Introduction to Statistics

Tomàs Aluja

Barcelona; December 16th, 2016

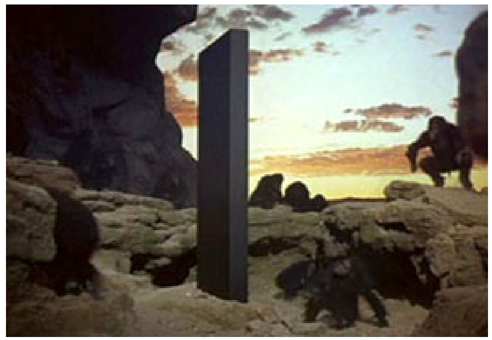
Outline

1. The beginning: The basic notions of statistics
2. Univariate description
   1. Detecting outliers
3. Properties of random sampling
4. Inference
   1. For an individual value
   2. For a mean
   3. Comparison of two means
   4. For a proportion

1. THE BEGINNING: THE BASIC NOTIONS OF STATISTICS

Statistics, a long history

Many years ago

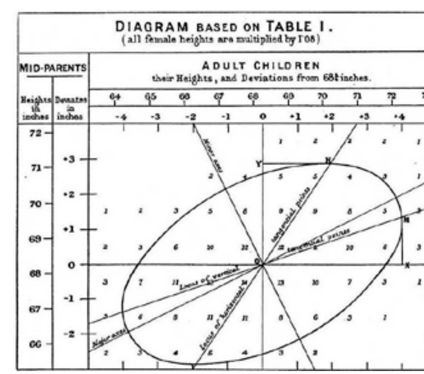


About 5500 years ago, in the Sumerian land men started to collect data of tax records onto dried clay tablets...

= 29086 measures of barley in 37 months. Signed Kushim

1450 First information revolution

and in XIX century we started to analyze it!



Regression origin, Sir Francis Galton (1886).
"Regression towards mediocrity in hereditary stature"

1990 Second information revolution

Thanks to the analysis of data, humanity is heading towards a new religion

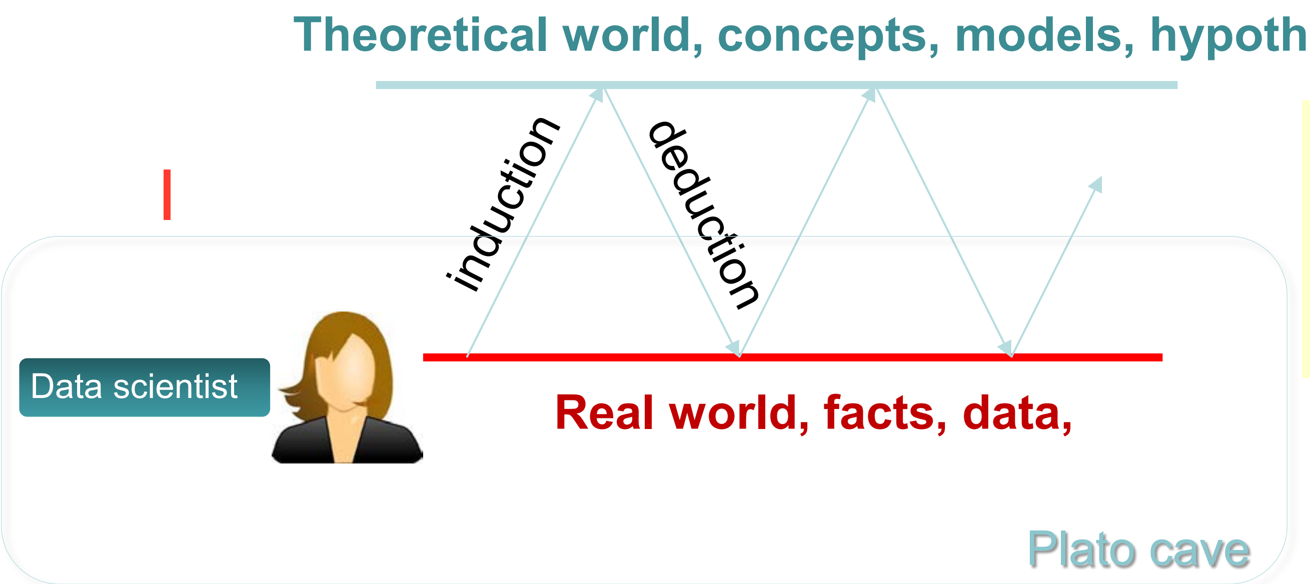
the Dataism

The intelligence is in the algorithms

Yuval Noah Harari
Homo Deus: A Brief History of tomorrow

What is Statistics about

- Etimologically: **accountability of states**
- **Description of data**:
  - Summary statistics: mean, median, variance, quantiles, correlation
  - Graphically.
- **Inference**: *learning from* data (*looking out* of the Plato cave)



Learning is an iteration process between the real world of facts and the hypothesized world of theories

Statistics is the Science to learn from Data

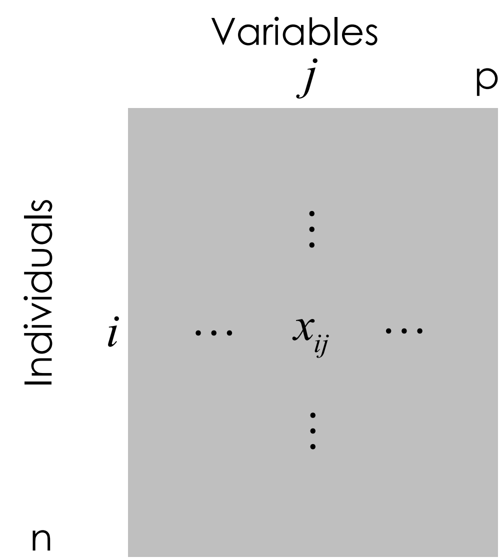
The elements of statistics

Variables: measures taken from individuals (*attributes*, *features*, *fields*,)

- Categorical
  - Binary
  - Nominal
  - Ordinal
- Quantitative
  - Frequencies
  - Continuous
    - Interval
    - Ratio

Individuals: units from which measures are taken (*instances*, *examples*, *records*,)

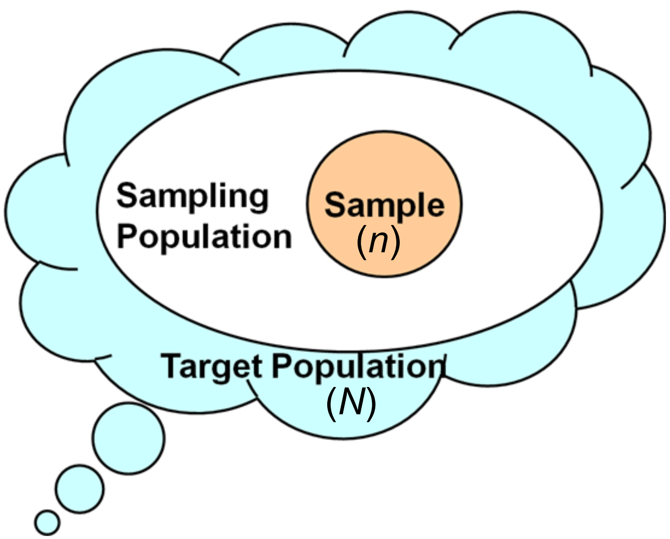
Represented in a **Data Matrix** (Table):



The elements of statistics

- **Sample**: group of individuals.
  - Intentional
  - Representative:
    - Simple random sampling. Stratified. Multi stage
  - Not representative (volunteer,)

- **Population**: *ideal* concept of all individuals concerned by a given problem.



- **Statistic**: i.e. sample average

$$\bar{x} = \frac{\sum x_i}{n}$$

- **Parameter**: i.e. population mean,

$$\mu = \frac{\sum x_i}{N}$$

from data to distributions

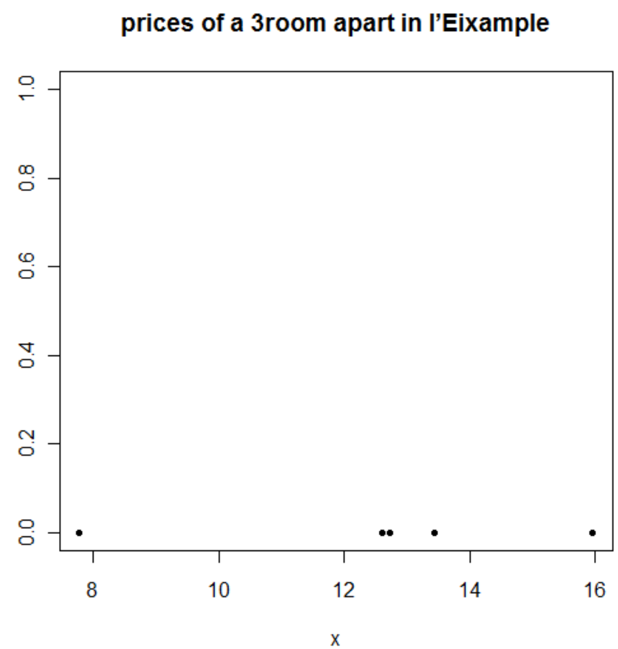
- You have to move to another city, so you decide to buy a nice three room apartment. What are the prices of a three room's apartment in a given district of the new city? (imagine, it is l'Eixample of Barcelona).

- You may start looking the price of some apartments in a real estate magazine (we are in the pre-internet era):

  `7.80, 12.60, 15.96, 12.75, 13.50`

- Are these 5 values at random? (the concept of **Randomisation)**.

- What if we collect a bigger sample?



from data to distributions

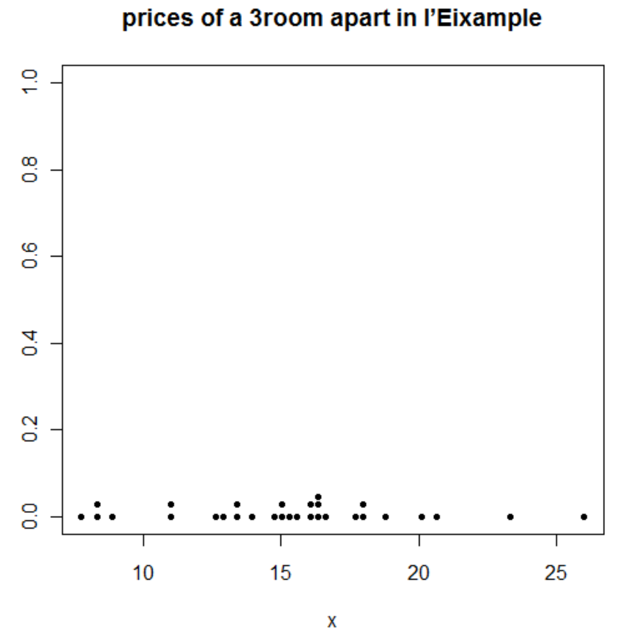
- You have to move to another city, so you decide to buy a nice three room apartment. What are the prices of a three room's apartment in a given district of the new city? (imagine, it is l'Eixample of Barcelona).

