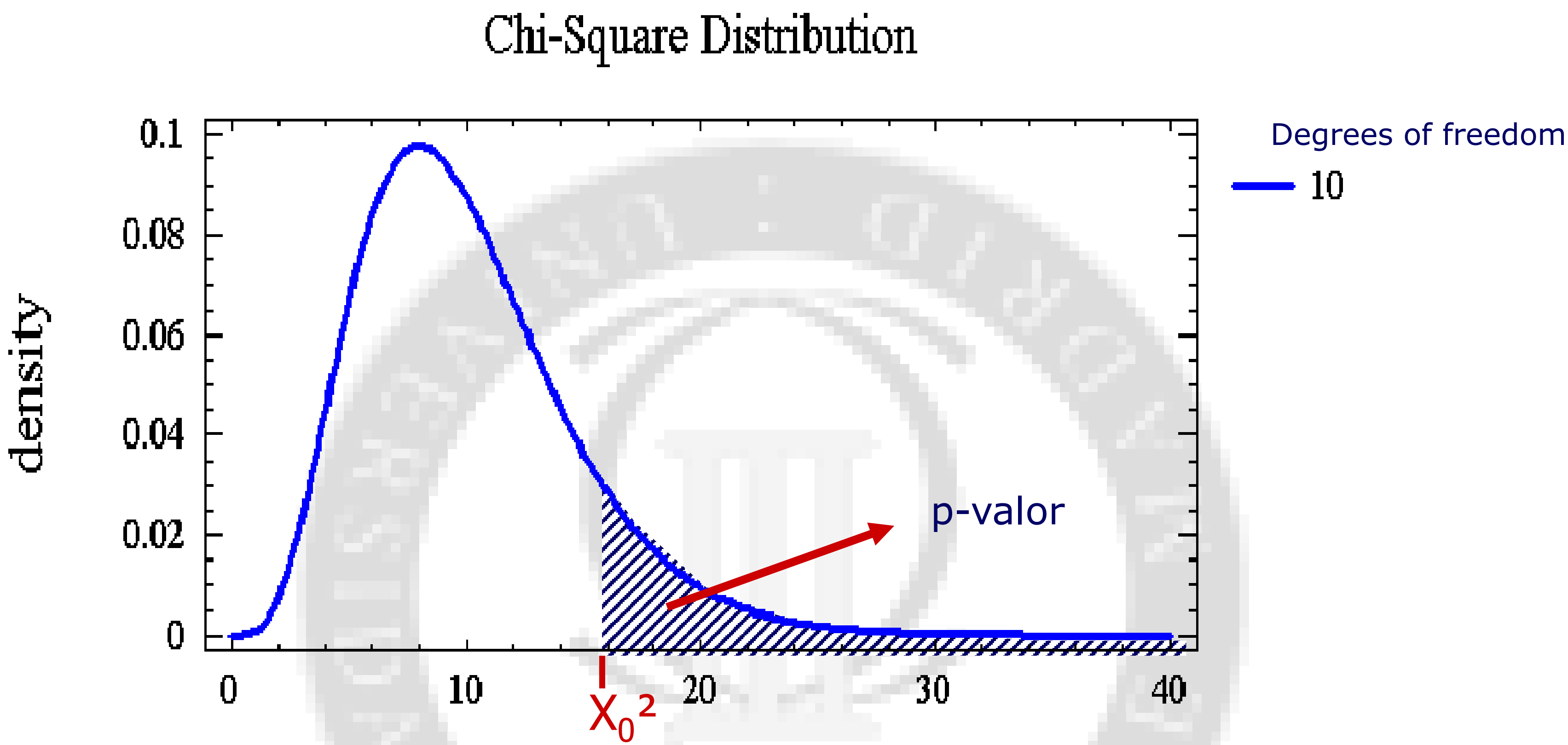
g= degrees of freedom

g=k-v-1

Number of bins

Number of estimated parameters


Chi-Square Distribution
density
0.1
0.08
0.06
0.04
0.02
0
0
10
20
30
40
X_0^2
X_0^2
Degrees of freedom
10
Adequate model
Inadequate model



