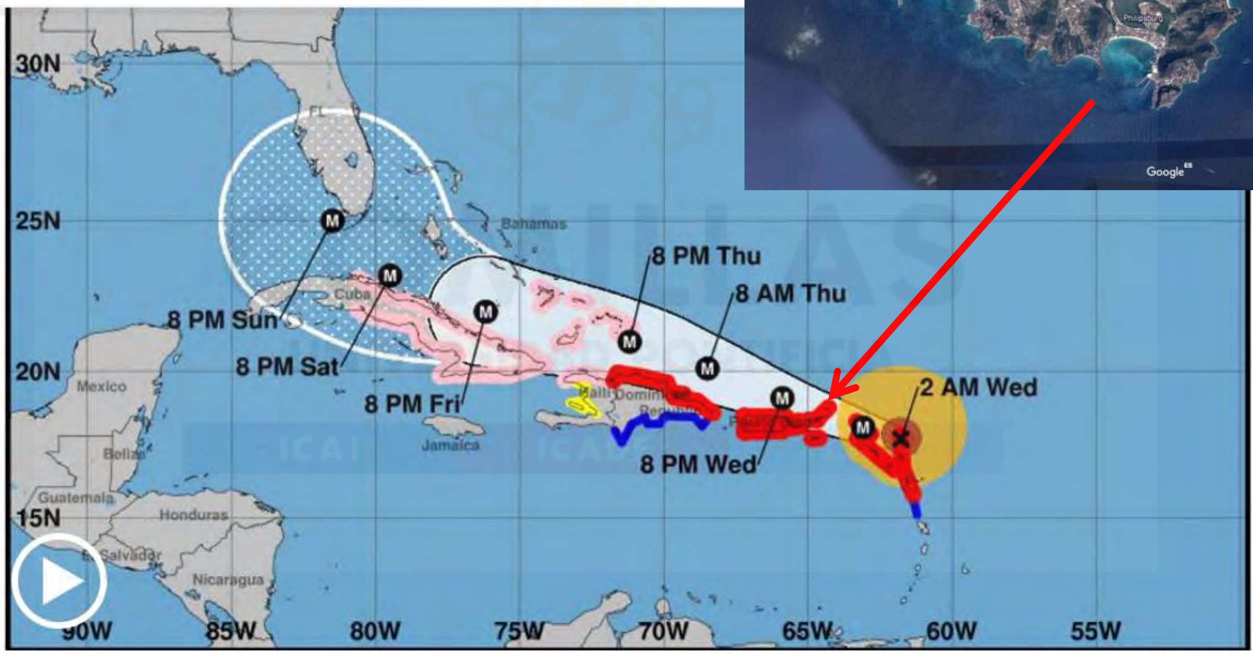
Fundamental Concepts in Statistics

Prof. Eugenio Sánchez Úbeda

January 2024

Irma

- Forecasted trajectory
  - What is represented?



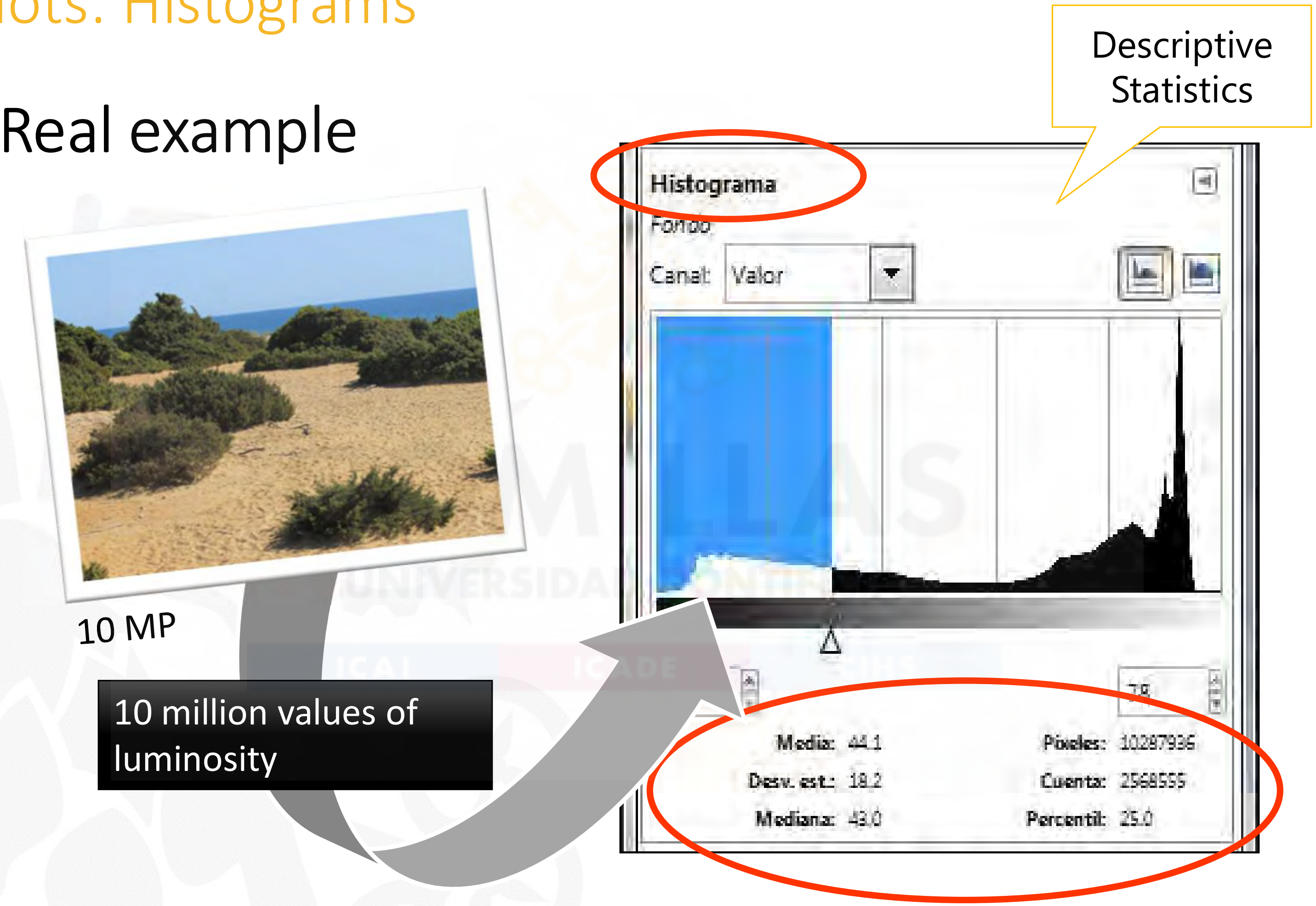
Trayectoria del huracán Irma. CENTRO NACIONAL DE HURACANES / VÍDEO: ATLAS

1

Descriptive Statistics

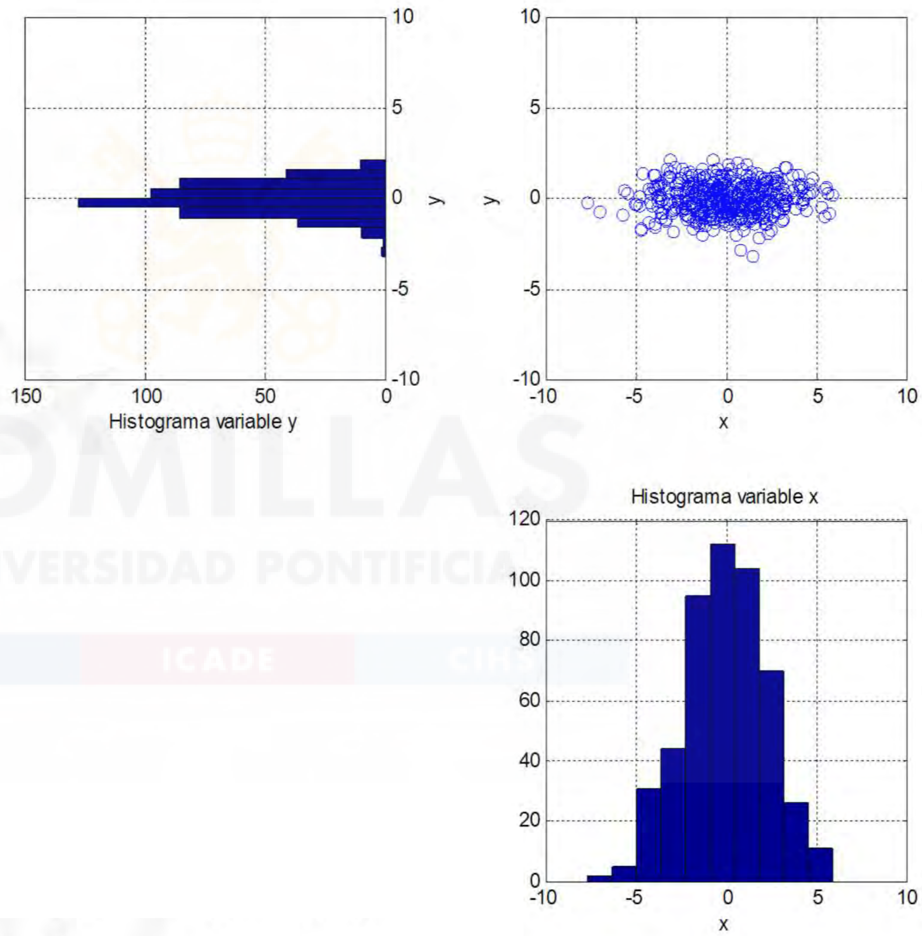
Plots: Histograms

- Real example



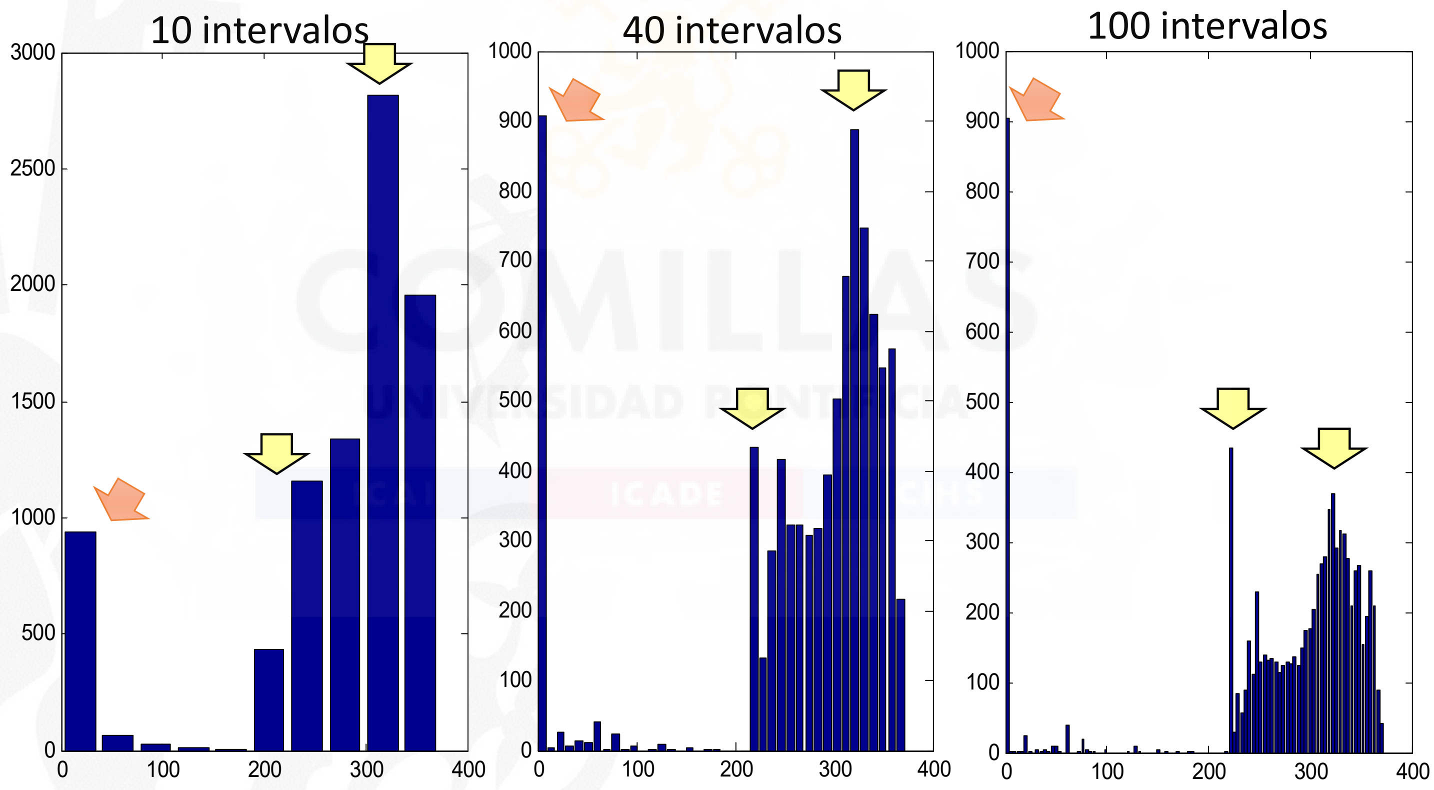
Descriptive Statistics
Plots: Histograms