- You may start looking the price of some apartments in a real estate magazine (we are in the pre-internet era):

- What if we collect a bigger sample?

n=30

7.80, 12.60, 15.96, 12.75, 13.50, 8.25, 8.25, 15.05, 13.83, 20.61, 16.46, 18.00, 15.40, 16.25, 13.39, 25.99, 16.64, 11.03, 16.28, 23.40, 14.81, 17.60, 20.21, 18.04, 18.70, 11.10, 15.60, 8.83, 15.05, 16.15



from data to distributions

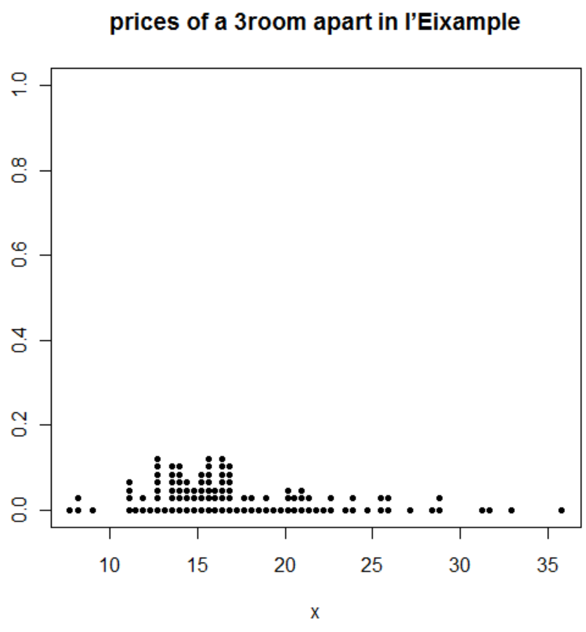
- You have to move to another city, so you decide to buy a new three room apartment. What are the prices of a three room's apartment in a given district of the new city? (imagine, it is l'Eixample of Barcelona).

- You may start looking the price of some apartments in a real estate magazine (we are in the pre-internet era):

- What if we collect a bigger sample?

n=107

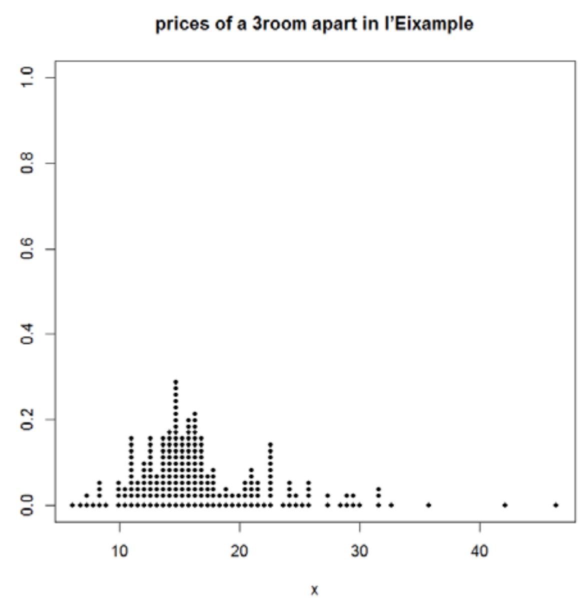
7.80, 12.60, 15.96, 12.75, 13.50, 8.25, 8.25, 15.05, 13.83,
20.61, 16.46, 18.00, 15.40, 16.25, 13.39, 25.99, 16.64, 11.03,
16.28, 23.40, 14.81, 17.60, 20.21, 18.04, 18.70, 11.10, 15.60,
8.83, 15.05, 16.15,
15.00, 11.10, 21.08, 28.66, 21.25, 20.47, 25.53, 10.89, 15.01,
11.78, 14.82, 12.17, 14.56, 16.96, 15.50, 14.43, 14.43, 12.71,
31.66, 15.75, 15.75, 11.69, 14.52, 17.35, 22.57, 20.00, 13.80,
13.68, 12.61, 19.00, 24.61, 16.80, 16.72, 23.95, 16.11, 19.41,
13.99, 16.48, 13.20, 13.47, 13.63, 14.76, 16.93, 31.31, 12.81,
21.81, 20.82, 35.77, 15.54, 12.62, 13.91, 21.18, 13.72, 12.00,
19.89, 16.46, 32.70, 22.73, 15.51, 16.26, 28.70, 18.90, 25.75,
16.89, 13.99, 13.99, 28.22, 20.79, 16.81, 20.25, 22.31, 24.03,
15.65, 27.28, 12.60, 17.55, 25.60



from data to distributions

- You have to move to another city, so you decide to buy a new three room apartment. What are the prices of a three room's apartment in a given district of the new city? (imagine, it is l'Eixample of Barcelona).

- You may start looking the price of some apartments in a real estate magazine (we are in the pre-internet era):



- What if we take all available data?

n=221

Historical available data

Prices of 3 room apartment in l'Eixample:

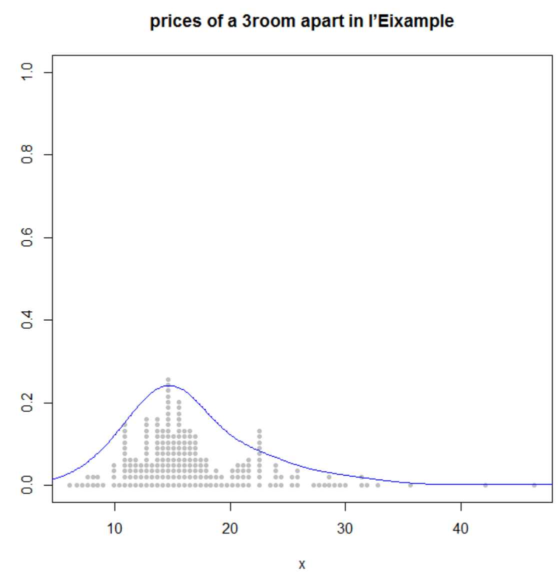
Large historical data recorded (n=221) in the last year, (in the magazine):

```
  [1] 11.10 11.05 16.90 13.91 14.43  8.25 13.87  8.58 12.17 22.31 16.20 14.28
 [13] 17.35 12.81 25.53  7.52 16.00 10.92 15.61 14.24 29.44 25.99 21.70 24.03
 [25] 14.36 23.96 18.90 23.40 24.61 14.80 16.25 20.21 10.06 12.61 13.99 14.76
 [37] 17.32 11.02 16.46 32.70 14.11 14.52  6.21 42.12 16.81 10.15 14.87 15.60
 [49] 12.08 12.10 15.05 15.00 22.31 20.82 18.26 46.32 14.43 10.89 20.25 16.48
 [61] 21.25 12.75 10.56 13.50 13.92 12.48 17.64  8.25 13.65  7.80 25.75 11.50
 [73] 22.62 31.35 27.28 15.69 16.46 19.00 14.81 22.57 19.42 14.53 14.51 27.44
 [85] 35.77 15.96 14.60 14.95 21.18 15.40 16.64 12.63 22.73 12.60 20.47 31.31
 [97] 14.35 13.47 14.43 12.62 21.08 22.44 12.75 22.75 16.46 16.26 13.44 21.81
[109] 15.72 16.14 18.00 16.80 15.73 28.22 14.79 12.60 10.65 12.00 17.55 15.65
[121] 14.82 17.75 16.23 17.60 16.96 14.54 15.50 10.00 15.51 11.78 11.10 15.20
[133] 13.12 10.66 15.00 13.39 28.70  7.14 13.80 19.89 13.63 16.59 11.11 20.00
[145] 13.68 14.43 16.72 16.59 11.20 16.11 15.75 11.69 13.57 25.40 13.83 15.05
[157] 22.60 13.99 14.56 15.91 15.75 20.79 31.66 10.71 29.29 18.70 11.84 12.68
[169] 23.95 16.83 13.99 21.42 14.72 21.68 13.32  8.83 30.10 12.53 13.89 15.50
[181] 14.51 15.39 17.10 14.80 18.56 19.41 16.01 13.20 15.99 12.96 24.00 15.01
[193] 15.54 12.71 11.10  8.40 15.40 13.72 25.60 22.36 15.11 17.55 10.14 16.89
[205] 18.00 11.03 28.66 12.00 16.15 21.75 22.70 14.62 18.04 20.61 11.31 24.60
[217]  6.76 21.12 16.93 16.28 11.38
```

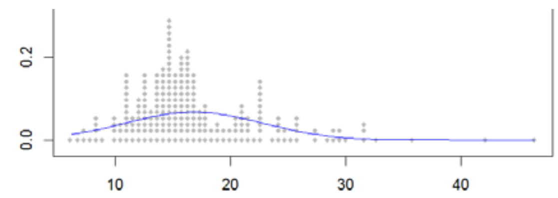
from data to distributions

Values of a variables are distributed according to ***a theoretical distribution***, some values are very popular, others very rare.

n=221

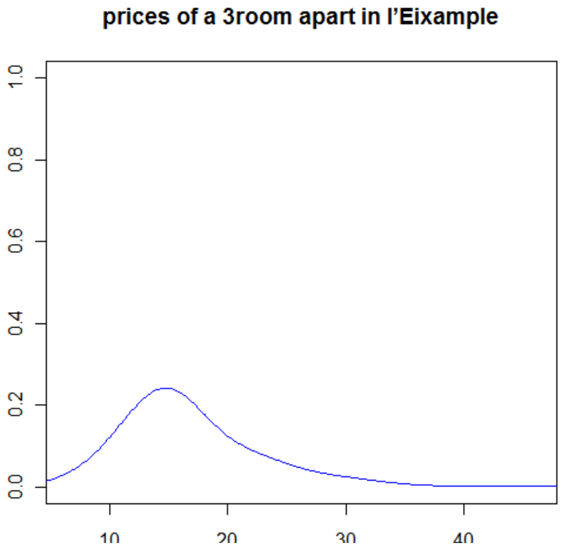


Theoretical distribution with gaussian assumption; (outliers problem)



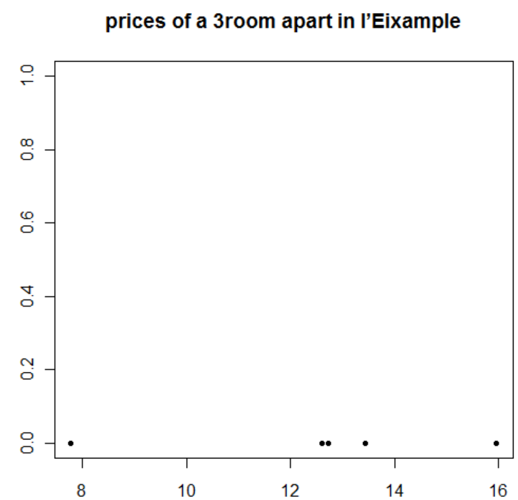
the data generating mechanism

We assume that data is generated by a certain *unique* mechanism.



Population of prices of 3 room apartments in l'Eixample

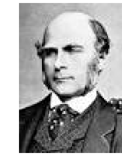
Plato cave



7.80 12.60 15.96 12.75 13.50

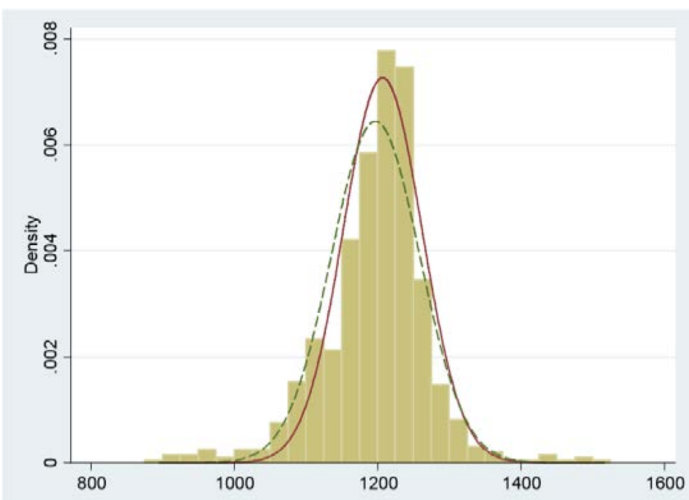
Data (sample, n=5)

Why *random* samples are representative?



(1822-1911)

In the fall of 1906 Sir Francis Galton went to country fair, the annual West England Fat Stock and Poultry Exhibition; as he walked down through the exhibition Galton came to a weight-judging competition "a fat ox had been selected and people placed their bets on the weight of the ox after having slaughtered and dressed on a written piece of paper. The best guesses receives prizes. 787 people tried their luck, many of them were butchers and farmers but there were also passing by people with no insider knowledge of cattle. Galton had curiosity about would be the "average voter", thinking that it was capable of very little. After all, a mix of mediocre people can't be better than the opinion of few very smart group of people. So, he borrowed all the tickets with the guesses to see if they formed a bell curve, and computed its mean, 1197 pounds. The true weight of the ox was 1198 pounds.