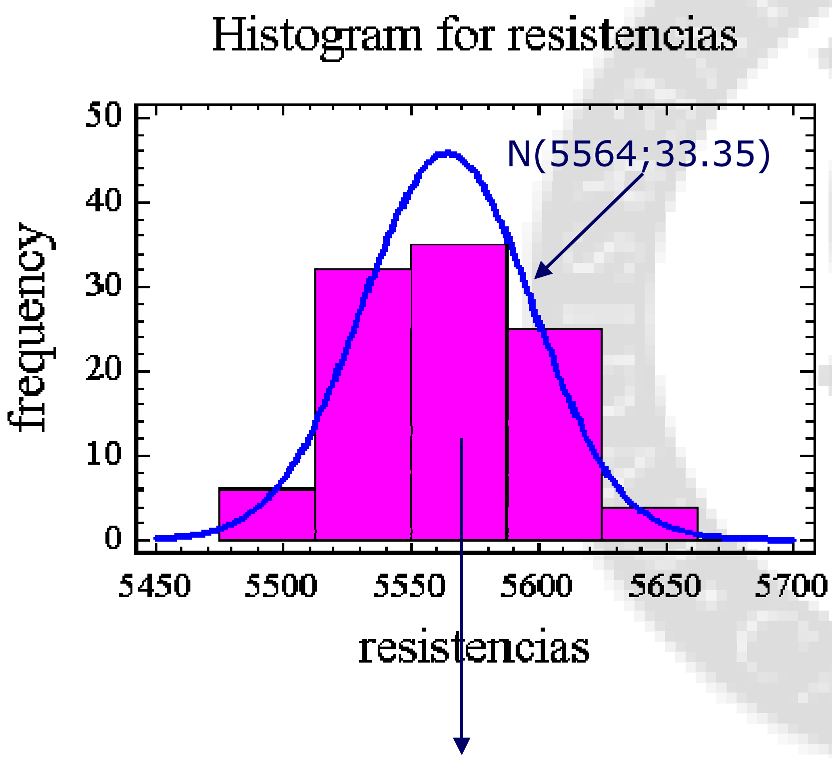
- The area on the right of X_0^2 is called p-value
- Its value is generally calculated by using a computer

In practice, if p-value<0.05 we reject the model

Example: Resistivity of 102 similar resistors

They are modeled by a Normal?

With the 102 observations: sample mean=5564. Sample standard deviation=33.35



Goodness-of-Fit Tests for resistencias

Chi-Square Test

Lower Limit	Upper Limit	Observed Frequency (O_i)	Expected Frequency (E_i)	Chi-Square $\frac{(O_i - E_i)^2}{E_i}$
at or below	5516.79	7	7.85	0.09
5516.79	5530.33	10	7.85	0.59
5530.33	5539.79	8	7.85	0.00
5539.79	5547.6	9	7.85	0.17
5547.6	5554.57	8	7.85	0.00
5554.57	5561.13	6	7.85	0.43
5561.13	5567.57	10	7.85	0.59
5567.57	5574.14	7	7.85	0.09
5574.14	5581.11	3	7.85	2.99
5581.11	5588.91	5	7.85	1.03
5588.91	5598.38	13	7.85	3.39
5598.38	5611.92	6	7.85	0.43
above 5611.92		10	7.85	0.59

Chi-Square = 10.4114 with 10 d.f. P-Value = 0.40517

X_0^2 p-value

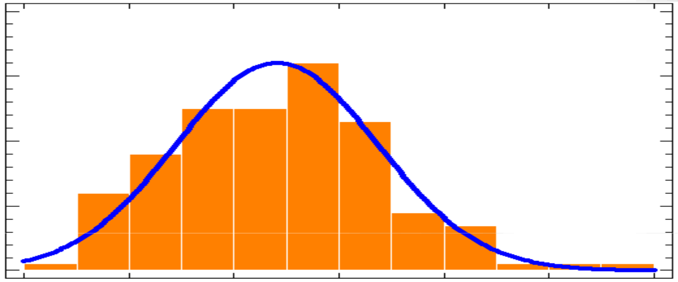
If p-value is sufficiently large, we can use the Normal distribution to represent the real population → Resist.~N(5564;33.35)

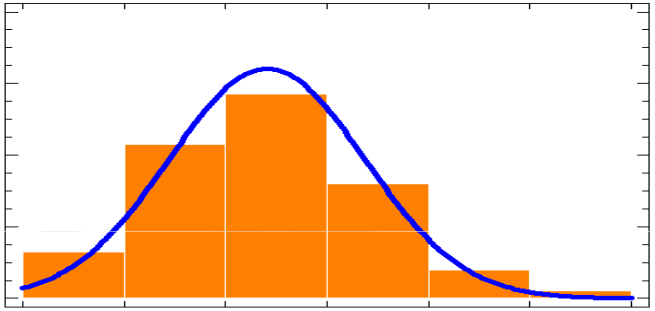
Chapter 5: Introduction to statistical inference