- Synthetic example



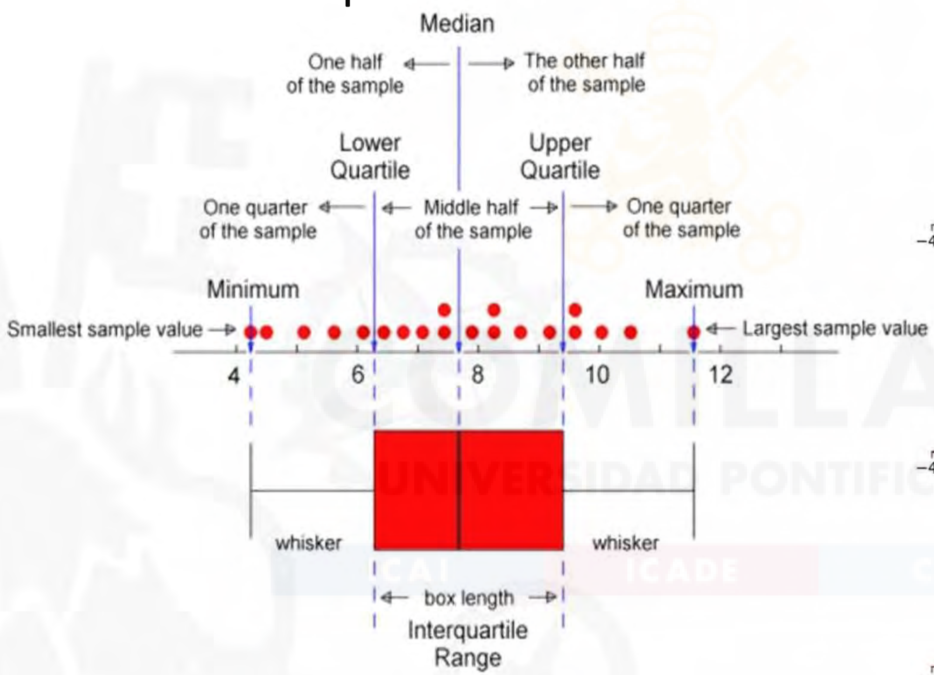
Descriptive Statistics
Plots: Histograms

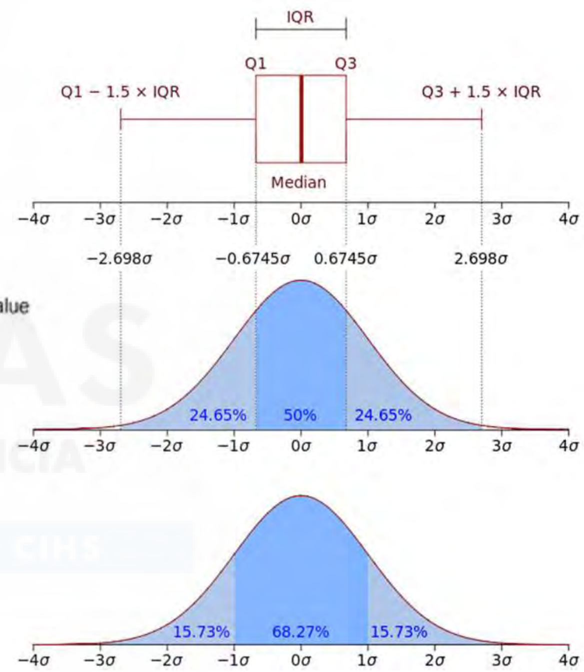
- The number of intervals may alter the perception of the distribution
- Example (Hourly production of ACE4 power plant, year 2011)



Descriptive Statistics
Plots: Box plot

- Classic box plot





- Ends of the whiskers can be:
  - 1.5 times the interquartile range (IQR) (there are outliers)
  - Max – Q3

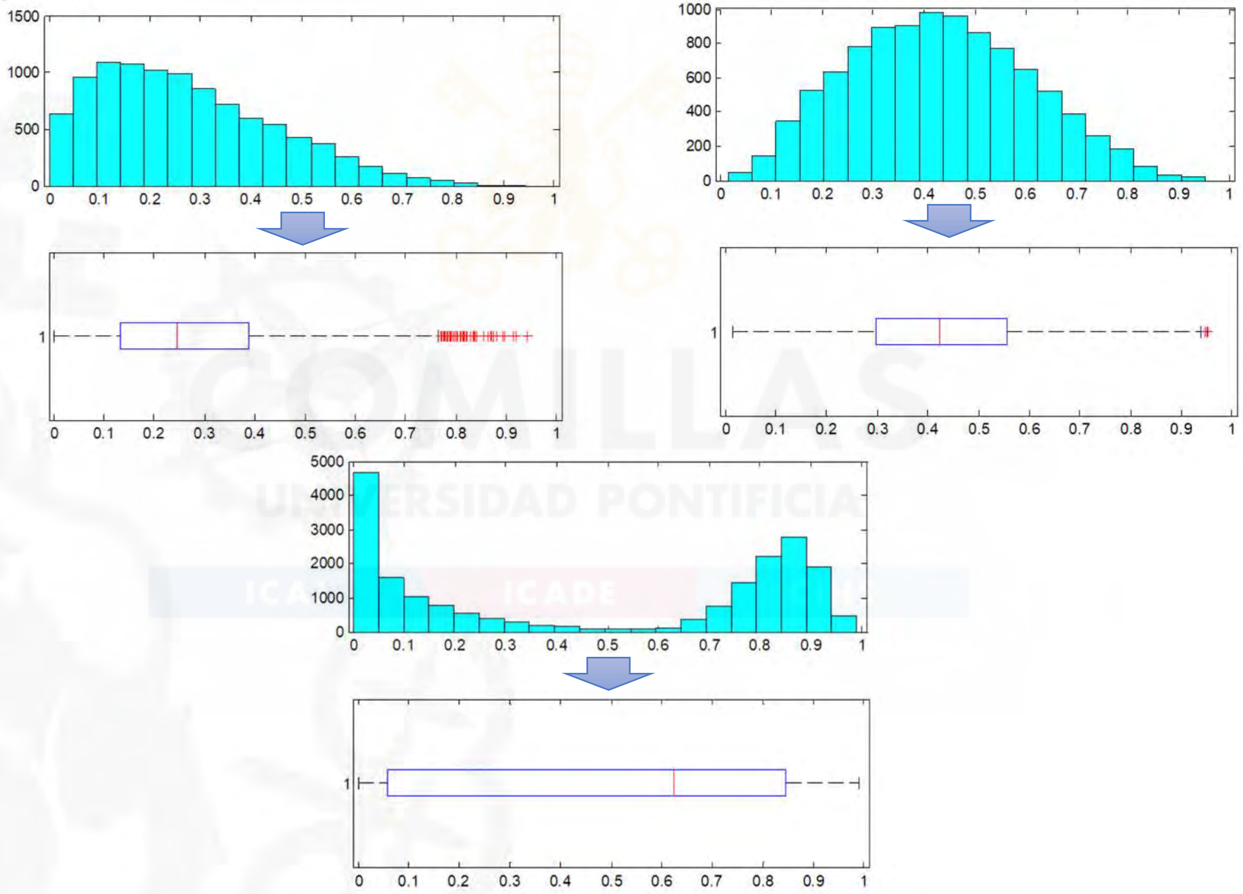
  (The lower one is symmetric)

[Source: Wikipedia]

Descriptive Statistics

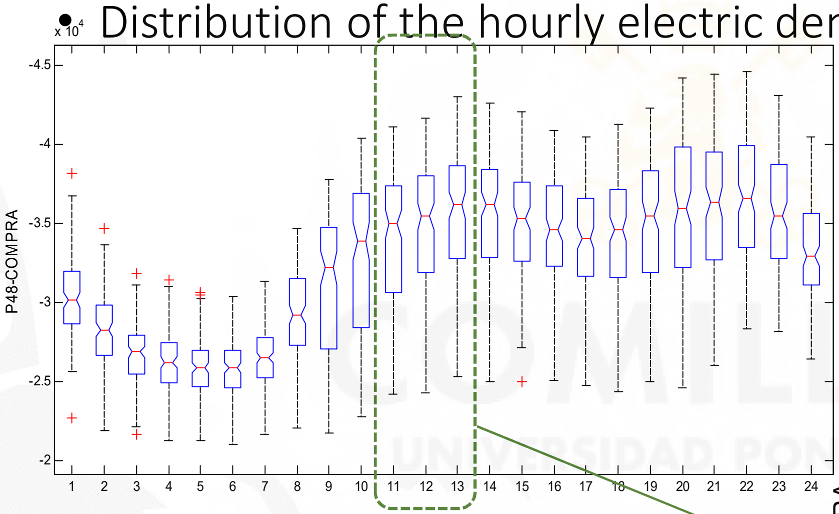
Plots: Box plot

- Examples

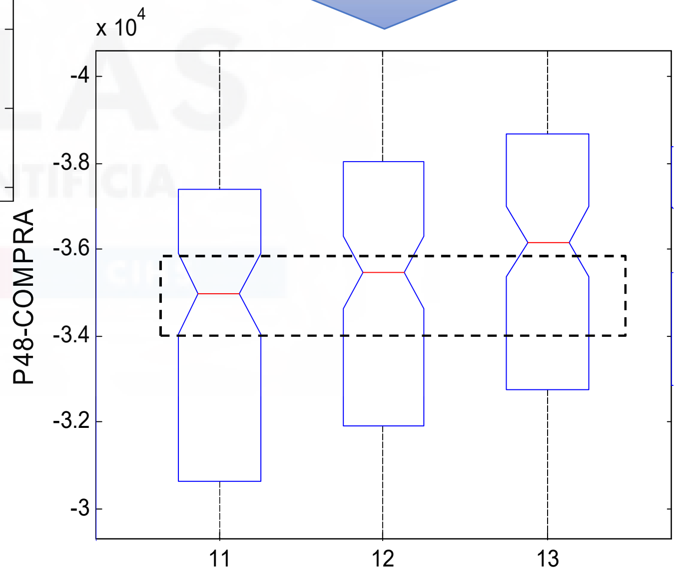


Descriptive Statistics
Plots: Box plot

- Example with notches:
  - Distribution of the hourly electric demand for each hour (2012)