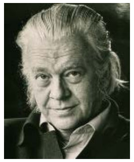
Conditions for the validity of the average estimate:

- Independence on the guess of each wager
- Diversity of people waging

Statisticians express this, saying that the sample should be ***random***

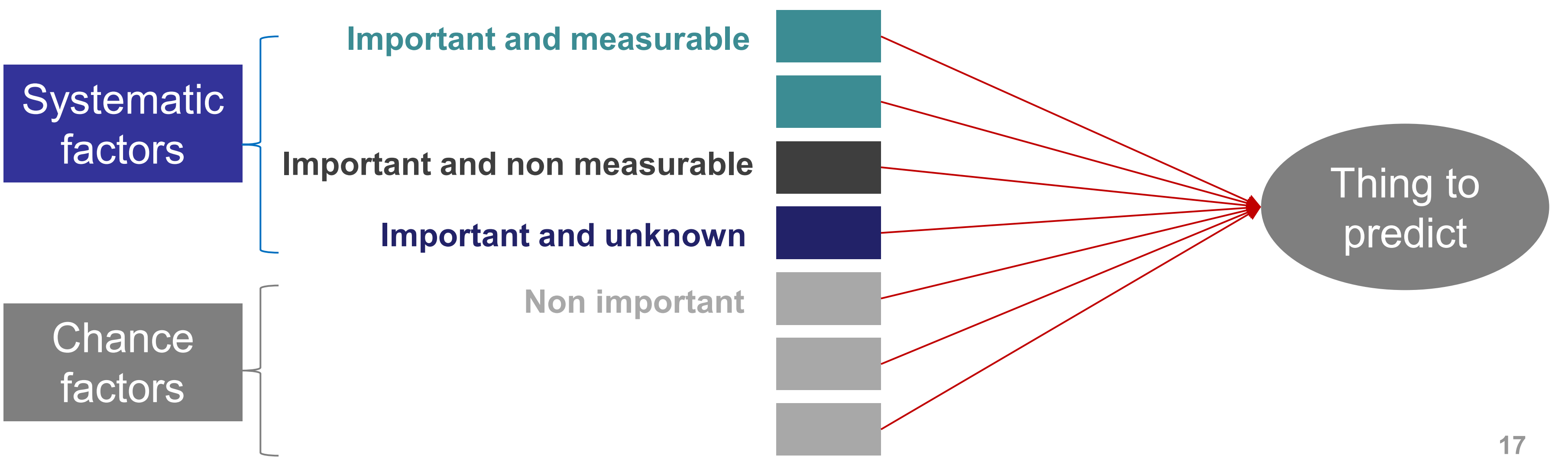
The Wisdom of Crowds, James Surowiecki. Abacus, 2004.

The paradigm of Statistics

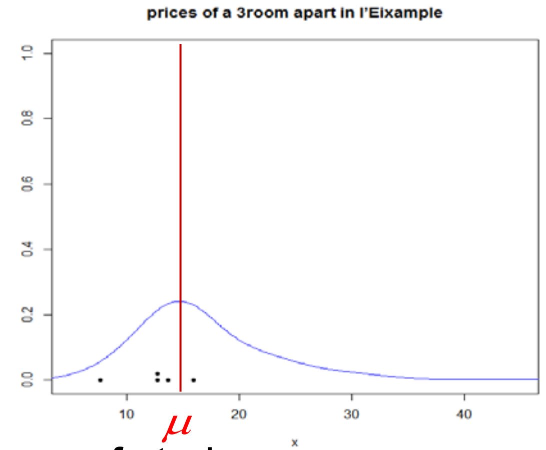


When the Lord created the world and people to live in –an enterprise which according modern science, took a very long time – I could well imagine that He reasoned with Himself as follows: “If I make everything predictable, these human beings, whom I have endowed with pretty good brains, will undoubtedly learn to predict everything, and they will thereupon have no motive to do anything at all, because they will recognize that the future is totally determined and cannot be influenced by any human action. On the other hand, if I make everything unpredictable, they will gradually discover that there is no rational basis for any decision whatsoever and, as in the first case, they will thereupon have no motive to do anything at all. Neither scheme would make sense. I must therefore create a mixture of the two. Let some things be predictable and let others be unpredictable. They will then, amongst many other things, have the very important task of finding out which is which”. *E.F. Schumacher*. **Small is beautiful**, 1973.

Factors influencing the thing to predict



The paradigm of Statistics



Data = Fit + Noise

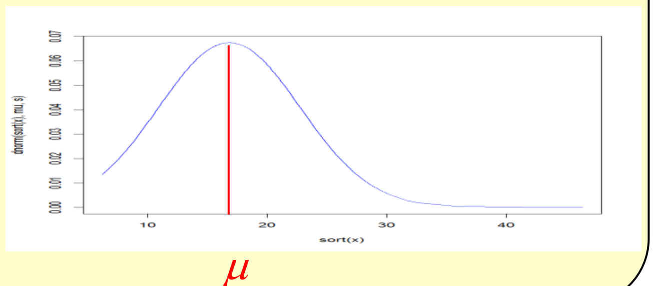
$$\boxed{x_i = \mu + \varepsilon_i}$$

Data: Result of all factors driving the phenomenon of study
Fit: Result of all relevant factors. *Systematic part*
Noise: Result of all irrelevant (small importance) factors driving the phenomenon of study. *Experimental error*. *Random fluctuation*

Central Limit Theorem:

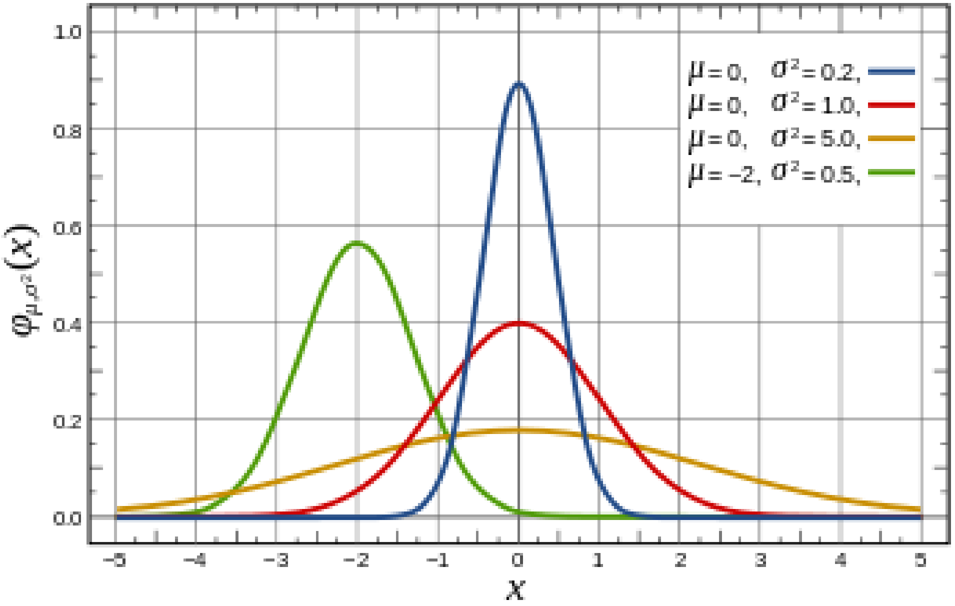
If there are plenty of irrelevant factors,
all of them with similar importance,
then, the generating mechanism is normal:

$$x \sim N(\mu, \sigma^2) \quad \equiv \quad \varepsilon \sim N(0, \sigma^2)$$



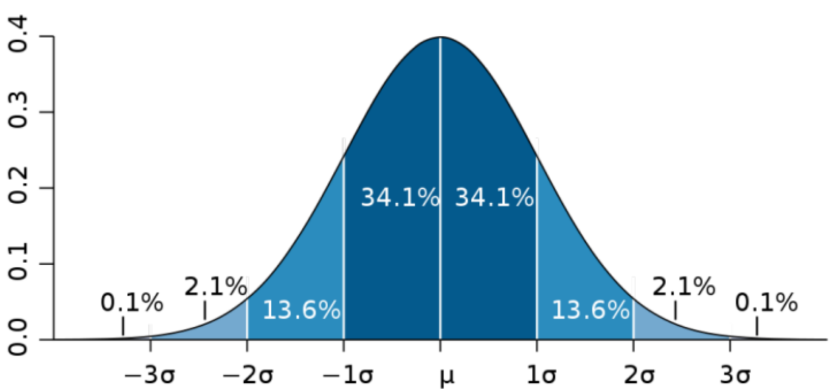
The standardized $N(0,1)$

$$x \sim N(\mu, \sigma^2)$$



$N(0,1)$

$$\rightarrow z = \frac{x-\mu}{\sigma} \sim N(0,1)$$



Every variable measured in standard deviations from its mean

2. UNIVARIATE DESCRIPTION

Numerical description of a variable

Summary statistics of a variable:

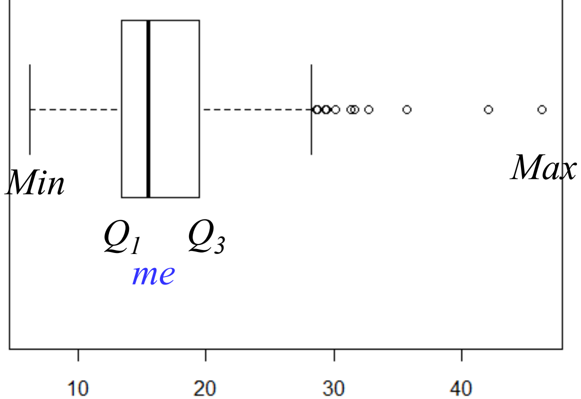
Central value:
Mean: $\overline{x} = \sum_{i=1}^{n} x_i \frac{1}{n}$

Population
$\mu = E[x] = \int x\, f(x)dx$

Median: $me = F^{-1}(0.50)$

```
summary(x)    = 221 prices of historical data
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   6.21   13.39   15.50   16.81   19.41   46.32
boxplot(x, horizontal=T)
```

Boxplot (Tukey, 1977)
Graphical description



Spread:
Variance: $s^2 = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n-1}$

Population
$\sigma^2 = \mathrm{var}[x] = \int (x-\mu)^2 f(x)dx$

$Standard\ deviation = \sqrt{Variance}$

```
sd(x)   5.915688
```

Interquartile range: $IQR = Q_3 - Q_1$

```
Q1  <- summary(x)[2]
Q3  <- summary(x)[5]
iqr <- Q3 - Q1
iqr    6.02
```

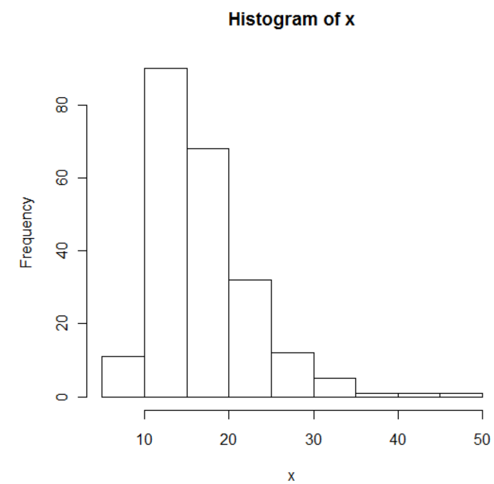
Useful for outlier detection

Graphical description of a variable

Histogram

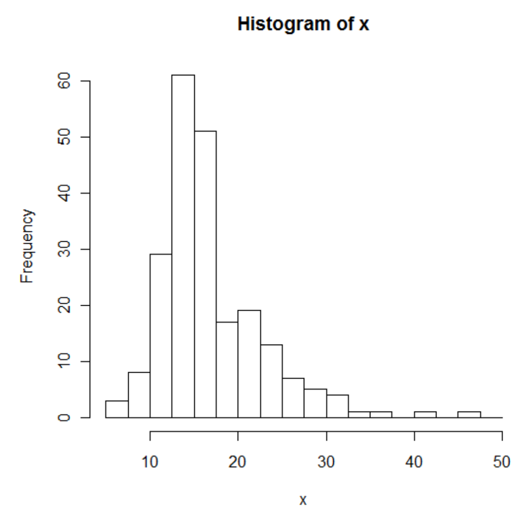
The classical graphical description of continuous variables

```
hist(x)
```