1. Statistical inference. Population and sample
2. Sampling distribution of a statistic
3. The sample mean distribution
4. Estimation and estimators
5. Method of moments
6. Diagnosis of the model
7. **Transformations that improve normality**

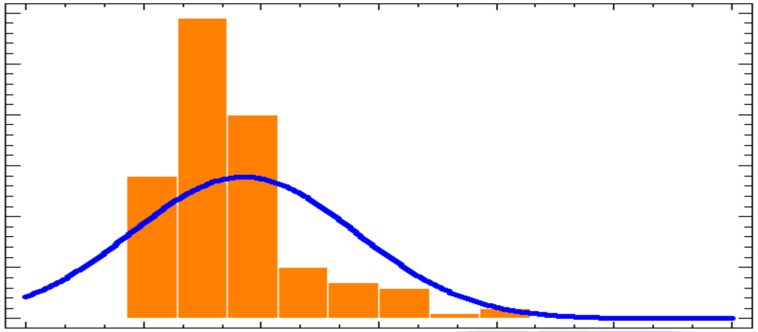
7. Transformations that improve normality

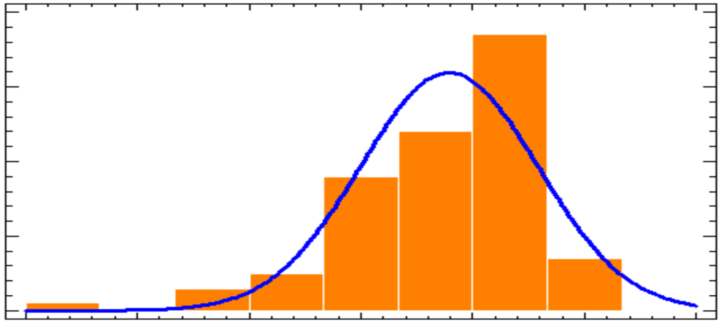
Most symmetric unimodal data can be well approximated by a Normal distribution





However it is common to find unimodal asymmetric data





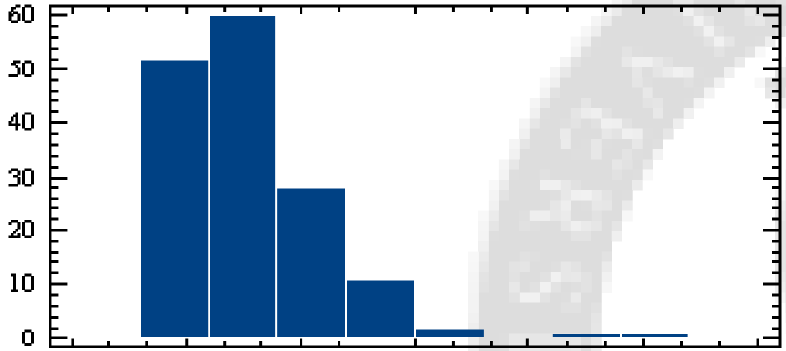
We transform the data in order to make them unimodal symmetric
Then we try to fit a normal model to the transformed data

7. Transformations that improve symmetry

Given a sample of data $x_1, x_2, x_3, \ldots, x_n$ with a distribution UNIMODAL asymmetric

We look for a transformation $y=h(x)$ such that y is the most symmetric as possible