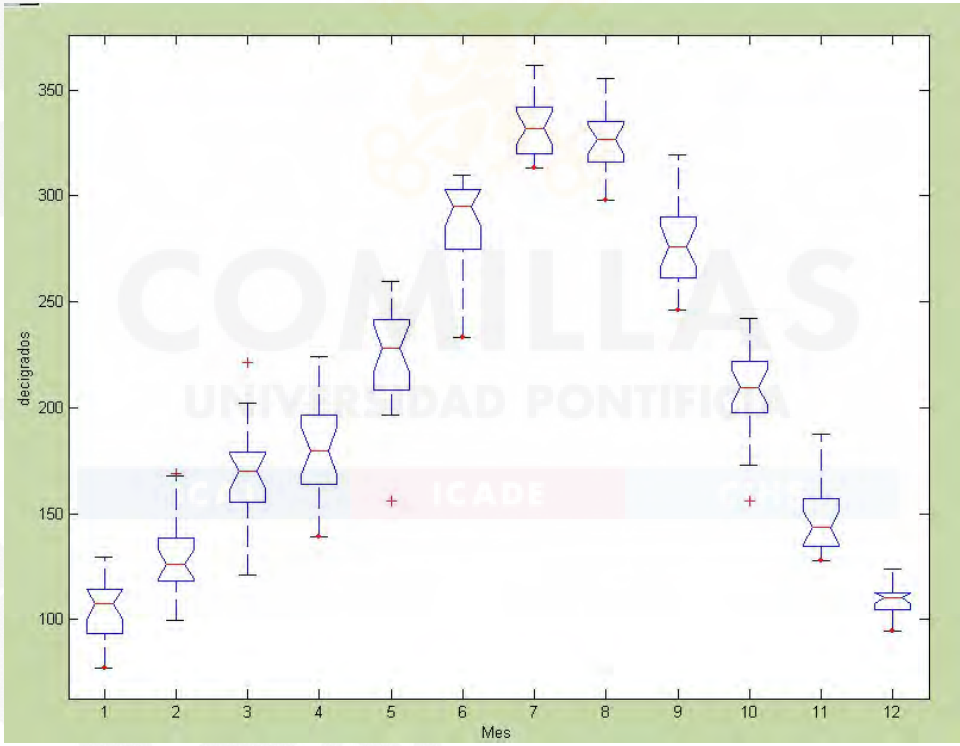


- Notch of box H11 shows that there are no significant differences between demand medians of H11 and H12



Descriptive Statistics
Plots: Box plot

- Example with notches:
  - Monthly distribution of maximum daily temperature in Madrid



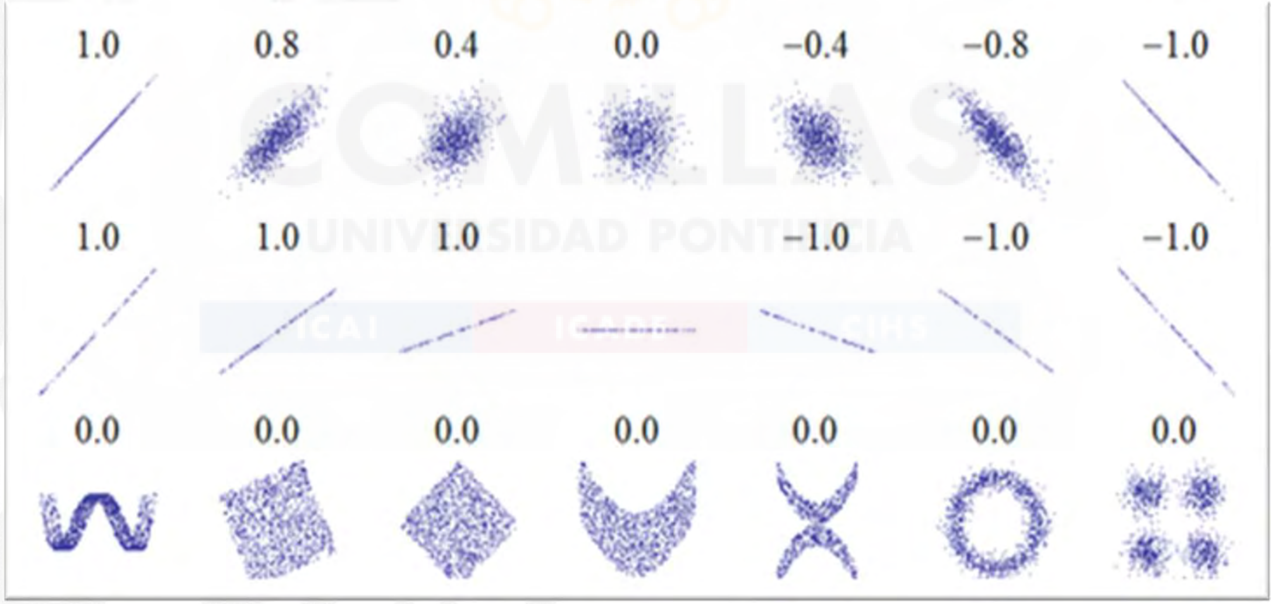
Descriptive Statistics

Pearson correlation coefficient

- Pearson linear correlation coefficient between two variables

$$\rho_{X,Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y},$$

- Takes values between -1 and 1



[Source: Wikipedia]

Pearson correlation coefficient

- If $r = 1$, there exist a perfect positive correlation. When a variable increases, the other one also does in a constant proportion.
- If $0 < r < 1$, there exist a positive correlation.
- If $r = 0$, there is no linear correlation, but there can be other nonlinear relations between both variables.
- If $-1 < r < 0$, there exist a negative correlation.
- If $r = -1$, there exist a perfect negative correlation. When a variable increases, the other one decreases in a constant proportion. It is called an inverse relation

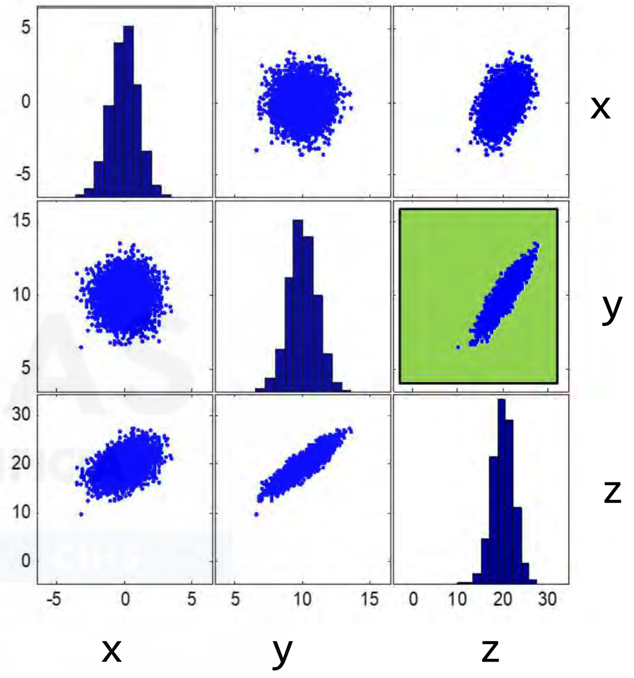
- If $r = 1$, there exist a perfect positive correlation. When a variable increases, the other one also does in a constant proportion.
- If $0 < r < 1$, there exist a positive correlation.
- If $r = 0$, there is no linear correlation, but there can be other nonlinear relations between both variables.
- If $-1 < r < 0$, there exist a negative correlation.
- If $r = -1$, there exist a perfect negative correlation. When a variable increases, the other one decreases in a constant proportion. It is called an inverse relation

Pearson correlation coefficient

- Example

```
>> cov([x y z]) % covariance matrix
    1.0041    0.0010    1.0062
    0.0010    1.0155    2.0321
    1.0062    2.0321    5.0704

>> corr([x y z]) % correlation matrix
    1.0000    0.0010    0.4459
    0.0010    1.0000    0.8955
    0.4459    0.8955    1.0000
```



2

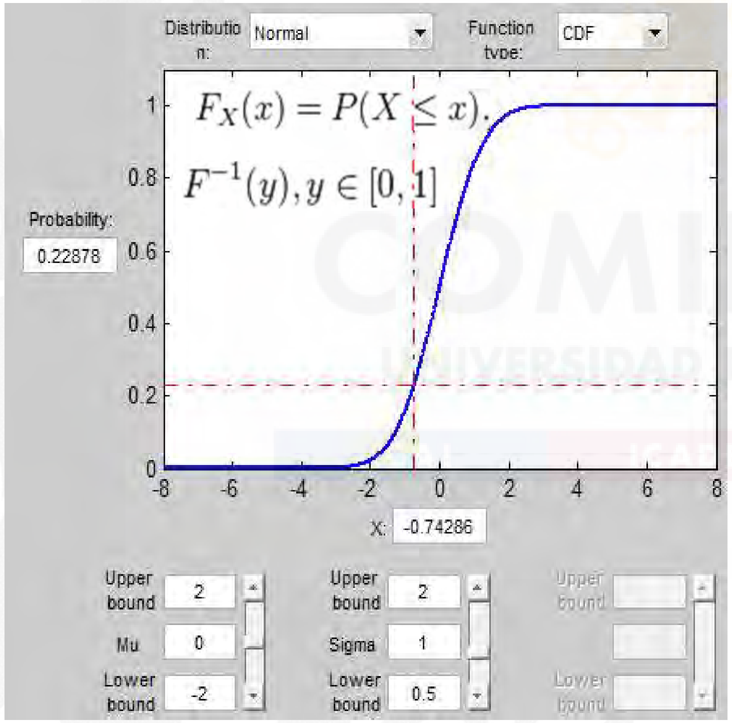
Probability Distributions

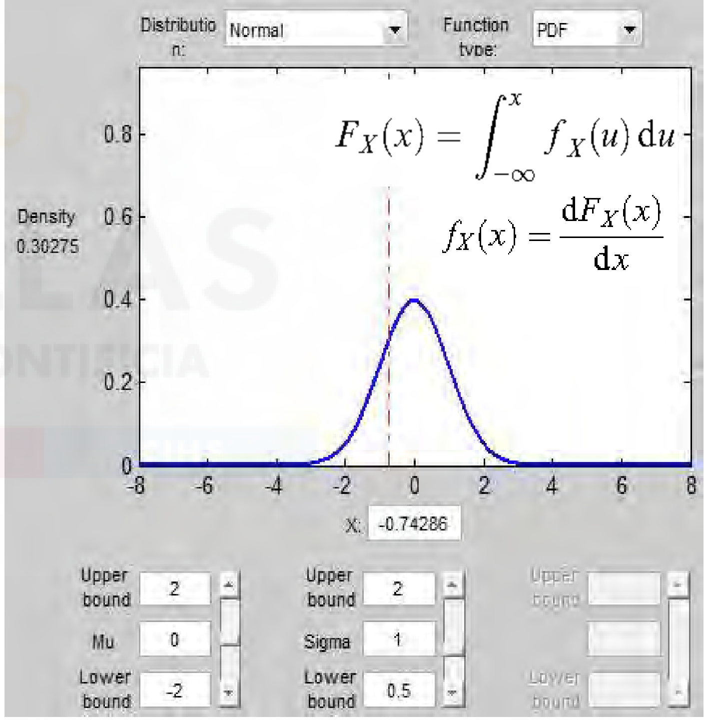
Probability
Random variable

- https://seeing-theory.brown.edu/probability-distributions/index.html#section1

Probability distributions

- Cumulative distribution function (CDF)
- Probability density function (PDF)
- Inverse cumulative distribution or quantile (ICDF)