The shape depends of the bin length

```
hist(x, breaks=seq(from=5, to=50, by=2.5))
```



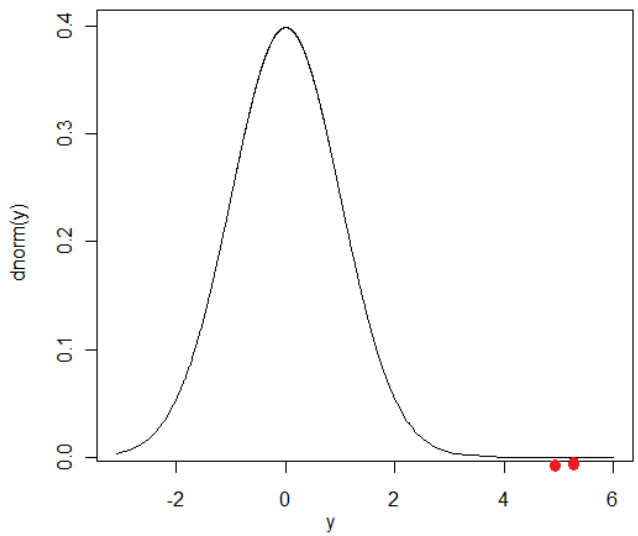
2.1 Outlier detection

What is an outlier? "An outlier is an observation which deviates so much from the other observations"

Statistics-based intuition. Data follow a "theoretical distribution generating data mechanism", defined by a given process. Outlying data may be due a:

- very unlikely events for the assumed generating mechanism
- data following a different generating mechanism

If Normal generating mechanism:



if x~N(0,1)	Prob(x≥X)
1	0.1586553
2	0.02275013
3	0.001349898
4	3.167124e-05
5	2.866516e-07

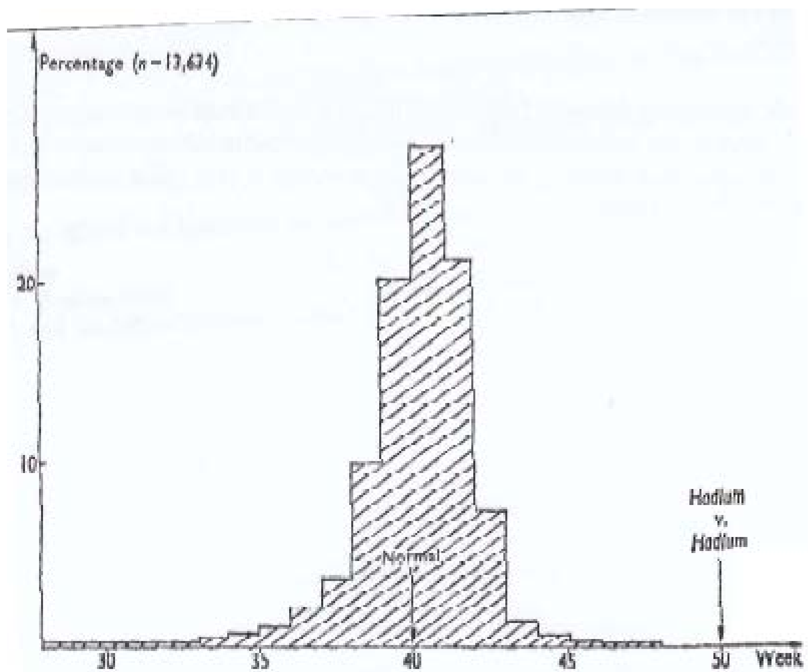
Hadlum versus Hadlum case (1949)

The study of outliers. Vic Barnett, (1978)

The birth of a child to Mrs. Hadlum happened 349 days after Mr. Hadlum left for military service.

Average human gestation period is 280 days (40 weeks).

Statistically, 349 days is an outlier.



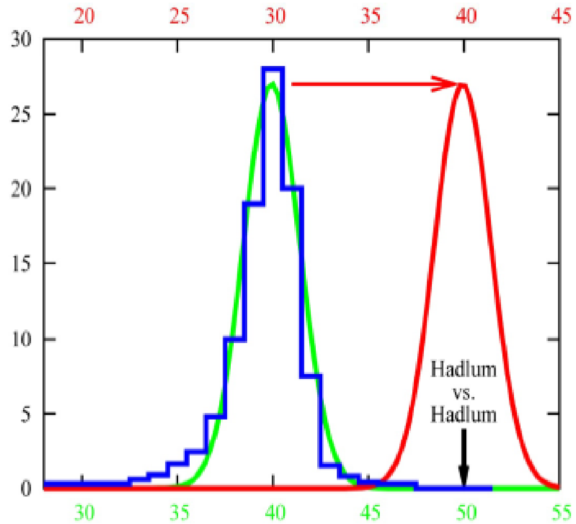
Distribution of human gestation periods.

The Hadlum vs. Hadlum case on statistical grounds

blue: statistical basis (13634 records of historical data of gestation periods)

green: assumed underlying Gaussian process. Very low probability for the birth of Mrs. Hadlums child for being generated by this process

red: assumption of Mr. Hadlum: Another Gaussian process responsible for the observed birth, where the gestation period starts later. Under this assumption the specific birthday has highest-probability.



Univariate outlier detection

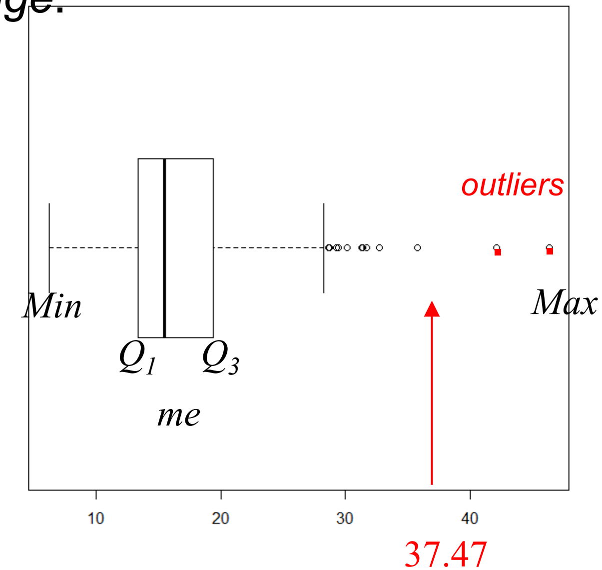
Using a Boxplot

An observation x_i is declared ***"potential" outlier***, if it lies outside of the interval:

$$\left[Q_1 - 3 \times IQR, Q_3 + 3 \times IQR\right]$$

where $IQR = Q_3 - Q_1$ is called the *Interquartile Range*.

```
Q1 - 3*iqr
  -4.67
Q3 + 3*iqr
  37.47
```




```
> x0[which(x0 > Q3 + 3*iqr)]
[1] 46.32 42.12
```

The number $3 * IQR$ are chosen by comparison with a normal distribution. If x ~ *Normal :*

Prob$(x \geq Q_3 + 3 \times IQR)$ = *1.170971e-06*

3. PROPERTIES OF RANDOM SAMPLING

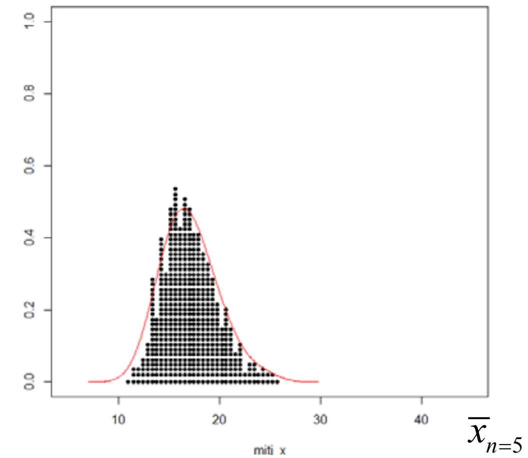
Do we really need normal generating mechanisms?

Usually we are only interested in the **mean** of a variable.

What if a take 1000 samples at random of apartments of L'Eixample of size 5 on our historical data and display the computed mean:

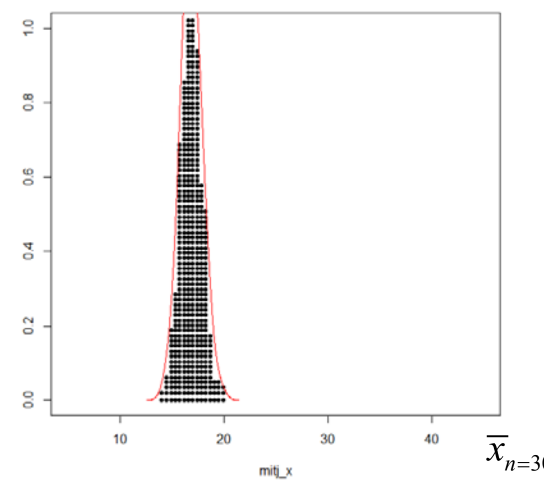
Then, we compute the mean and variance of the distribution of means:

```
mean(x)        16.81149
mean(mitj_x)   16.79399
var(x)         34.99536
var(mitj_x)     6.958161
var(x)/5        6.999073
```



What if a take 1000 samples at random of apartments of L'Eixample of size 30 :

```
mean(mitj_x)   16.78824
var(mitj_x)     1.018578
var(x)/30       1.166512
```



We see that increasing n, the distribution of the mean $\overline{x}_n$ tends towards a Normal distribution

$$x \sim Any(\mu, \sigma^2)$$

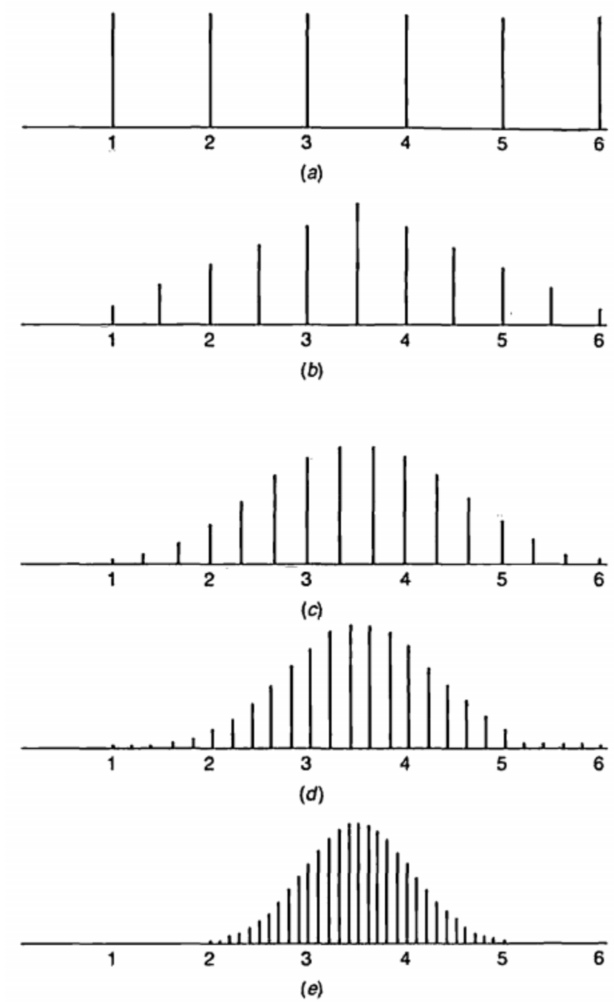
$$\overline{x}_n = \frac{x_1 + \cdots + x_n}{n} \quad \rightarrow CLT: \ \overline{x}_n \sim N(\mu, \sigma^2/n)$$