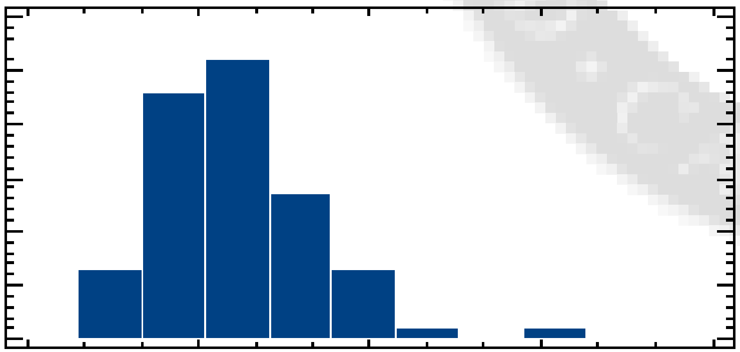
x



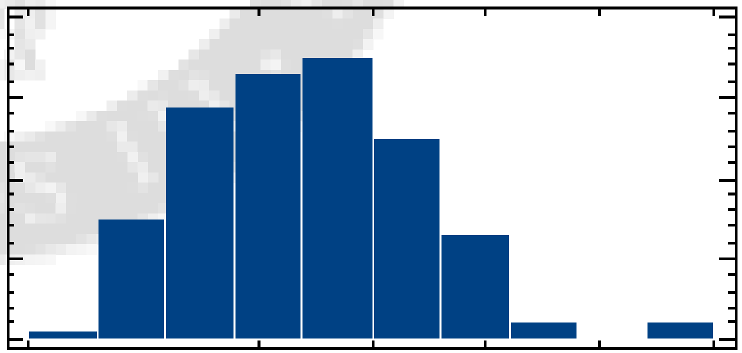
Positive asymmetries (very common)

- Use transformations of the kind $y=x^c$, $c<1$
- $y=\ln(x)$
- This transformations are no lineal with the property to compress a lot the large values and less the small values

$\sqrt{x}$



$\ln(x)$



- $\ln(x)$ can be interpreted as the limit of a transformation of the kind $y=x^c$ as $c \to 0$
- The greater is the asymmetry the smaller will be the value of the exponent c

7. Transformations that improve symmetry

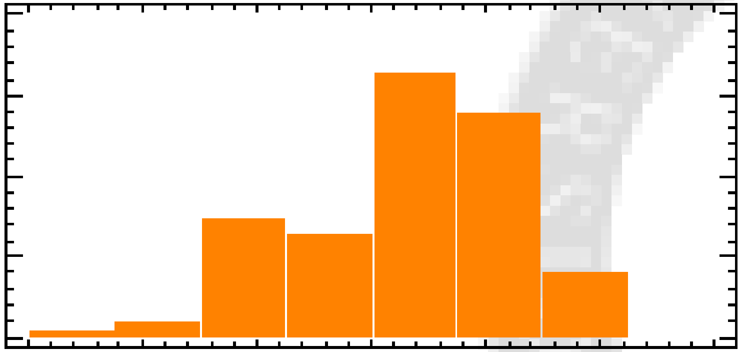
Given a sample of data x_1, x_2, x_3, ..., x_n with a distribution UNIMODAL asymmetric

We look for a transformation $y=h(x)$ such that y is the most symmetric as possible

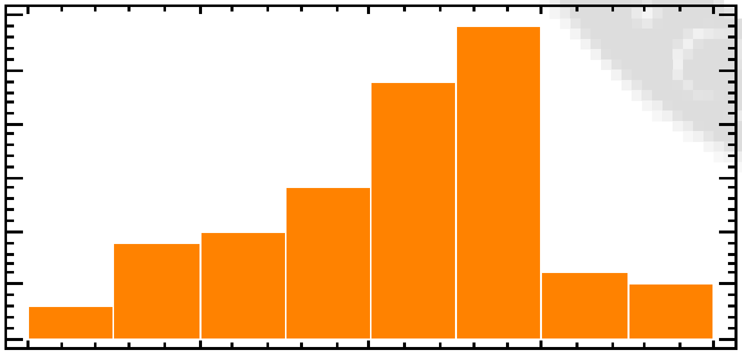
Negative Asymmetries

- Use transformations of the kind $y=x^c$, $c>1$
- This transformations are no lineal with the property to expand a lot the large values and less the small values