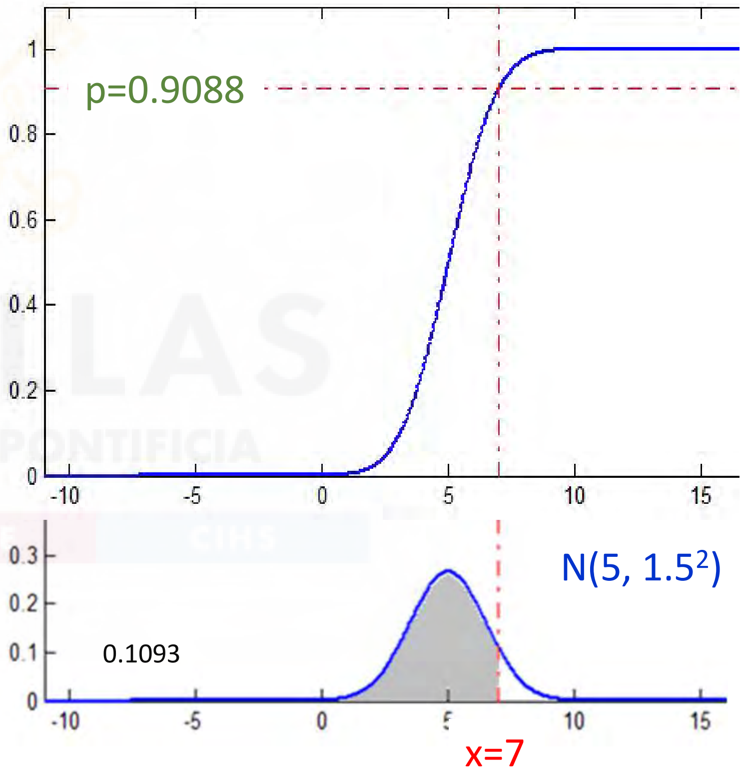
Probability distributions

- Probability distributions in Matlab (Statistics Toolbox)

```
% CDF
p = normcdf(7, 5, 1.5);
```

```
% ICDF
x = norminv(0.9088, 5, 1.5);
```

```
% PDF
% (useful with discrete vars)
d = normpdf(7, 5, 1.5);
```



Probability

Discrete distributions: summary

	$F_X(x)$	$f_X(x)$	$\mathbb{E}[X]$	$\mathbb{V}[X]$	$M_X(s)$
Uniform$\{a,\ldots,b\}$	$\begin{cases} 0 & x < a \\ \frac{\lfloor x \rfloor - a + 1}{b-a} & a \le x \le b \\ 1 & x > b \end{cases}$	$\frac{I(a < x < b)}{b-a+1}$	$\frac{a+b}{2}$	$\frac{(b-a+1)^2-1}{12}$	$\frac{e^{as}-e^{-(b+1)s}}{s(b-a)}$
Bernoulli(p)	$(1-p)^{1-x}$	$p^x(1-p)^{1-x}$	p	$p(1-p)$	$1-p+pe^s$
Binomial(n,p)	$I_{1-p}(n-x, x+1)$	$\binom{n}{x}p^x(1-p)^{n-x}$	np	$np(1-p)$	$(1-p+pe^s)^n$
Multinomial(n,p)		$\frac{n!}{x_1!\ldots x_k!}p_1^{x_1}\cdots p_k^{x_k} \quad \sum_{i=1}^{k} x_i = n$	np_i	$np_i(1-p_i)$	$\left(\sum_{i=0}^{k} p_i e^{s_i}\right)^n$
Hypergeometric(N,m,n)	$\approx \Phi\left(\frac{x-np}{\sqrt{np(1-p)}}\right)$	$\frac{\binom{m}{x}\binom{m-x}{n-x}}{\binom{N}{x}}$	$\frac{nm}{N}$	$\frac{nm(N-n)(N-m)}{N^2(N-1)}$	N/A
NegativeBinomial(r,p)	$I_p(r, x+1)$	$\binom{x+r-1}{r-1}p^r(1-p)^x$	$r\frac{1-p}{p}$	$r\frac{1-p}{p^2}$	$\left(\frac{p}{1-(1-p)e^s}\right)^r$
Geometric(p)	$1-(1-p)^x \quad x \in \mathbb{N}^+$	$p(1-p)^{x-1} \quad x \in \mathbb{N}^+$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{p}{1-(1-p)e^s}$
Poisson(λ)	$e^{-\lambda}\sum_{i=0}^{x}\frac{\lambda^i}{i!}$	$\frac{\lambda^x e^{-\lambda}}{x!}$	λ	λ	$e^{\lambda(e^s-1)}$

Continuous distributions: summary

	$F_X(x)$	$f_X(x)$	$\mathbb{E}[X]$	$\mathbb{V}[X]$	$M_X(s)$
$\text{Uniform}(a,b)$	$\begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a < x < b \\ 1 & x > b \end{cases}$	$\frac{I(a < x < b)}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{sb}-e^{sa}}{s(b-a)}$
$\text{Normal}(\mu,\sigma^2)$	$\Phi(x) = \int_{-\infty}^{x} \phi(t)\,dt$	$\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$	μ	σ^2	$\exp\left\{\mu s + \frac{\sigma^2 s^2}{2}\right\}$
$\text{Log-Normal}(\mu,\sigma^2)$	$\frac{1}{2} + \frac{1}{2}\text{erf}\left[\frac{\ln x - \mu}{\sqrt{2\sigma^2}}\right]$	$\frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(\ln x - \mu)^2}{2\sigma^2}\right\}$	$e^{\mu+\sigma^2/2}$	$(e^{\sigma^2}-1)e^{2\mu+\sigma^2}$	
$\text{Multivariate Normal}(\mu,\Sigma)$		$(2\pi)^{-k/2}\lvert\Sigma\rvert^{-1/2}e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$	μ	Σ	$\exp\left\{\mu^T s + \frac{1}{2}s^T\Sigma s\right\}$
$\text{Chi-square}(k)$	$\frac{1}{\Gamma(k/2)}\gamma\left(\frac{k}{2},\frac{x}{2}\right)$	$\frac{1}{2^{k/2}\Gamma(k/2)}x^{k/2}e^{-x/2}$	k	$2k$	$(1-2s)^{-k/2}\ s < 1/2$
$\text{Exponential}(\beta)$	$1-e^{-x/\beta}$	$\frac{1}{\beta}e^{-x/\beta}$	β	β^2	$\frac{1}{1-\beta s}\ (s < 1/\beta)$
$\text{Gamma}(\alpha,\beta)$[1]	$\frac{\gamma(\alpha,x/\beta)}{\Gamma(\alpha)}$	$\frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}e^{-x/\beta}$	$\alpha\beta$	$\alpha\beta^2$	$\left(\frac{1}{1-\beta s}\right)^\alpha\ (s < 1/\beta)$
$\text{InverseGamma}(\alpha,\beta)$	$\frac{\Gamma\left(\alpha,\frac{\beta}{x}\right)}{\Gamma(\alpha)}$	$\frac{\beta^\alpha}{\Gamma(\alpha)}x^{-\alpha-1}e^{-\beta/x}$	$\frac{\beta}{\alpha-1}\ \alpha > 1$	$\frac{\beta^2}{(\alpha-1)^2(\alpha-2)^2}\ \alpha > 2$	$\frac{2(-\beta s)^{\alpha/2}}{\Gamma(\alpha)}K_\alpha\left(\sqrt{-4\beta s}\right)$
$\text{Dirichlet}(\alpha)$		$\frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)}\prod_{i=1}^k x_i^{\alpha_i-1}$	$\frac{\alpha_i}{\sum_{i=1}^k \alpha_i}$	$\frac{\mathbb{E}[X_i](1-\mathbb{E}[X_i])}{\sum_{i=1}^k \alpha_i + 1}$	
$\text{Beta}(\alpha,\beta)$[2]	$I_x(\alpha,\beta)$	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	$1+\sum_{k=1}^{\infty}\left(\prod_{r=0}^{k-1}\frac{\alpha+r}{\alpha+\beta+r}\right)\frac{s^k}{k!}$
$\text{Weibull}(\lambda,k)$	$1-e^{-(x/\lambda)^k}$	$\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}e^{-(x/\lambda)^k}$	$\lambda\Gamma\left(1+\frac{1}{k}\right)$	$\lambda^2\Gamma\left(1+\frac{2}{k}\right)-\mu^2$	$\sum_{n=0}^{\infty}\frac{s^n\lambda^n}{n!}\Gamma\left(1+\frac{n}{k}\right)$
$\text{Pareto}(x_m,\alpha)$	$1-\left(\frac{x_m}{x}\right)^\alpha\ x \geq x_m$	$\alpha\frac{x_m^\alpha}{x^{\alpha+1}}\ \ x \geq x_m$	$\frac{\alpha x_m}{\alpha-1}\ \alpha > 1$	$\frac{x_m^\alpha}{(\alpha-1)^2(\alpha-2)}\ \alpha > 2$	$\alpha(-x_m s)^\alpha\Gamma(-\alpha,-x_m s)\ s < 0$

Probability distributions in Matlab

Statistics and Machine Learning Toolbox™ supports more than 30 probability distributions, including parametric, nonparametric, continuous, and discrete distributions.

The toolbox provides several ways to work with probability distributions.

- Use *probability distribution objects* to fit a probability distribution object to sample data, or to create a probability distribution object with specified parameter values. The Using Objects page for each distribution provides information about the object's properties and the functions you can use to work with the object.
- Use *probability distribution functions* to work with data input from matrices, tables, and dataset arrays. Some of the supported distributions have distribution-specific functions. These functions use the following abbreviations:
  - pdf — Probability density functions
  - cdf — Cumulative distribution functions
  - inv — Inverse cumulative distribution functions
  - stat — Distribution statistics functions
  - fit — Distribution fitting functions
  - like — Negative log-likelihood functions
  - rnd — Random number generators

  You can also use the following generic functions to work with most of the distributions:
  - `pdf` — Probability density function
  - `cdf` — Cumulative distribution function
  - `icdf` — Inverse cumulative distribution function
  - `mle` — Distribution fitting function
  - `random` — Random number generating function
- Use *probability distribution apps and user interfaces* to interactively fit, explore, and generate random numbers from probability distributions. Available apps and user interfaces include:
  - The Distribution Fitting app, to interactively fit a distribution to sample data, and export a probability distribution object to the workspace.
  - The Probability Distribution Function user interface, to visually explore the effect on the pdf and cdf of changing the distribution parameter values.
  - The Random Number Generation user interface (`randtool`), to interactively generate random numbers from a probability distribution with specified parameter values and export them to the workspace.