Illustration of the CLT

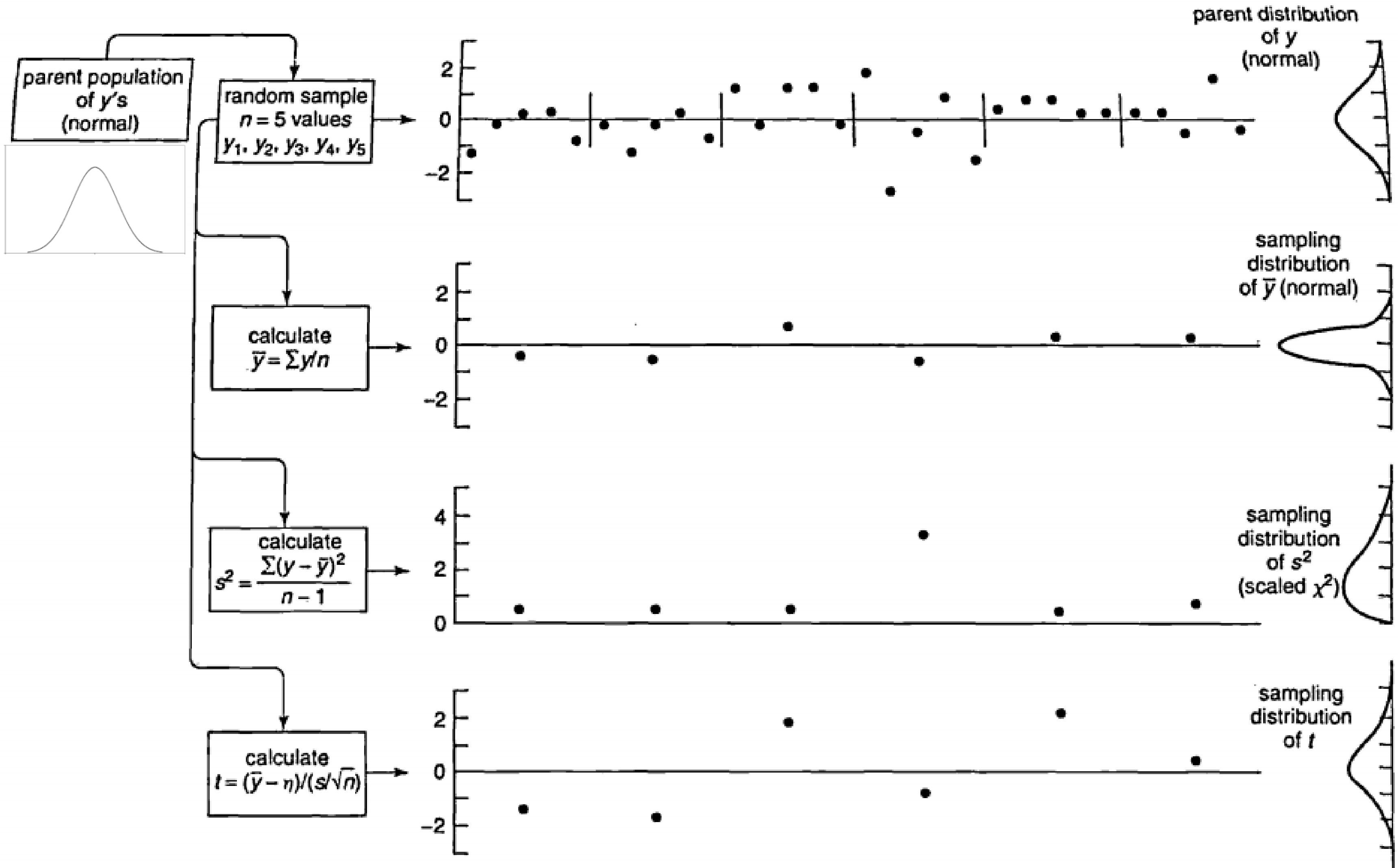
Tendency to the normal distribution of averages

Distribution of averages scores from throwing various numbers of dice:

(a) $n=1$
(b) $n=2$
(c) $n=3$
(d) $n=5$
(e) $n=10$



Random Sampling from $N(\mu,\sigma^2)$



4. INFERENCE

IMAGINE OUTSIDE THE CAVE

The tool: Statistical testing

Statistical tests are the most important statistical contribution to scientific progress. They constitute the tool for any scientific advance of all experimental sciences. They follow the Karl Popper principle that only a hypothesis can be taken as scientific if it is falsifiable.

Elements of a Hypothesis Test:

Null hypothesis (H_0): Default hypothesis.

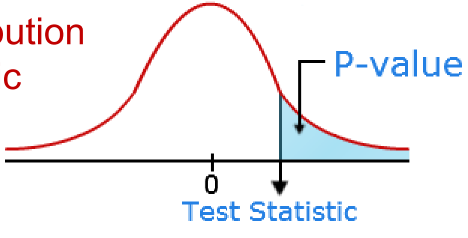
Alternative hypothesis (H_1): The plausible variation of the null hypothesis that we want to validate.

Test statistic: A statistic value suitable to compare the null hypothesis respect to its alternative *(it depends on the problem)*.

Reference distribution: Distribution of the test statistic if the H_0 is true.

p.value: Probability of the observed test statistic if H_0 is true. If it is too low, the null hypothesis is disaccredited, hence rejected, (we say that the test statistic *is significant*).

Reference distribution of the test statistic under the null hypothesis



Some relevant statistical tests:

- Salk vaccine trial for the polio, `p.value=6.56` $\times 10^{-11}$ (1954).
- Talpiot tomb (of Jesus family), `p.value=0.0017` (1980)
- Discovery of the Higgs boson, `p.value=3` $\times 10^{-7}$ (2012).

4.1 Inference for individual values

From historical data

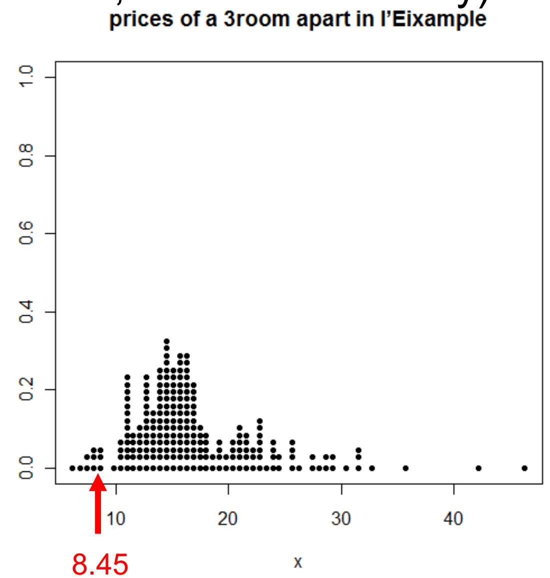
A friend of mine offers me a 3room apart. in l'Eixample for 8.45MPts. Should I buy it?

To decide whether 8.45 is an opportunity or not (skip all other factors for buying), we need a ***Reference Distribution*** of prices (always we have it, but unconsciously).

Reference Distribution from historical data:
If we have collected a large enough data on prices of 3room aparts in l'Eixample, we could use these values as *Reference Distribution*.

There are 8 aparts. out of 221 with a lower price.

$$p.value = \frac{8}{221} = 0.0362$$



The ***p.value*** represents how likely is the 8.45 value in the Reference Distribution of historical prices. 3.62% of times we could expect to find an apart. with this or lower price.

Inference for individual values

Without historical data

If we don't have historical data, we must build the Reference Distribution making assumptions and using theory.

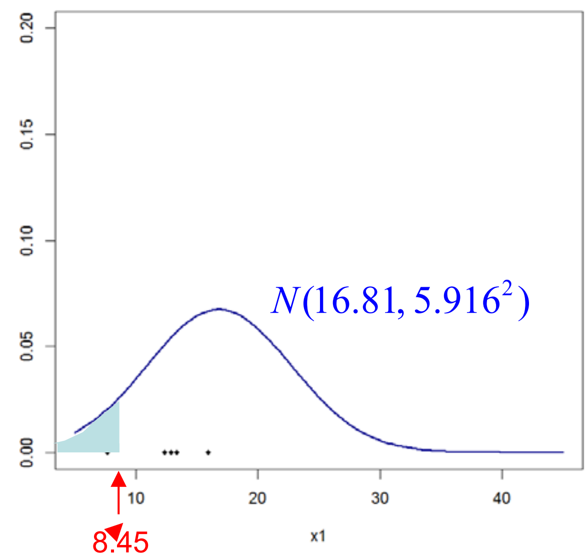
Lets assume

$prices = x \sim N(\mu, \sigma^2)$ *convenience assumption*

Then, if we don't have historical data BUT we have a fairly good estimation of the mean of 3room aparts. and its standard deviation. Thus, we assume μ=16.81 and σ=5.916

$$z_i = \frac{x_i - \mu}{\sigma} \sim N(0,1)$$

```
z <- (8.45 - 16.81)/5.916
pnorm(z)
0.07876
```



Signal to noise statistics

Most statistical tests are based in a Signal/Noise ratio:

$$\frac{Signal}{Noise} = \frac{Detected\ effect}{random\ \ fluctuation\ measure} = \frac{statistic - E\left[statistic\right]}{\sqrt{\operatorname{var}\left(statistic\right)}} = \cdots$$

$$z_i = \frac{x_i - \mu}{\sigma} \qquad t_i = \frac{x_i - \overline{x}}{s} \qquad t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \qquad F = \frac{\sum_{k=1}^{q} n_k (\overline{x}_k - \overline{x})^2 \Big/ q-1}{s^2_{res}}$$

Inference for individual values

Without historical data, and without any idea of the generating process of data

If we don't have historical data without any idea about the central value and dispersion of the distribution. We need to collect a ***random sample*** to estimate μ and σ^2.
Lets be our sample of 5 prices at random: `7.80, 12.60, 15.96, 12.75, 13.50`

The *unique* information comes from the sample

```
x <- c(7.80, 12.60, 15.96, 12.75, 13.50)
```

```
mean(x)   12.522
sd(x)      2.963599
```

$$\overline{x} \to \mu$$

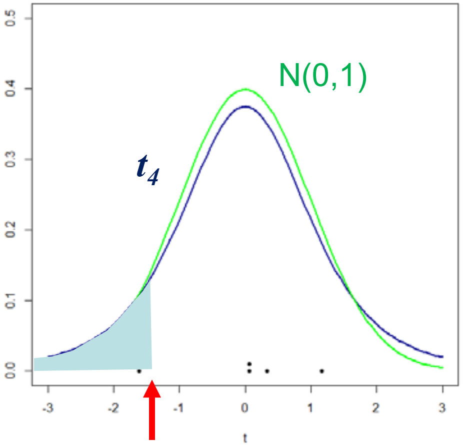
$$s^2 \to \sigma^2$$

$$t_i = \frac{x_i - \overline{x}}{s} \sim Student's\ t_{n-1}$$

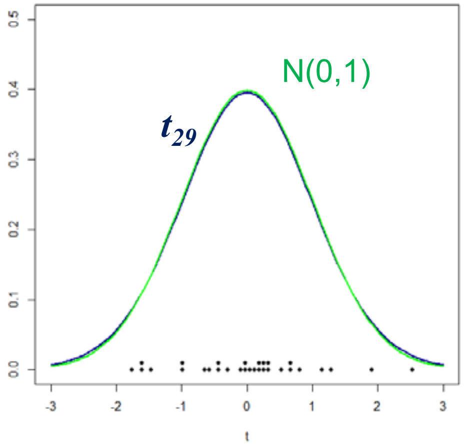
William Gosset "Student", English, 1876, 1937

$$t = \frac{8.45 - 12.522}{2.9636} = -1.374$$

```
pt(t, df=(n-1))
0.1207
```



What if we take a bigger sample, n=30:



4.2 Inference for a mean

Comparison with a nominal value

Now, the problem is to compare a *raw batch* of actual data with a established nominal value, for instance we may ask whether the prices of L'Eixample on average are equal to the overall price of Barcelona which is 15.289. This allows us to assess whether apartments in l'Eixample are more expensive, equal or less expensive than the overall Barcelona.