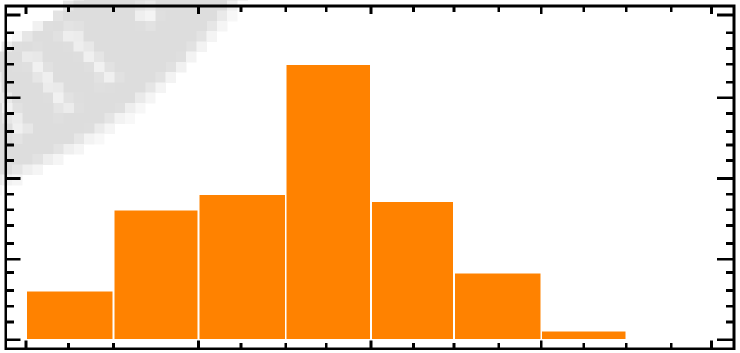
x



$x^{1,5}$



x^2

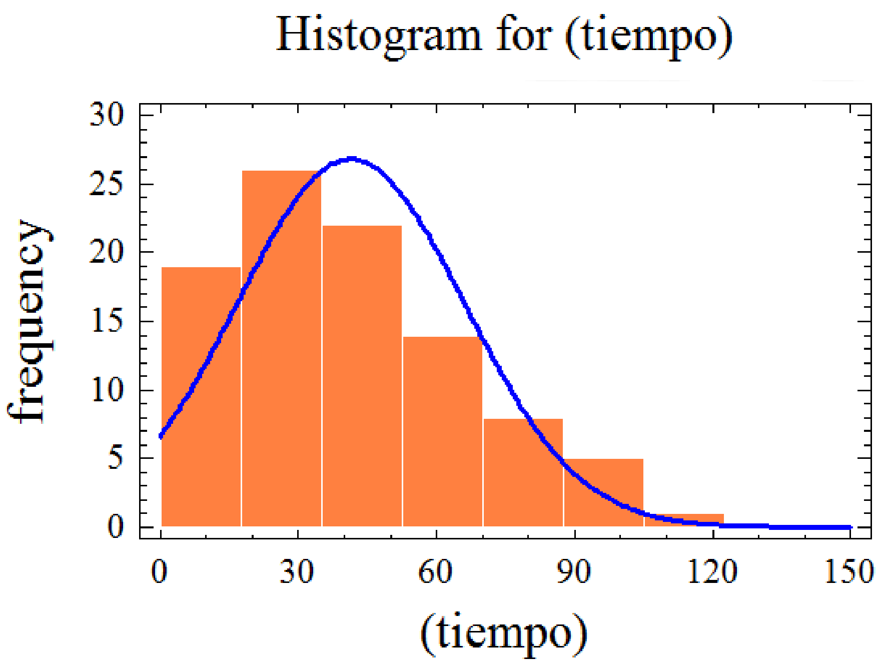


The greater is the asymmetry the larger will be the value of the exponent c

Example

The file AlumnosIndustriales.sf3 contains a sample of Industrial Engineering students. The variable **tiempo** records the time (in minutes) that these students take to get at the University. What is the probability that a student will take more than 50 minutes to arrive to the University?

(Note: we are not asking for the proportion of students who takes more than 60 minute, what we are looking for is the proportion as it were computed over the whole population of students)



Goodness-of-Fit Tests for (tiempo)

Chi-Square Test

Lower Limit	Upper Limit	Observed Frequency	Expected Frequency	Chi-Square
at or below	7,20844	4	7,92	1,94
7,20844	17,4889	15	7,92	6,34
17,4889	24,7355	8	7,92	0,00
24,7355	30,7657	13	7,92	3,26
30,7657	36,2154	5	7,92	1,07
36,2154	41,4211	6	7,92	0,46
41,4211	46,6267	11	7,92	1,20
46,6267	52,0764	5	7,92	1,07
52,0764	58,1066	2	7,92	4,42
58,1066	65,3532	11	7,92	1,20
65,3532	75,6337	8	7,92	0,00
above 75,6337		7	7,92	0,11

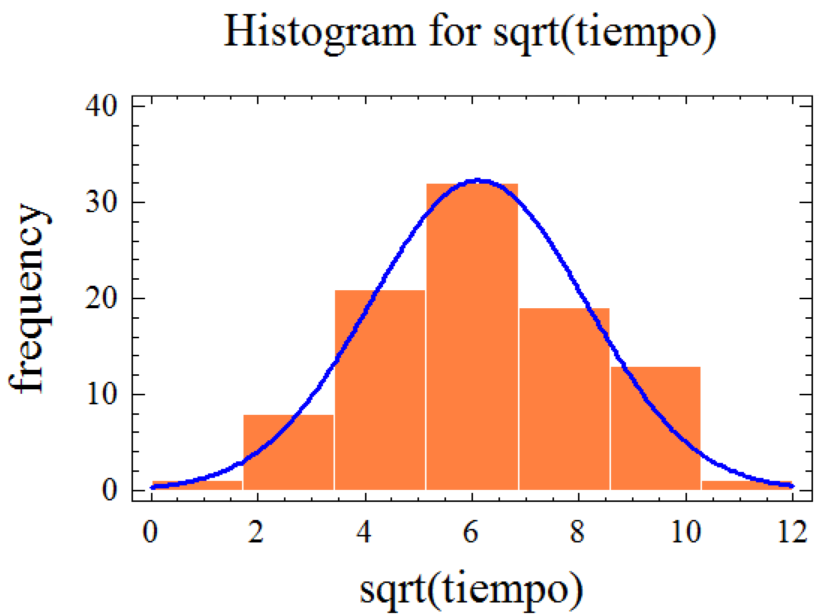
Chi-Square = 21,0842 with 9 d.f.   P-Value = 0,0122818