Probability

Continuous distributions: Normal

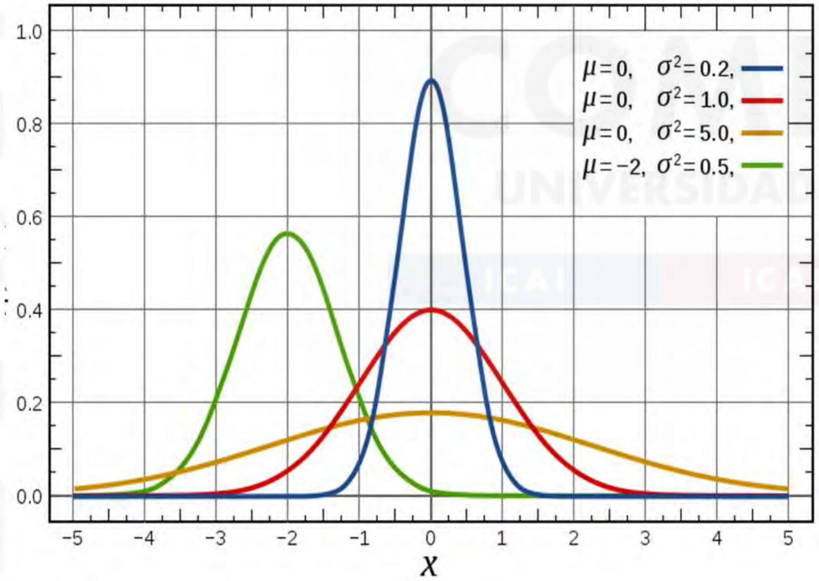
$$X \sim \mathcal{N}(\mu, \sigma^2)$$

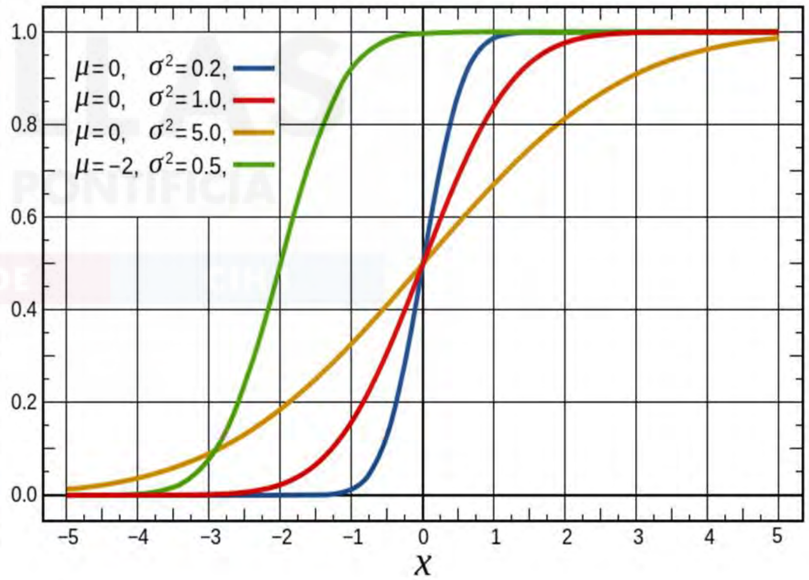
- PDF and CDF

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right).$$

$$F(x;\mu,\sigma^2) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt$$

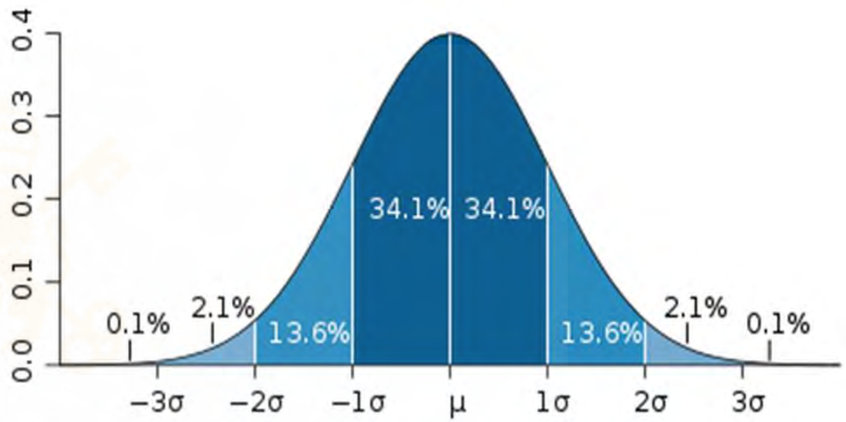


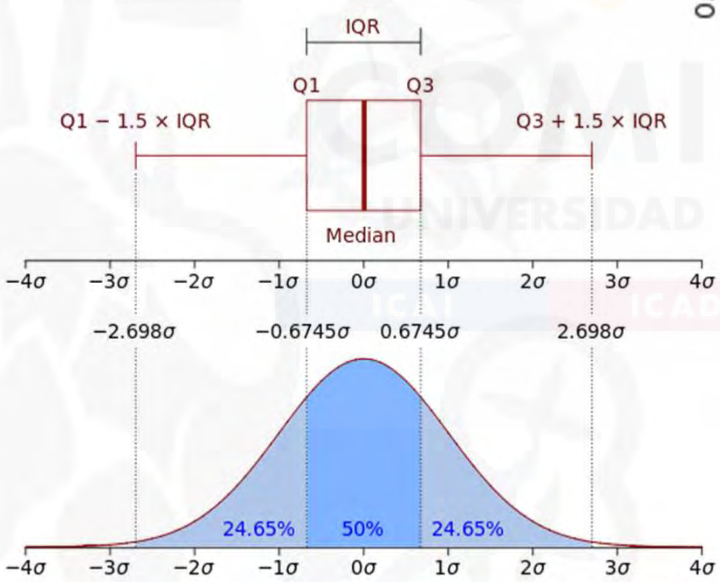


Probability

Continuous distributions: Normal

- 68.2% of the distribution is within the interval of 2 standard deviations





- 95.4% of the distribution is within the interval of 4 standard deviations

[Source: wikipedia]

Mean and variance

- Mean

$$\alpha_n = E\{X^n\} = \int_{-\infty}^{\infty} x^n f_X(x)\mathrm{d}x$$

La esperanza es un operador lineal, ya que:

$$\mathrm{E}(X+c) = \mathrm{E}(X) + c$$
$$\mathrm{E}(X+Y) = \mathrm{E}(X) + \mathrm{E}(Y)$$
$$\mathrm{E}(aX) = a\,\mathrm{E}(X)$$

Combinando estas propiedades, podemos ver que -

$$\mathrm{E}(aX+b) = a\,\mathrm{E}(X) + b$$
$$\mathrm{E}(aX+bY) = a\,\mathrm{E}(X) + b\,\mathrm{E}(Y)$$

- Variance

$$E\{(X-m)^n\} = \int_{-\infty}^{\infty} (x-m)^n f_X(x)\mathrm{d}x$$

- $V(X) \geq 0$
- $V(aX+b) = a^2V(X)$ siendo *a* y *b* números reales cualesquiera. De esta propiedad se deduce que la varianza de una constante es cero, es decir, $V(b) = 0$
- $V(X+Y) = V(X) + V(Y) + 2Cov(X,Y)$, donde Cov(*X*,*Y*) es la covarianza de *X* e *Y*.
- $V(X-Y) = V(X) + V(Y) - 2Cov(X,Y)$, donde Cov(*X*,*Y*) es la covarianza de *X* e *Y*.

Fuente: wikipedia

Probability

Linear combination of variables

- Independent variables (null covariance)

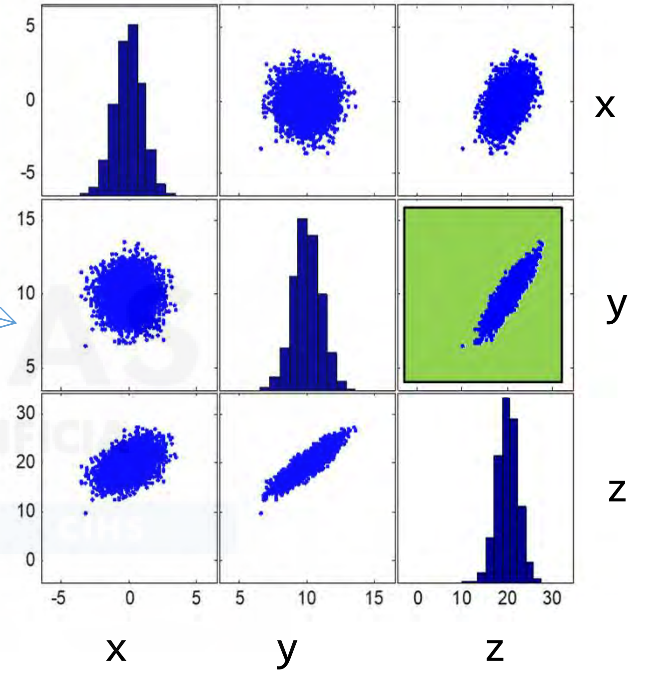
```
x = randn(5000,1);
y = 10 + randn(5000,1);

z= x + 2*y;

figure; plotmatrix([x y z])
```

```
>> cov([x y z]) % covariance matrix
    1.0041    0.0010    1.0062
    0.0010    1.0155    2.0321
    1.0062    2.0321    5.0704

>> corr([x y z]) % correlation matrix
    1.0000    0.0010    0.4459
    0.0010    1.0000    0.8955
    0.4459    0.8955    1.0000
```



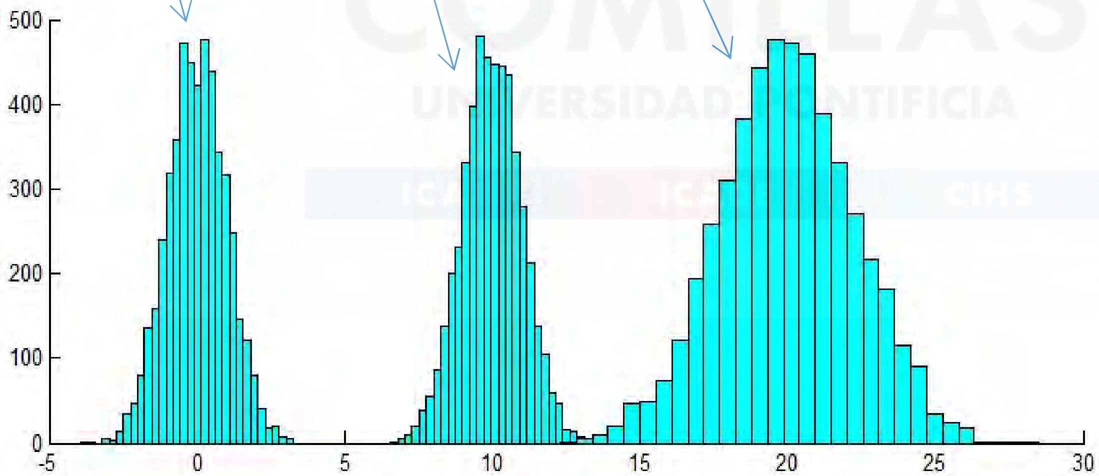
Linear combination of variables

- Independent variables (null covariance)

```
x = randn(5000,1);
y = 10 + randn(5000,1);
z= x + 2*y;

figure; hold on; colormap cool;
hist(x,30);hist(y,30);hist(z,30);
```

$$X \sim N(\mu_X, \sigma_X^2)$$
$$Y \sim N(\mu_Y, \sigma_Y^2)$$
$$Z \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2).$$




```
var(x) -> 1.0041
var(y) -> 1.0155

var(z) -> 5.0704
```

$1 + 2^2 * 1 + 0$

Probability

Linear combination of variables

- Dependent variables (non null covariance)

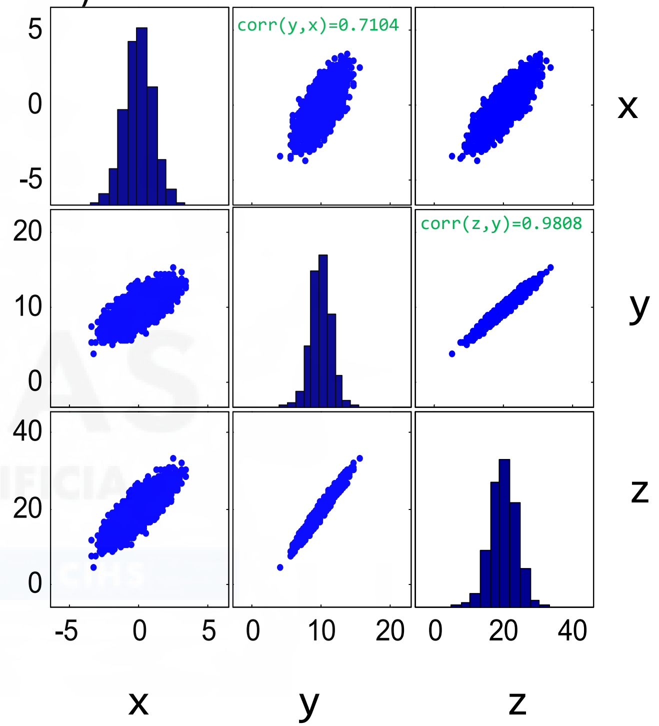
```
x = randn(5000,1);
y = 10 + x + randn(5000,1);

z= x + 2*y;

figure; plotmatrix([x y z])
```

```
>> cov([x y z]) % covariance matrix
    1.0080    1.0117    3.0314
    1.0117    2.0119    5.0355
    3.0314    5.0355   13.1025

>> corr([x y z]) % correlation matrix
    1.0000    0.7104    0.8341
    0.7104    1.0000    0.9808
    0.8341    0.9808    1.0000
```