We will say that L'Eixample is equal priced if: $H_0 : \quad \mu_{EIX} = \mu_{BCN}$ $\qquad \mu_{BCN} = 15.289 = \mu_0$

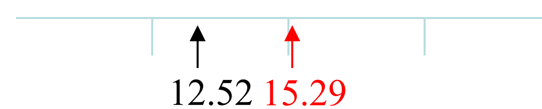
Our suspicion is L'Eixample is overpriced : $H_1 : \quad \mu_{EIX} > \mu_{BCN}$

We take our sample: `x <- c(7.80, 12.60, 15.96, 12.75, 13.50)`

```
mean(x)   12.522
sd(x)      2.963599
```

$\overline{x} \rightarrow \mu_{EIX}$

our data says: $\overline{x} < \mu_{BCN}$ Does it imply that H_0 holds?



To answer that question we need the ***Reference Distribution*** of $\overline{x}_{n=5}$

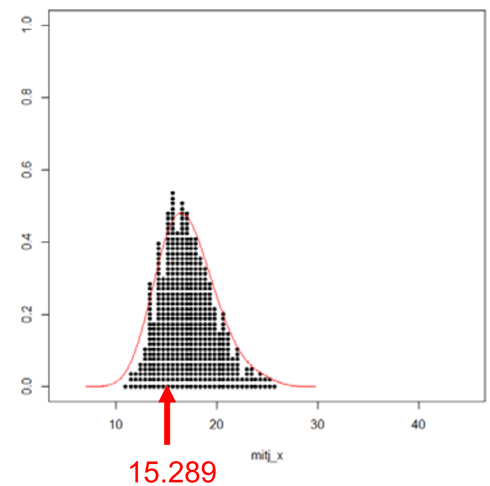
Inference for a mean

Comparison with a nominal value

- Reference distribution from historical data

From the historical data, we take, say 1000 samples of size 5 at random, and we compute their average. Then, we count how many samples have an average lower of equal than 15.289 (=mu0)

```
sum(mitj_x < mu0)/1000
0.277
```



- Reference distribution without historical data

$$\bar{x}_n \sim N(\mu, \sigma^2/n) \leftarrow N(\bar{x}, s^2/n)$$

$$\bar{x} \to \mu$$

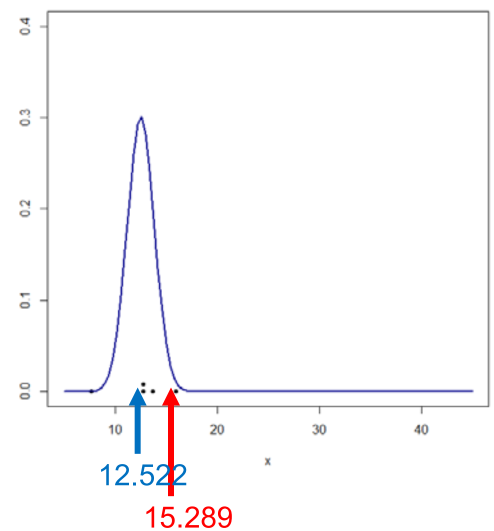
$$s^2 \to \sigma^2$$

Assuming random sample and $x \sim N(\mu, \sigma^2)$

$$t = \frac{\mu_0 - \bar{x}}{s/\sqrt{n}} \sim t_{df=n-1}$$

```
t <- sqrt(n)*(mu0 - mean(x))/sd(x)
pt(t, df=4)
0.9474
```

Classical critical threshold: 0.05



4.3 Comparison of two means

Very often, the problem consist of comparing two different methods, two treatments (clinical data, chemistry), two processes (in management) ... Usually, one corresponds to the standard method, whereas the other represents the improved one and we want to assess whether the improvement is worth or not.

Two solve this problem we need two samples, one for the method called *A*, and the other for method called *B*.

$$(x_{1A}, x_{2A}, \ldots, x_{n_a A})$$

$$(x_{1B}, x_{2B}, \ldots, x_{n_b B})$$

Lets take method *A* as the standard one and method *B* as the reengineered, let's *x* represents a measure of performance for each method. The obtained data has been:

x_A:	89.7	81.4	84.5	84.8	87.3	79.7	85.1	81.7	83.7	84.5
x_B:	84.7	86.1	83.2	91.9	86.3	79.3	82.6	89.1	83.7	88.5

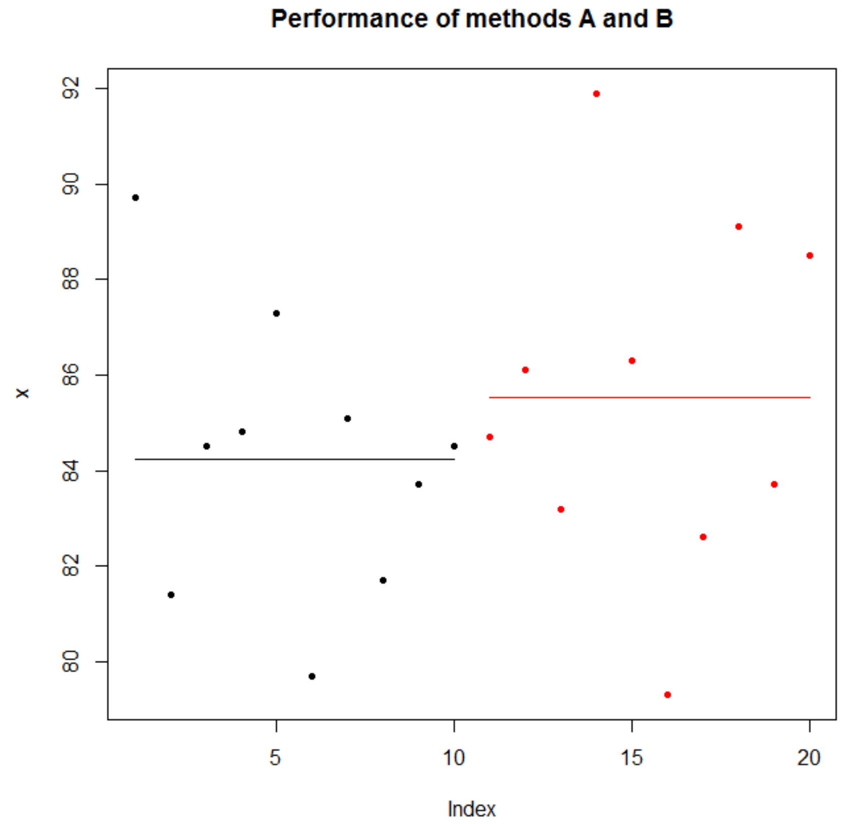
Can we say that improved method *B* is better than the current *A*?

Comparison of two means

Runs have been obtained sequentially.
Is this at random?

$n_a = 10$ $\quad \bar{x}_A = 84.24$ $\quad \bar{x}_B - \bar{x}_A = 1.30 > 0$

$n_b = 10$ $\quad \bar{x}_B = 85.54$



Can we assess that method *B* is more efficient than method *A*?

We will say that method *B* is more efficient than method *A* if :

$$H_1 : \quad \mu_B - \mu_A > 0$$

Whereas, will say that method *B* is equal efficient than method *A* if :

$$H_0 : \quad \mu_B - \mu_A = 0$$

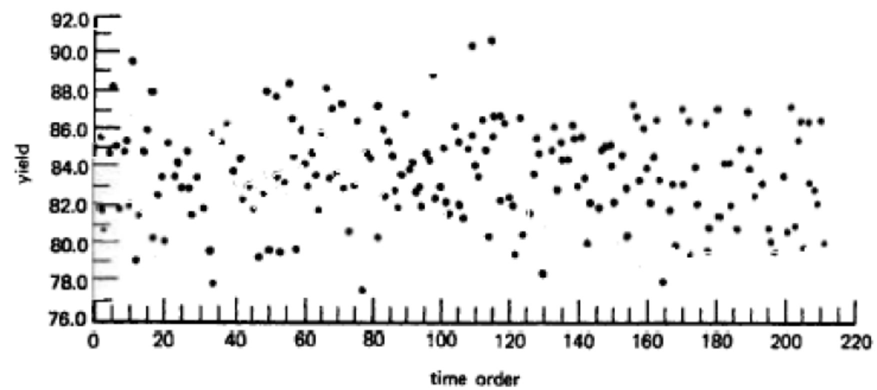
To assess this, we need the **Reference Distribution** of $\bar{x}_B - \bar{x}_A$ i.e, what would be the values of $\bar{x}_B - \bar{x}_A$ with different samples, if method *B* is equal efficient than method *A*?.

Comparison of two means

From historical data

From our past records we have 210 measures of the performance obtained before the trial, using the standard method A:

```
  [1] 85.5 81.7 80.6 84.7 88.2 84.9 81.8 84.9 85.2 81.9
 [11] 89.4 79.0 81.4 84.8 85.9 88.0 80.3 82.6 83.5 80.2
 [21] 85.2 87.2 83.5 84.3 82.9 84.7 82.9 81.5 83.4 87.7
 [31] 81.8 79.6 85.8 77.9 89.7 85.4 86.3 80.7 83.8 90.5
 [41] 84.5 82.4 86.7 83.0 81.8 89.3 79.3 82.7 88.0 79.6
 [51] 87.8 83.6 79.5 83.3 88.4 86.6 84.6 79.7 86.0 84.2
 [61] 83.0 84.8 83.6 81.8 85.9 88.2 83.5 87.2 83.7 87.3
 [71] 83.0 90.5 80.7 83.1 86.5 90.0 77.5 84.7 84.6 87.2
 [81] 80.5 86.1 82.6 85.4 84.7 82.8 81.9 83.6 86.8 84.0
 [91] 84.2 82.8 83.0 82.0 84.7 84.4 88.9 82.4 83.0 85.0
[101] 82.2 81.6 86.2 85.4 82.1 81.4 85.0 85.8 84.2 83.5
[111] 86.5 85.0 80.4 85.7 86.7 86.7 82.3 86.4 82.5 82.0
[121] 79.6 86.7 80.5 91.7 81.6 83.9 85.6 84.8 78.4 89.9
[131] 85.0 86.2 83.0 85.4 84.4 84.5 86.2 85.6 83.2 85.7
[141] 83.5 80.1 82.2 88.6 82.0 85.0 85.2 85.3 84.3 82.3
[151] 89.7 84.8 83.1 80.6 87.4 86.8 83.5 86.2 84.1 82.3
[161] 84.8 86.6 83.5 78.1 88.8 81.9 83.3 80.0 87.2 83.3
[171] 86.6 79.5 84.1 82.2 90.8 86.5 79.7 81.0 87.2 81.6
[181] 84.4 84.4 82.2 88.9 80.9 85.1 87.1 84.0 76.5 82.7
[191] 85.1 83.3 90.4 81.0 80.3 79.8 89.0 83.7 80.9 87.3
[201] 81.1 85.6 86.6 80.0 86.6 83.3 83.1 82.3 86.7 80.2
```



Comparison of two means

Reference Distribution from historical data

To build the reference distribution, we subtract the mean of 10 sequential observations from the batch of the previous 10 sequential observations.