The positive asymmetry does not allow us to approximate the distribution to the one of a normal random variable. We can try to use a transformation of the kind x^c, with $c<1$

Example

The file AlumnosIndustriales.sf3 contains a sample of Industrial Engineering students. The variable **tiempo** records the time (in minutes) that these students take to get at the University. What is the probability that a student will take more than 50 minutes to arrive to the University?

(Note: we are not asking for the proportion of students who takes more than 60 minute, what we are looking for is the proportion as it were computed over the whole population of students)



Analysis Summary

Data variable: sqrt(tiempo)

95 values ranging from 1,0 to 10,9545

Fitted normal distribution:
mean = 6,11693
standard deviation = 2,01167

Goodness-of-Fit Tests for sqrt(tiempo)

Chi-Square Test

Lower Limit	Upper Limit	Observed Frequency	Expected Frequency	Chi-Square
at or below	3,3348	9	7,92	0,15
3,3348	4,1708	10	7,92	0,55
4,1708	4,76008	8	7,92	0,00
4,76008	5,25045	3	7,92	3,05
5,25045	5,69362	10	7,92	0,55
5,69362	6,11693	5	7,92	1,07
6,11693	6,54025	6	7,92	0,46
6,54025	6,98341	11	7,92	1,20
6,98341	7,47378	7	7,92	0,11
7,47378	8,06307	11	7,92	1,20
8,06307	8,89906	8	7,92	0,00
above 8,89906		7	7,92	0,11

Chi-Square = 8,45304 with 9 d.f. P-Value = 0,489212

The transformation $x^{0.5}$ seems to work.
It gives a p-value 0.49 that is acceptable.

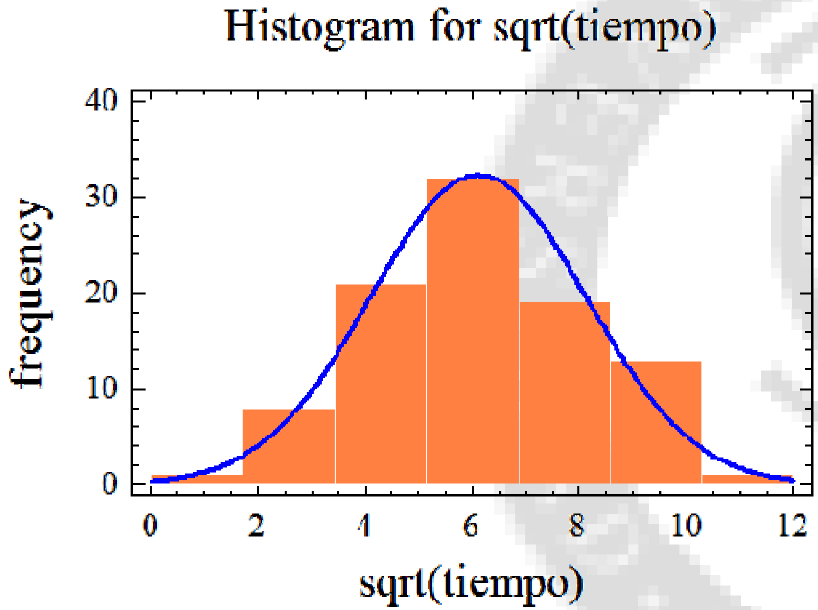
The model that fits to a Normal distribution is

$$X^{0.5} \sim N(6.12; 2.01^2)$$

Example

The file AlumnosIndustriales.sf3 contains a sample of Industrial Engineering students. The variable **tiempo** records the time (in minutes) that these students take to get at the University. What is the probability that a student will take more than 50 minutes to arrive to the University?

(Note: we are not asking for the proportion of students who takes more than 60 minute, what we are looking for is the proportion as it were computed over the whole population of students)



$$X^{0.5} \sim N(6.12; 2.01^2)$$

$$P(X > 60) = P(X^{0.5} > 60^{0.5})$$

$$= P(X^{0.5} > 7.746) = 0.21$$

We **estimate** that 21% of the students of Industrial Engiinering takes more than one hour to get at the University.