Probability

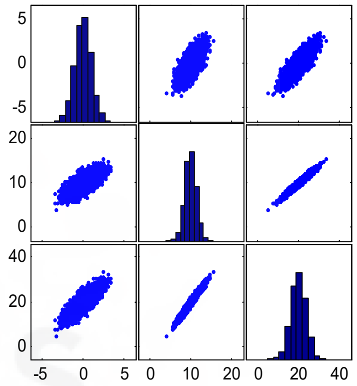
Linear combination of variables

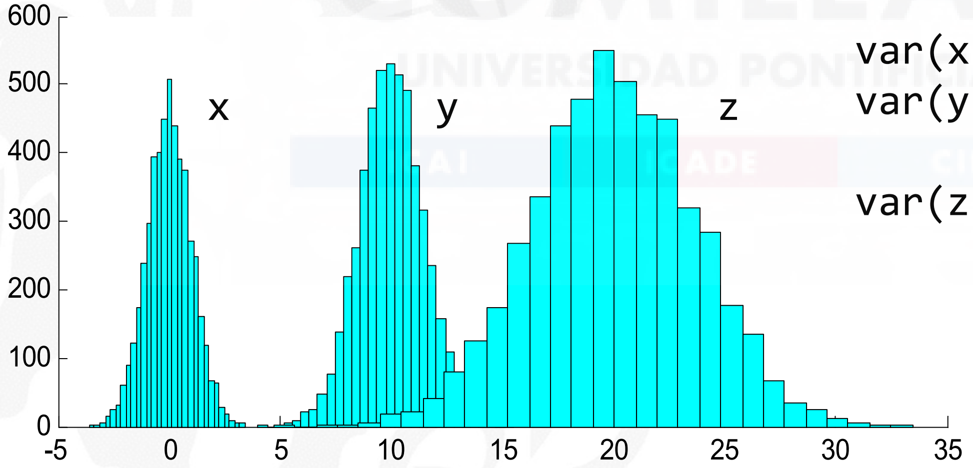
- Dependent variables (non null covariance)

```
x = randn(5000,1);
y = 10 + x + randn(5000,1);
z= x + 2*y;

figure; plotmatrix([x y z]);

figure; hold; colormap cool;
hist(x,30);hist(y,30);hist(z,30);
```





var(x) -> 0.9912   1

var(y) -> 2.0290   1+1

var(z) -> 13.1005 $1+2^2*2+2*cov(x,y)$

Probability

Linear functions of random variables

- Linear combination of independent and normally distributed random variables

$$Z = X + Y \begin{cases} X \sim N(\mu_X, \sigma_X^2) \\ Y \sim N(\mu_Y, \sigma_Y^2) \end{cases}$$

$$Z \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2).$$

- In general

$$\sum_{i=1}^{n} \text{Normal}(\mu_i, \sigma_i^2) \sim \text{Normal}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

Central limit theorem

- The sum (mean) of a high number (>30) of independent and identically distributed (iid) random variables is distributed according to a Normal distribution independently of the type of distribution of the random variables

Teorema del límite central: Sea $X_1, X_2, ..., X_n$ un conjunto de variables aleatorias, independientes e idénticamente distribuidas con media μ y varianza σ^2 distinta de cero. Sea

$$S_n = X_1 + \cdots + X_n$$

Entonces

$$\lim_{n\to\infty} \Pr\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq z\right) = \Phi(z)$$

Teorema (del límite central): Sea $X_1, X_2, ..., X_n$ un conjunto de variables aleatoria, independientes e idénticamente distribuidas de una distribución con media μ y varianza $\sigma^2 \neq 0$. Entonces, si **n** es suficientemente grande, la variable aleatoria

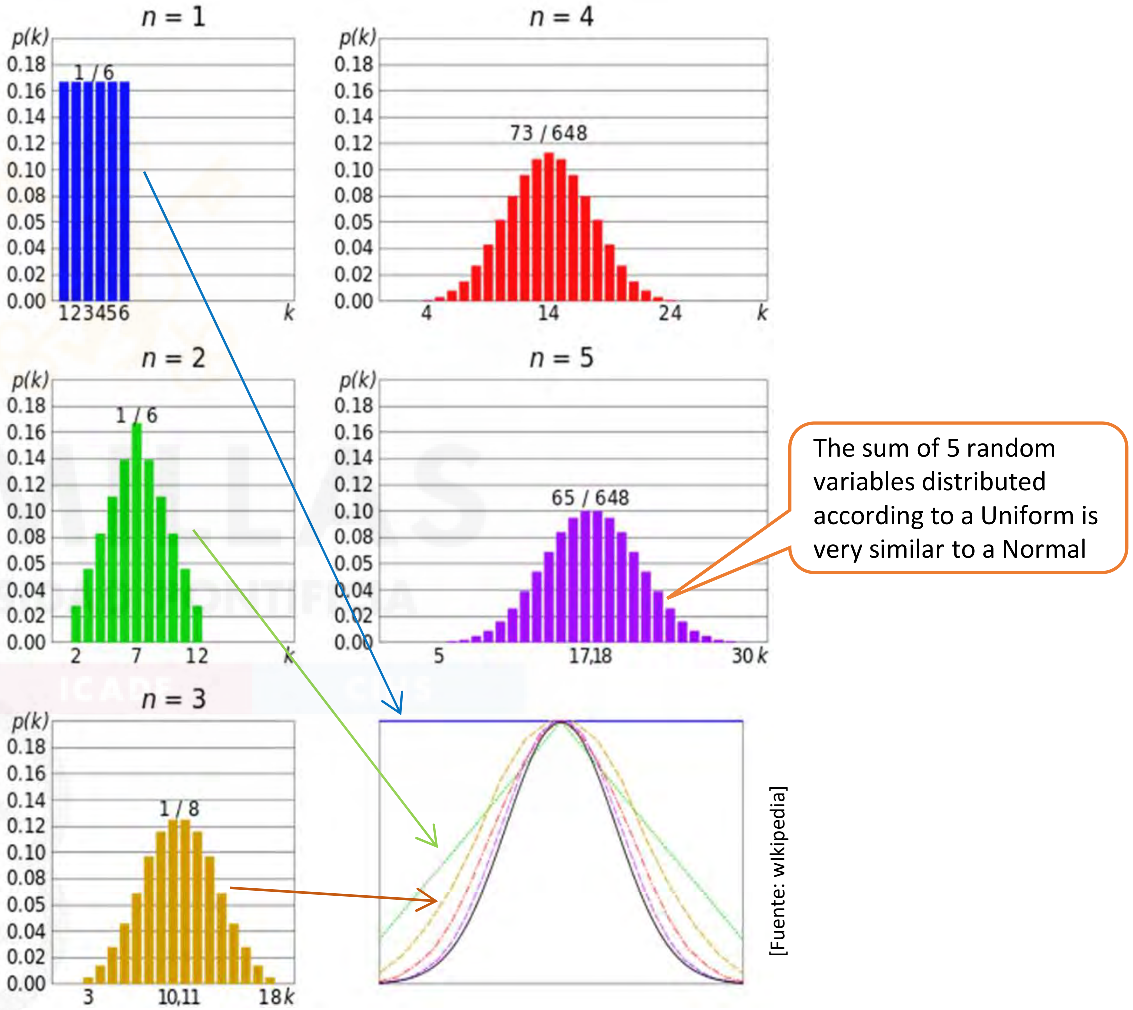
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

tiene aproximadamente una distribución normal con $\mu_{\bar{X}} = \mu$ y $\sigma^2_{\bar{X}} = \sigma^2/n$.

Central limit theorem

- Example
  - Sum of Uniform

$$S_n = X_1 + \cdots + X_n$$

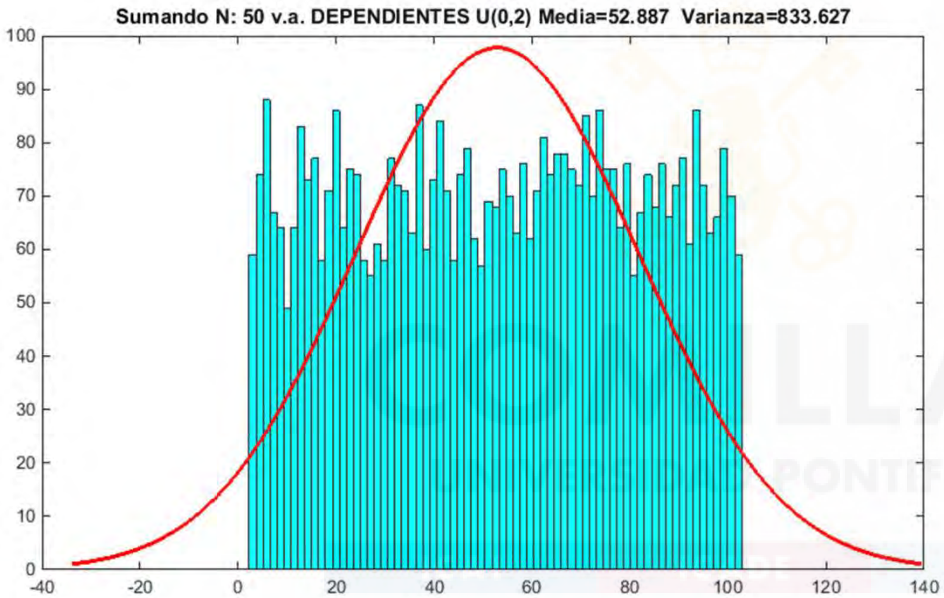


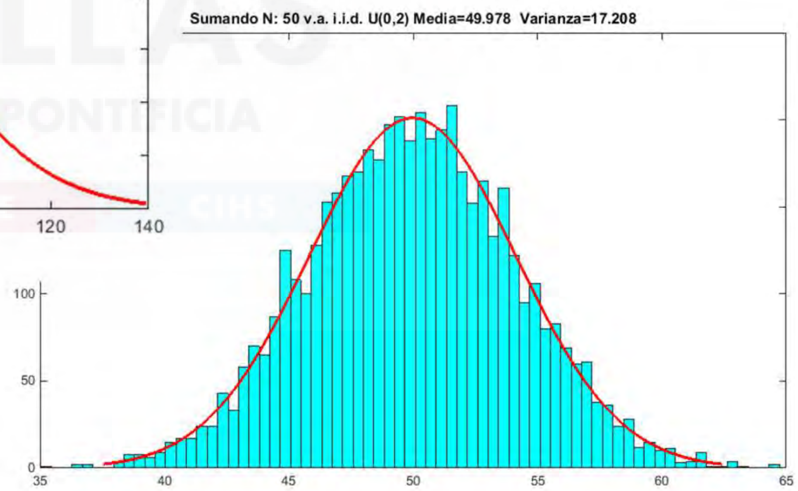
[Fuente: wikipedia]