x_i	$\overline{x}_{i\cdots i+9}$	x_i	$\overline{x}_{i\cdots i+9}$	x_i	$\overline{x}_{i\cdots i+9}$	x_i	$\overline{x}_{i\cdots i+9}$	x_i	$\overline{x}_{i\cdots i+9}$	x_i	$\overline{x}_{i\cdots i+9}$
85.5		84.5	84.42	80.5	84.53	79.5	83.72	84.8	84.36	81.1	83.68
81.7		82.4	84.70	86.1	84.09	86.7	83.89	86.6	84.54	85.6	83.91
80.6		86.7	84.79	82.6	84.28	80.5	83.90	83.5	84.58	86.6	83.53
84.7		83.0	85.30	85.4	84.51	91.7	84.50	78.1	84.33	80.0	83.43
88.2		81.8	84.51	84.7	84.33	81.6	83.99	88.8	84.47	86.6	84.06
84.9		89.3	84.90	82.8	83.61	83.9	83.71	81.9	83.98	83.3	84.41
81.8		79.3	84.20	81.9	84.05	85.6	84.04	83.3	83.96	83.1	83.82
84.9		82.7	84.40	83.6	83.94	84.8	83.88	80.0	83.34	82.3	83.68
85.2		88.0	84.82	86.8	84.16	78.4	83.47	87.2	83.65	86.7	84.26
81.9	83.94	79.6	83.73	84.0	83.84	89.9	84.26	83.3	83.75	80.2	83.55
89.4	84.33	87.8	84.06	84.2	84.21	85.0	84.81	86.6	83.93		
79.0	84.06	83.6	84.18	82.8	83.88	86.2	84.76	79.5	83.22		
81.4	84.14	79.5	83.46	83.0	83.92	83.0	85.01	84.1	83.28		
84.8	84.15	83.3	83.49	82.0	83.58	85.4	84.38	82.2	83.69		
85.9	83.92	88.4	84.15	84.7	83.58	84.4	84.66	90.8	83.89		
88.0	84.23	86.6	83.88	84.4	83.74	84.5	84.72	86.5	84.35		
80.3	84.08	84.6	84.41	88.9	84.44	86.2	84.78	79.7	83.99		
82.6	83.85	79.7	84.11	82.4	84.32	85.6	84.86	81.0	84.09		
83.5	83.68	86.0	83.91	83.0	83.94	83.2	85.34	87.2	84.09		
80.2	83.51	84.2	84.37	85.0	84.04	85.7	84.92	81.6	83.92		
85.2	83.09	83.0	83.89	82.2	83.84	83.5	84.77	84.4	83.70		
87.2	83.91	84.8	84.01	81.6	83.72	80.1	84.16	84.4	84.19		
83.5	84.12	83.6	84.42	86.2	84.04	82.2	84.08	82.2	84.00		
84.3	84.07	81.8	84.27	85.4	84.38	88.6	84.40	88.9	84.67		
82.9	83.77	85.9	84.02	82.1	84.12	82.0	84.16	80.9	83.68		

$$\overline{x}_{11\cdots 20} - \overline{x}_{1\cdots 10}$$
$$\overline{x}_{12\cdots 21} - \overline{x}_{2\cdots 11}$$
$$\overline{x}_{13\cdots 22} - \overline{x}_{3\cdots 12}$$
$$\cdots$$
$$\overline{x}_{201\cdots 210} - \overline{x}_{191\cdots 200}$$

	dif_A
83.51-83.94=	-0.43
83.09-84.33=	-1.24
83.91-84.06=	-0.15
84.12-84.14=	-0.02
84.07-84.15=	-0.08
83.77-83.92=	-0.15

Comparison of two means

Reference Distribution from historical data

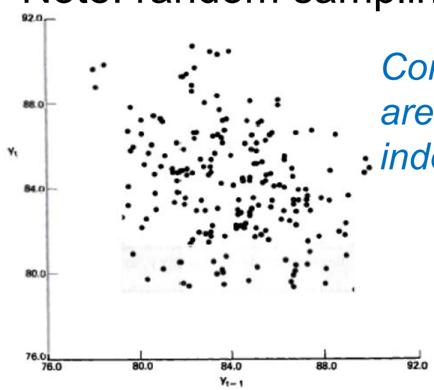
The Reference Distribution of the difference of two consecutive batches of 10 observations each, using the standard method A:

$$\overline{x}_B - \overline{x}_A = 1.30$$

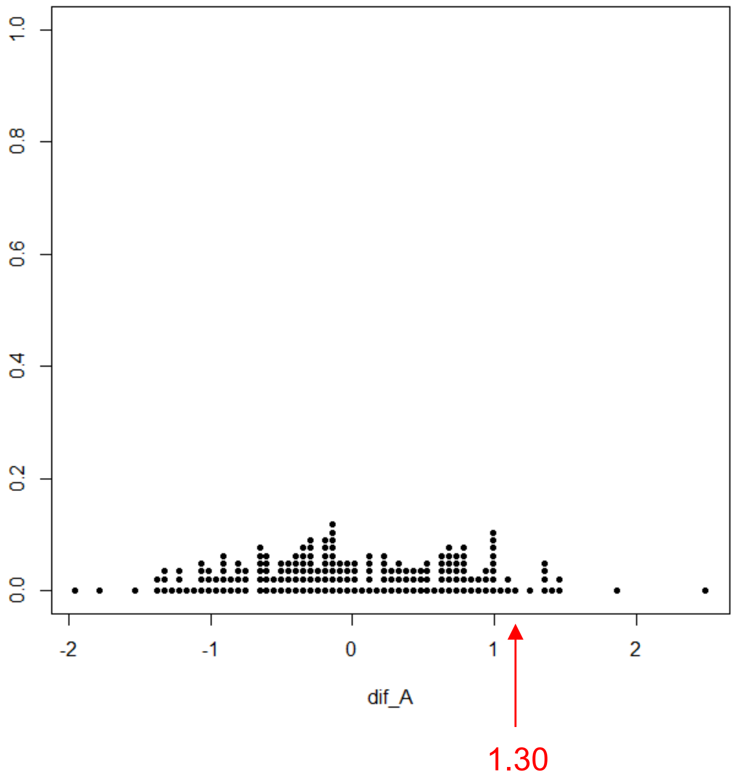
We count how many times we have observed a difference of 1.30 or larger

```
sum(dif_A>=1.30)/length(dif_A)
0.04712042
```

Note: random sampling not assumed.



Consecutive runs, are they independent?.



Comparison of two means

Without historical data

We substitute historical data by assumptions about the generating mechanism:

1. Each sample is a random sample.

$$x_A \sim Any\left(\mu_A, \sigma_A^2\right) \qquad x_B \sim Any\left(\mu_B, \sigma_B^2\right)$$

CLT

$$\bar{x}_A \sim N\left(\mu_A, \sigma_A^2/n_a\right) \qquad \bar{x}_B \sim N\left(\mu_B, \sigma_B^2/n_b\right)$$

2. Both samples are independent

$$\bar{x}_B - \bar{x}_A \sim N\left(\mu_A - \mu_A, \sigma_A^2/n_a + \sigma_B^2/n_b\right)$$

3. If the variances of both processes are equal: $\longrightarrow$ $\bar{x}_B - \bar{x}_A \sim N\left(\mu_A - \mu_A, \sigma^2\left(\frac{1}{n_a} + \frac{1}{n_b}\right)\right)$

Then, under H_0, that is, if there are no improvement in the *B* method:

$$H_0: \quad \mu_B - \mu_A = 0$$

$$\boxed{\bar{x}_B - \bar{x}_A \sim N\left(0, \sigma^2\left(\frac{1}{n_a} + \frac{1}{n_b}\right)\right)}$$

The z-test and the t-test

If we know the common variance σ^2 of the generating mechanism

z-test

$$z = \frac{\bar{x}_B - \bar{x}_A}{\sigma\sqrt{\frac{1}{n_a} + \frac{1}{n_b}}} \sim N(0,1)$$

```
sigma_com
2.880124
```

=sd of the historical data

```
z <- (mitj_B-mitj_A)/(sigma_com*sqrt((1/nA)+(1/nB)))
pnorm(z,lower.tail=F)
0.1564171
```

If we don't know the common variance σ^2

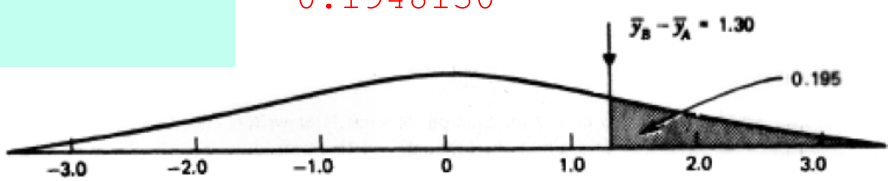
t-test

$$s^2_{pool} = \frac{(n_a - 1)s_A^2 + (n_b - 1)s_B^2}{n_a + n_b - 2}$$

```
s_pool <- sqrt(((nA-1)*var_A+(nB-1)*var_B)/(nA+nB-2))
s_pool
3.297373
```

$$t = \frac{\bar{x}_B - \bar{x}_A}{s_{pool}\sqrt{\frac{1}{n_a} + \frac{1}{n_b}}} \sim t_{df = n_a + n_b - 2}$$

```
t <- (mitj_B - mitj_A)/(s_pool*sqrt((1/nA)+(1/nB)))
pt(t, df=(nA+nB-2), lower.tail=F)
0.1948130
```



Is the t-test a correct solution for this problem?

t-test for comparing two means

Method A	Method B	$n_A = 10$	$n_B = 10$
89.7	84.7	Sum $= 842.4$	Sum $= 855.4$
81.4	86.1	Average $\bar{y}_A = 84.24$	Average $\bar{y}_B = 85.54$
84.5	83.3		
84.8	91.9		
87.3	86.3		
79.7	79.3		
85.1	82.6		
81.7	89.1		
83.7	83.7		
84.5	88.5		

Difference $\bar{y}_B - \bar{y}_A = 1.30$

$$\sum y_A^2 - \left(\sum y_A\right)^2/n_A = 75.784 \qquad \sum y_B^2 - \left(\sum y_B\right)^2/n_B = 119.924$$

Pooled estimate of σ^2: $s^2 = \dfrac{75.784 + 119.924}{10 + 10 - 2} = \dfrac{195.708}{18} = 10.8727$

with $\nu = 18$ degrees of freedom

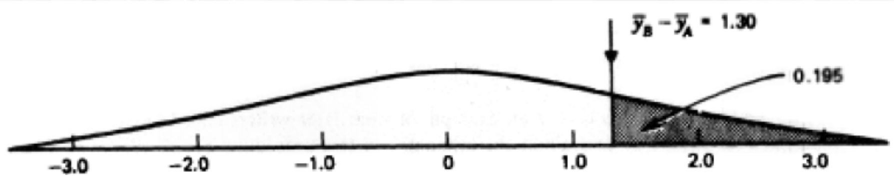
Estimated variance of $\bar{y}_B - \bar{y}_A$: $s^2\left(\dfrac{1}{n_A} + \dfrac{1}{n_B}\right) = \dfrac{2s^2}{10}$

Estimated standard error of $\bar{y}_B - \bar{y}_A$: $\sqrt{\dfrac{s^2}{5}} = \sqrt{\dfrac{10.8727}{5}} = 1.47$

$$t_0 = \frac{(\bar{y}_B - \bar{y}_A) - \delta_0}{s\sqrt{1/n_B + 1/n_A}}$$

For $\delta_0 = 0$, $t_0 = \dfrac{1.30}{1.47} = 0.88$ with $\nu = 18$ degrees of freedom

$$\Pr(t \geq 0.88) = 19.5\%$$



Is the *t*-test a correct solution for this problem?