Probability

Central limit theorem

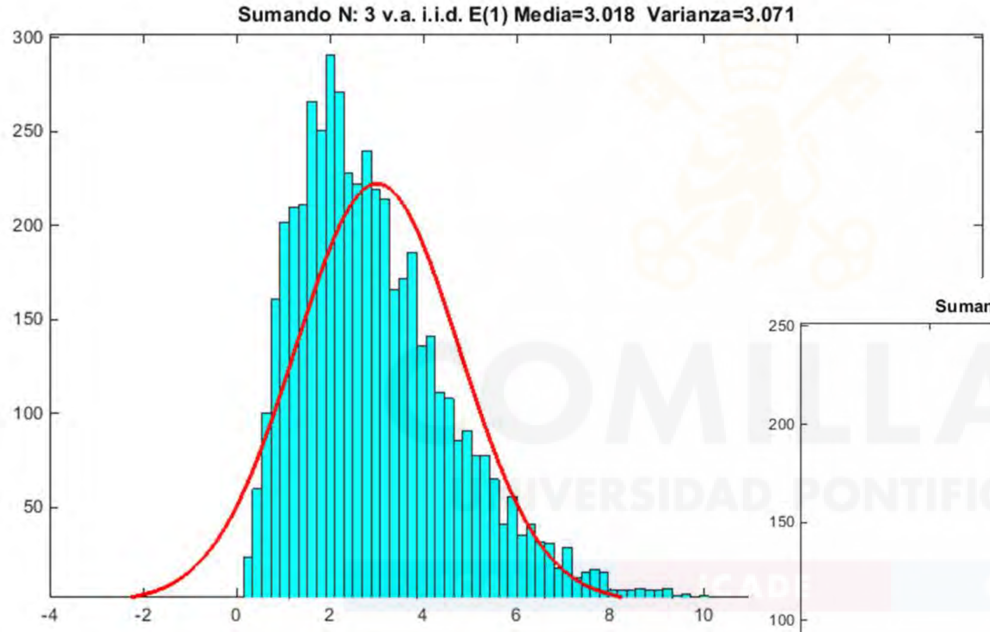
- Impact of independence

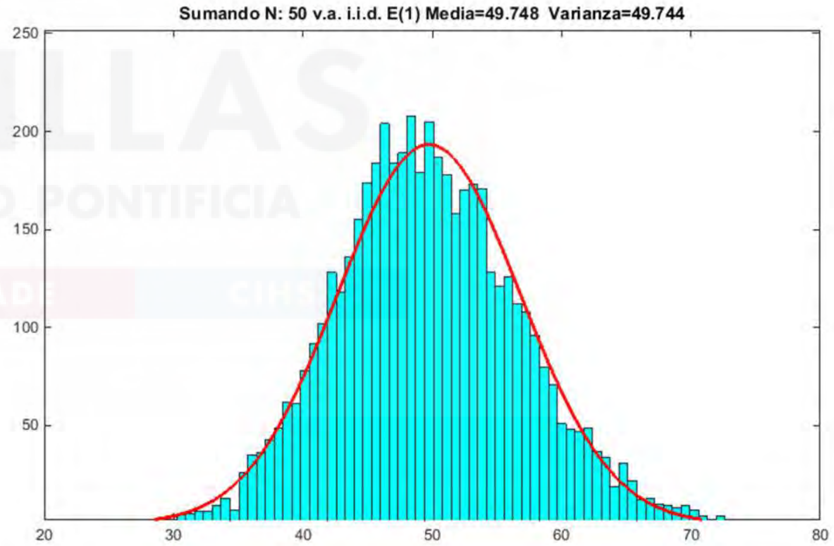




Central limit theorem

- Impact of number of variables

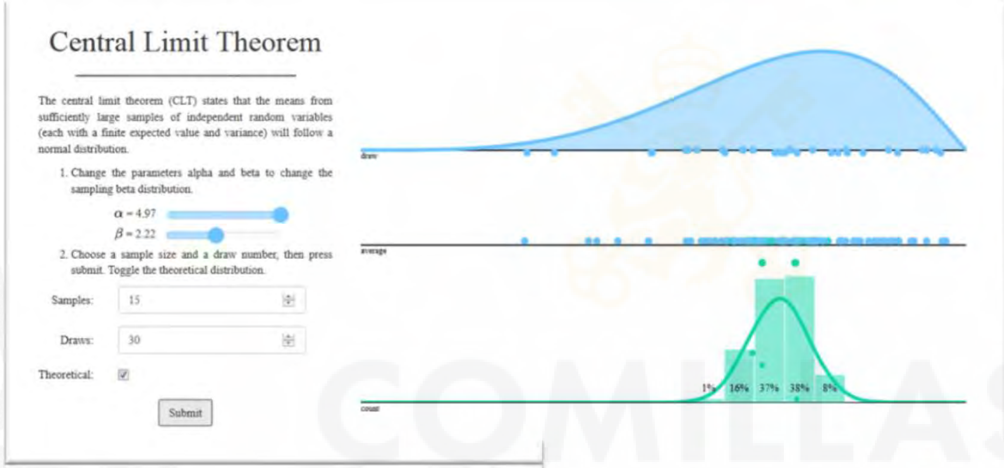


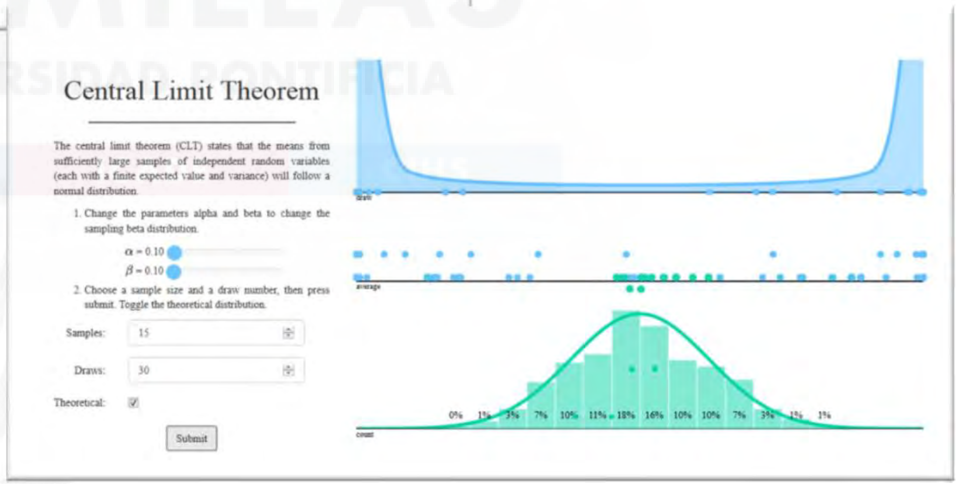


Probability

Central limit theorem

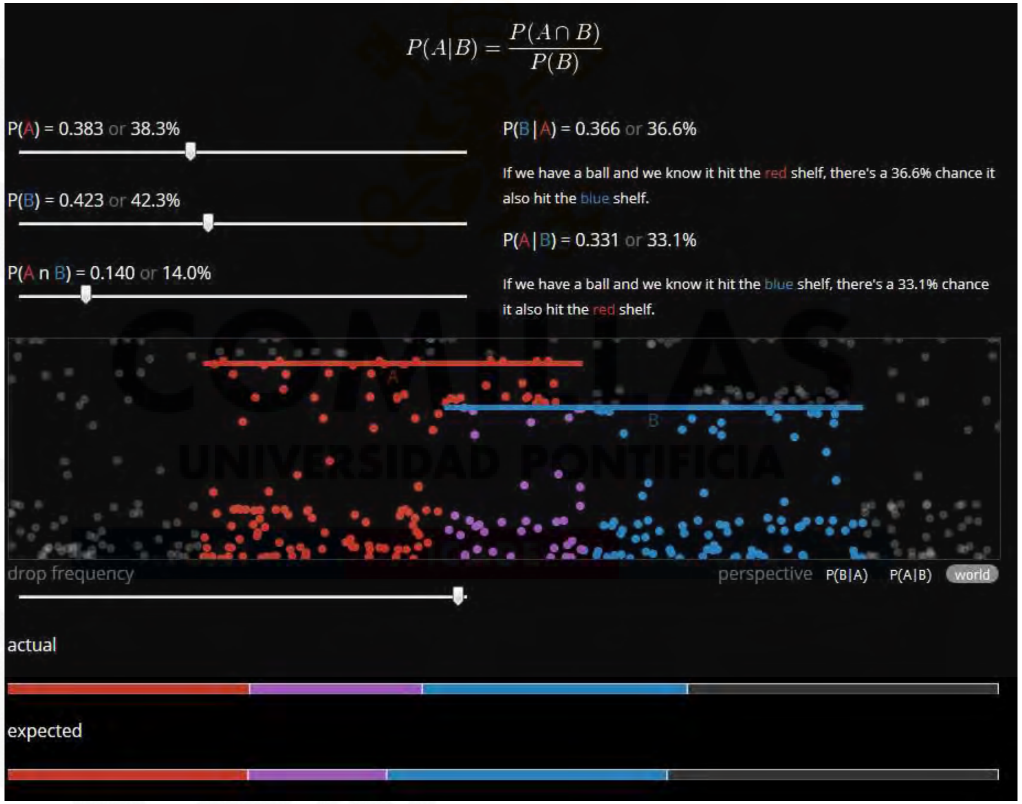
- https://seeing-theory.brown.edu/probability-distributions/index.html#section3





Probability
Conditional probability

- http://setosa.io/conditional/



Joint probability distribution, marginal and conditional

- Joint probability distribution of random discrete variables X and Y (probability mass function)

$$\begin{aligned} \mathrm{P}(X = x \text{ y } Y = y) &= \mathrm{P}(Y = y \mid X = x) \cdot \mathrm{P}(X = x) \\ &= \mathrm{P}(X = x \mid Y = y) \cdot \mathrm{P}(Y = y). \end{aligned}$$

- Joint probability distribution of random continuous variables X and Y (probability density function)

$$f_{X,Y}(x,y) = f_{Y|X}(y|x) f_X(x) = f_{X|Y}(x|y) f_Y(y)$$

Joint probability distribution, marginal and conditional

- The distribution of a variable is conditioned by the value of the other one

$$F_{XY}(x|y) = P(X \leq x | Y = y)$$

- With PDF (continuous variables)

$$f_{X,Y}(x,y) = f_{Y|X}(y|x) f_X(x) = f_{X|Y}(x|y) f_Y(y)$$

Joint PDF XY

PDF of Y **conditioned to** a X

Marginal PDF of X

- Besides

X and Y are independent

⇕

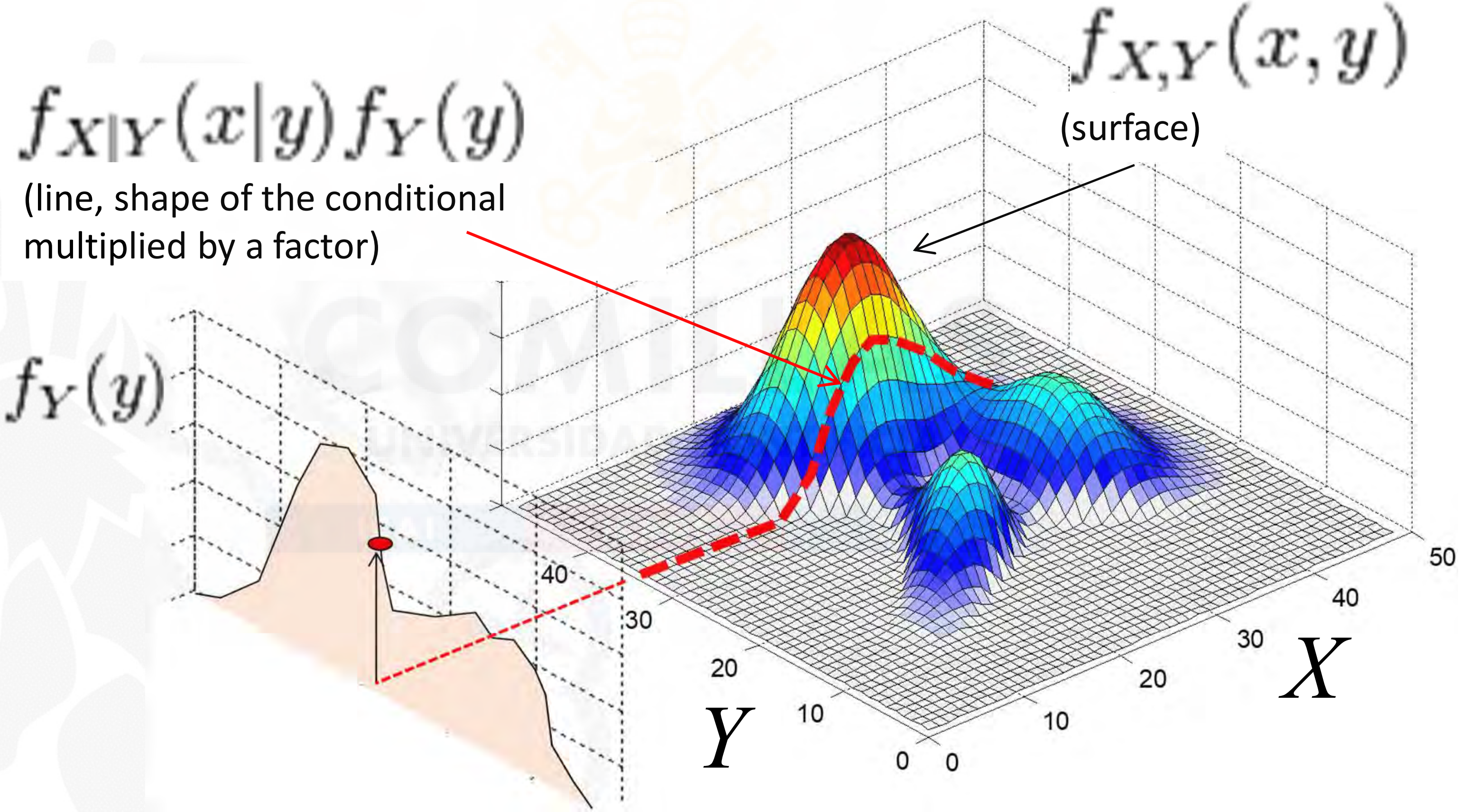
$$f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y)$$