The alternative: permutation test

What if we don’t have historical data and we don’t want to make probabilistic assumptions about the data generating mechanism

Data:

89.7 81.4 84.5 84.8 87.3 79.7 85.1 81.7 83.7 84.5 84.7 86.1 83.2 91.9 86.3 79.3 82.6 89.1 83.7 88.5

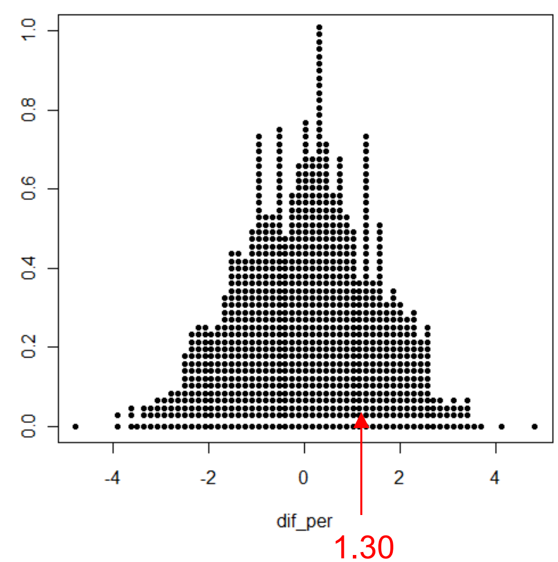
Method:

A A A A A A A A A A B B B B B B B B B B

Under the Null Hypothesis, we assume that both methods have been generated with the same mechanism, so *“method labels*” are just random labels among all the possible permutations, of 10 “A” and 10 “B”.

```
for (i in 1:1000)
      {rnd <- sample(1:20,10)
       dif_per[i] = mean(x[rnd])-mean(x[-rnd])}

sum(dif_per>=1.30)/length(dif_per)
0.209
```



We can assess how likely is the actual difference of 1.30 if both labels are asssigned at random

4.4 Dealing with proportions

Data sometimes occur as the proportions of times a certain event happens.
i.e. proportion of times that a server lasts more than 30sec. to answer a query, Proportion of e-buyers in visitors, proportion of mutations in a genomic sequence, churning proportion of clients,

Let call p the proportion of interest.
To compute it, we need a random sample of n events, where we have observed x successes.

The theoretical model: Binomial process.
n independent events, each with two possibilities, red or black, being p the constant probability of obtaining a red outcome (and hence $1\text{-}p$ being the probability of a black outcome)

●●●●●●●●●●●●●●●●●●●●●●●●●●●●●●●●●●●

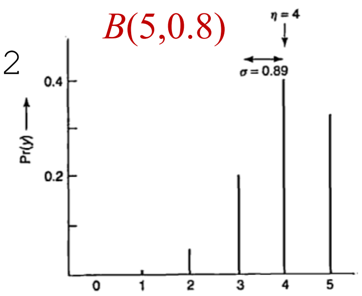
$$B(n, p)$$

How can I estimate the proportion p in such process

Let x be the number of “red” outcomes obtained out of n events

$$\hat{p} = \frac{x}{n}$$

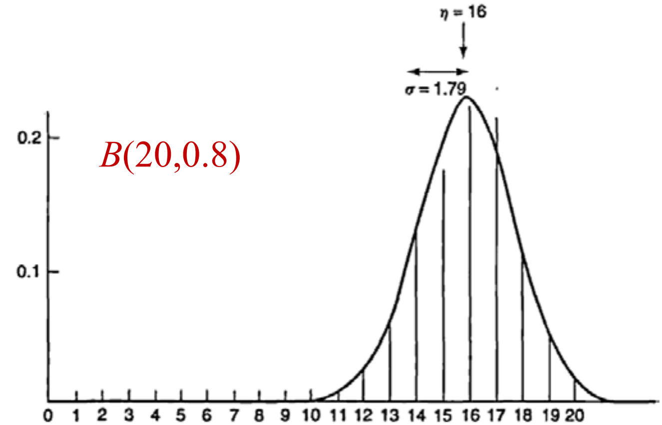
Tendency to the normal distribution



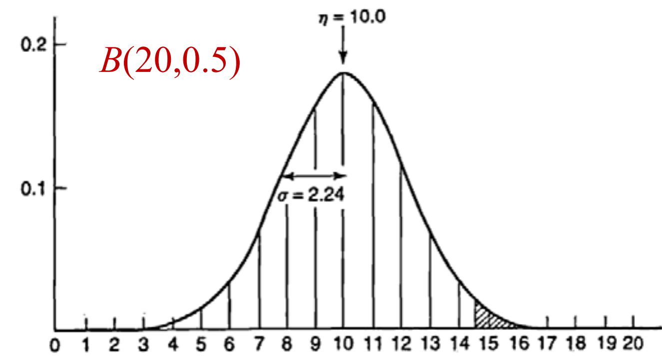
Probability of having 0 “reds” in 5 events: `0.2*0.2*0.2*0.2*0.2 = 0.00032`

```
x = c(0,1,2,3,4,5)
dbinom(x,size=5,prob=0.8)
[1] 0.00032 0.00640 0.05120 0.20480 0.40960 0.32768

sum(x*dbinom(x,size=5,prob=0.8)) # MEAN
[1] 4     = 5*0.8
sum((x-4)^2*dbinom(x,size=5,prob=0.8)) # VARIANCE
[1] 0.8   = 5*0.8*(1-0.8)
```




```
x=0:20
sum(x*dbinom(x,size=20,prob=0.8))
[1] 16   = 20*0.8
> sum((x-16)^2*dbinom(x,size=20,prob=0.8))
[1] 3.2  = 20*0.8*(1-0.8)
```




```
sum(x*dbinom(x,size=20,prob=0.5))
[1] 10   = 20*0.5
sum((x-10)^2*dbinom(x,size=20,prob=0.5))
[1] 5    = 20*0.5*(1-0.5)
```

- Closer to 0.5 more normal is the distribution
- Bigger is n, closer to the normal is the distribution

Distribution of a proportion

$$x \sim B(n,p) \quad \rightarrow \quad N(\mu,\sigma^2)$$

$$\mu = E[x] = n \times p$$

$$\sigma^2 = n \times p \times (1-p)$$

$$\hat{p} = \frac{x}{n} \sim \quad \rightarrow \quad N\left(\mu/n, \sigma^2/n^2\right) = N\left(p, \frac{p \times (1-p)}{n}\right)$$

$$E[\hat{p}] = p$$

$$\operatorname{var}[\hat{p}] = \frac{p \times (1-p)}{n}$$

$$p = 0.1297$$

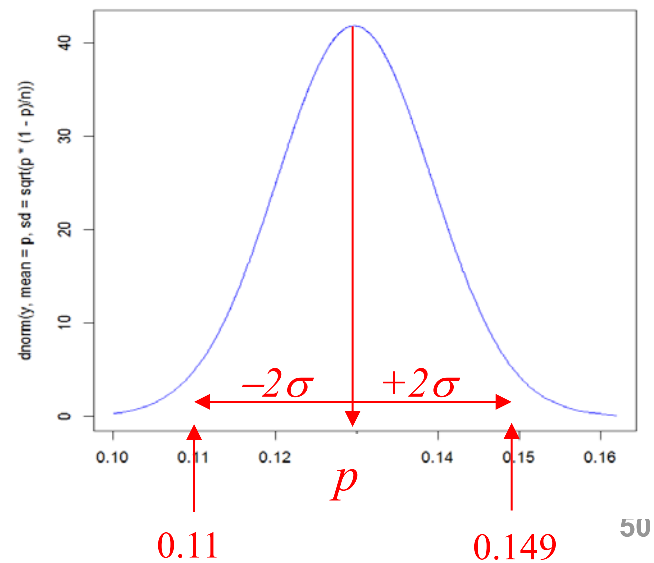
$$n = 1240$$

$$\hat{p} \sim N\left(0.1297, \frac{0.1297 \times (1-0.1297)}{1240}\right)$$

$$N\left(0.1297, \frac{0.1297 \times (1-0.1297)}{1240}\right) = N\left(0.1297, 0.00954^2\right) =$$

$$2 \times 0.00954 = 0.019$$

$$p + 0.019 = 0.149; \quad p - 0.019 = 0.11$$



How reliable is a proportion

Centre d'Estudis d'Opinió

REO núm. 763

Data 30 de gener de 2015

☐ Personal ☒ Telefònica CATI ☐ Internet / on-line

	Escolliu opció:	*Observacions:*
A.1.1) **Durada del qüestionari:**	de 16 a 30 minuts	
A.1.2) **Grandària mostra:**	De 1501 a 2500 entrevistes	n=1600 Barcelona= 603 Girona= 349 Lleida= 300 Tarragona=348 Ponderació: Sí Error = ± 3,09 %
A.1.3) **Àmbit geogràfic:**	Catalunya	
A.1.4) **Univers a entrevistar:**	Població general	Població amb ciutadania espanyola de 18 i més anys resident a Catalunya
A.1.5) **Tipus de mostreig**	Quotes	Estratificat per província i dimensió de municipi amb quotes encreuades de sexe, edat i lloc de naixement.

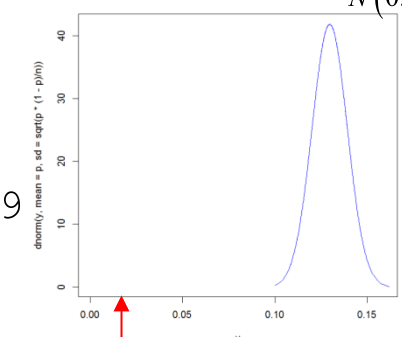
Total Mostra real	**1600**
17b. I em podria dir a quin partit o coalició va votar en les darreres eleccions al Parlament de Catalunya?	
B: Segur/a va votar M Real	**1240**
PPC	**29**

```
p_hat=29/1240 = 0.0233871

sqrt(p_true*(1-p_true)/n) = 0.00954
pnorm(p_hat,mean=p_true,sd=0.00954) = 3.88e-29
```

$$z = \frac{0.02339 - 0.1297}{0.00954} = -11.15$$

$N\left(0.1297, 0.00954^2\right)$



0.0234

Do you think that respondents have enough good memory?

What type of problem is this?

UNIVERSITAT POLITÈCNICA DE CATALUNYA
School of Professional & Executive Development

Finding the significant words of a document

Miguel Hernandez poems	AMOR	CORAZON	HUERTO	MUERTE	SANGRE	LUZ	HOMBRE	Total
La Morada	41	3	32	21	8	52	5	162
Perito en Lunas	4	1	3	3	1	12	0	186
Oda a la Higuera	37	6	11	27	14	35	6	160
Rayo que no cesa	17	26	0	8	12	1	1	201
Mi sangre es un camino	7	16	0	9	26	1	2	126
Vientos del pueblo	3	23	2	61	35	3	22	210
Romancero de ausencias	44	20	2	38	25	19	19	316
Hijo de la luz y de la sombra	14	11	2	15	13	25	8	255
Total	167	106	52	182	134	148	63	940

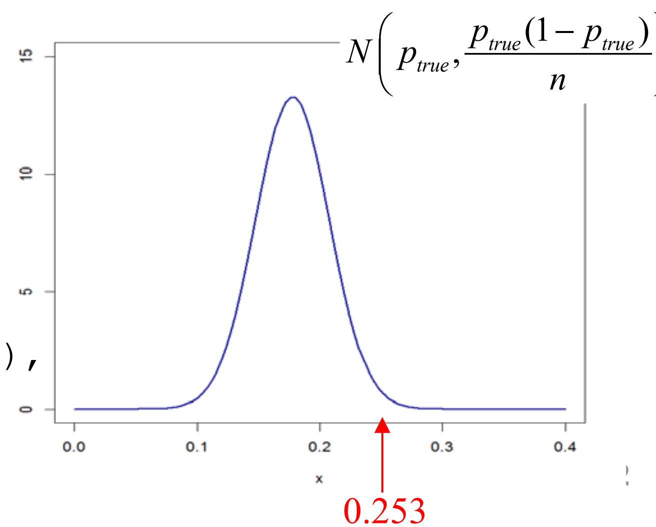
How characteristic is a word for a poem, i.e. AMOR for *La Morada*

```
p_true = 167/940 = 0.1776596

x = 41
n = 162
p_hat = x/n = 0.2530864

pnorm(p_hat,mean=p_true,sd=sqrt(p_true*(1-p_true)/n),
lower.tail=F) = 0.006007997

sqrt(p_true*(1-p_true)/n) = 0.0300305
```



References

- Introducción a la estadística. Thomas H. Wonnacott, Ronald J. Wonnacott. LIMUSA.
- Estadística para investigadores : diseño, innovación y descubrimiento. George E. P. Box, J. Stuart Hunter, William Gordon Hunter. REVERTE, 2008.
- Introductory Statistics with R. Peter Dalgaard. Springer 2008.