$$\int_x \int_y f_{X,Y}(x,y)\; dy\, dx = 1.$$

Probability

Joint probability distribution, marginal and conditional

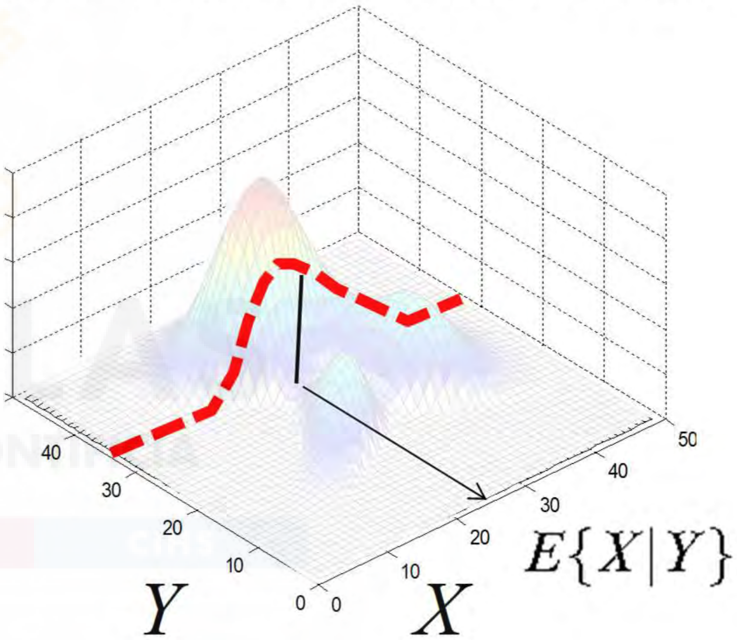
- **Example:** relation among joint, marginal, and conditional distributions

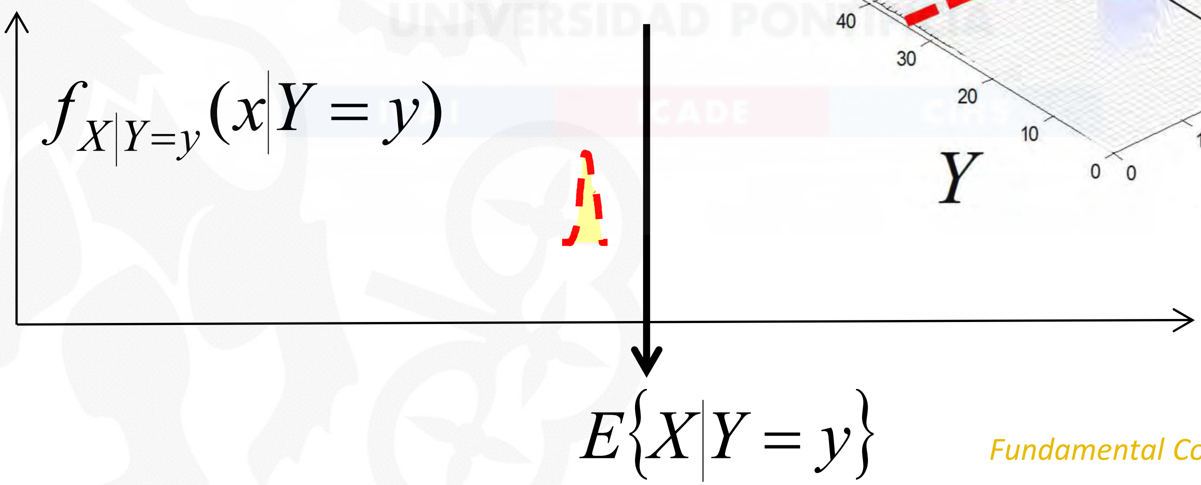


Conditional mean

- Represents the mean value of X, given a value of Y
  - I.e., expected value of $\boldsymbol{X}$ distribution conditioned to the value of $\boldsymbol{Y}$ $E\{X|Y\}$

$$E\{X|Y=y\}=\int_{-\infty}^{+\infty} x\cdot f_{X|Y=y}(x|Y=y)dx$$

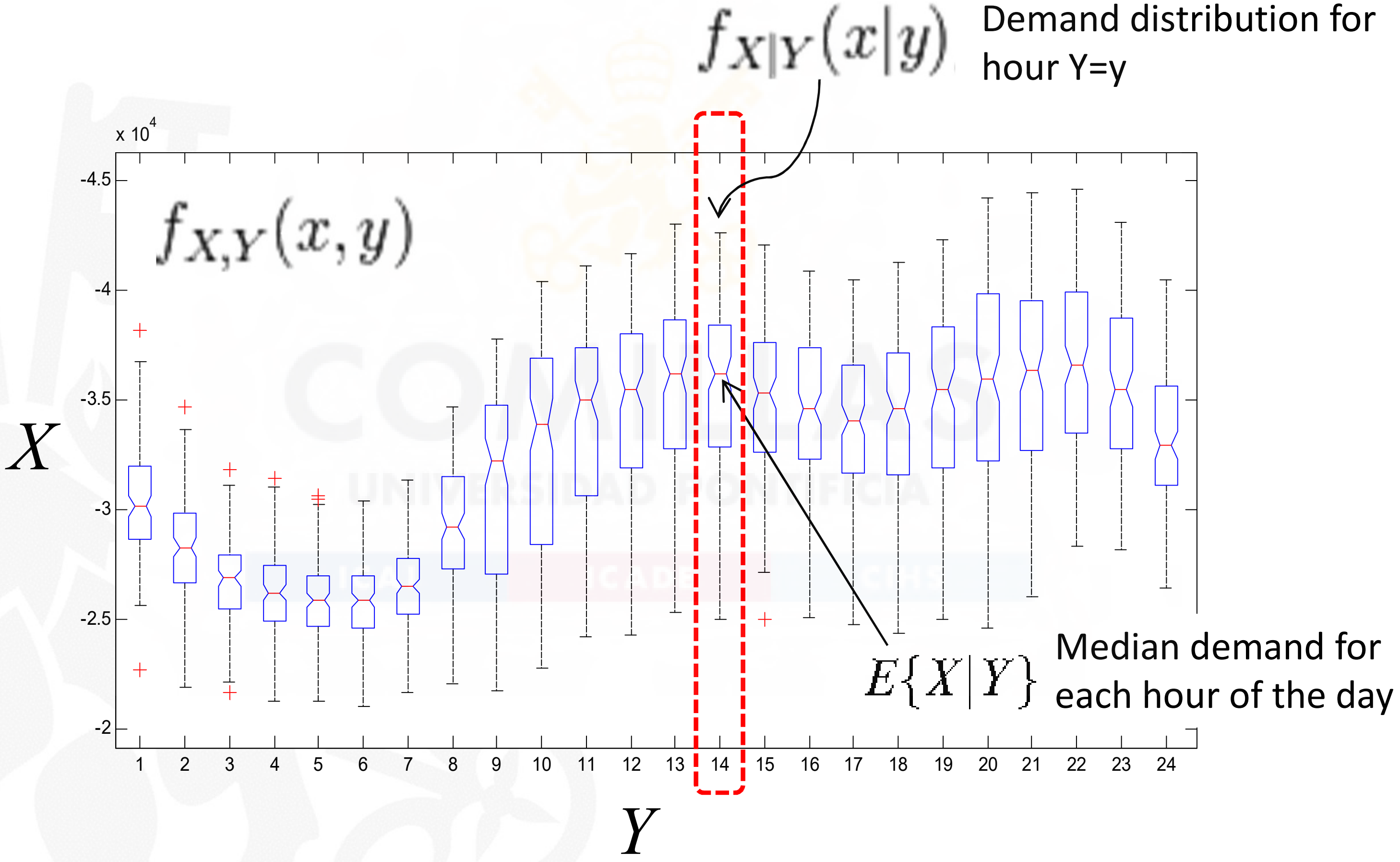




Probability

Conditional mean

- Example (empirical) distribution of the demand for each hour (2012)

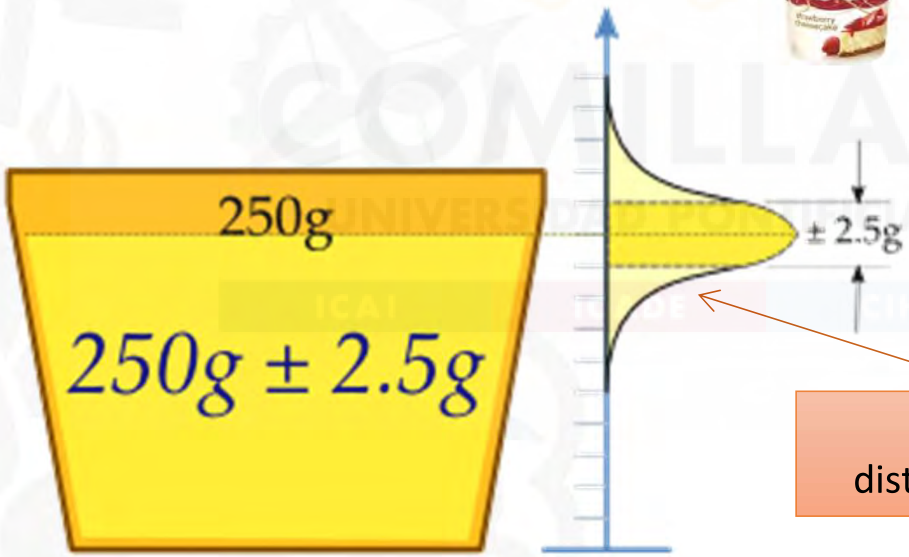


3

Statistical Inference

Idea

- Expected weight?
- Tolerance?



What does this distribution represent?

Statistical inference

Population vs. sample

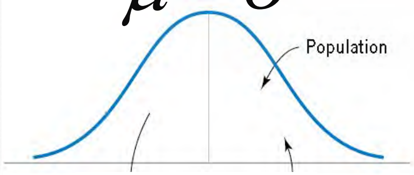
- Population $\boldsymbol{N}$
  - Set of all possible observations
  - May have a finite or infinite size
  - We assume that it is unknown

Sampling

- Observed sample $\boldsymbol{n}$
  - Any subset of the population
  - The greater the sample size, the more accurate and reliable will be the inferences about the population
  - It is possible to obtain different samples. Each sample may give a different estimation

$x_1, x_2, \ldots, x_n$

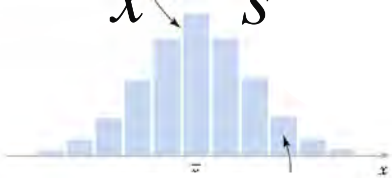
$\mu \quad \sigma^2$



Distribution

Estimation

$\bar{x} \quad s^2$

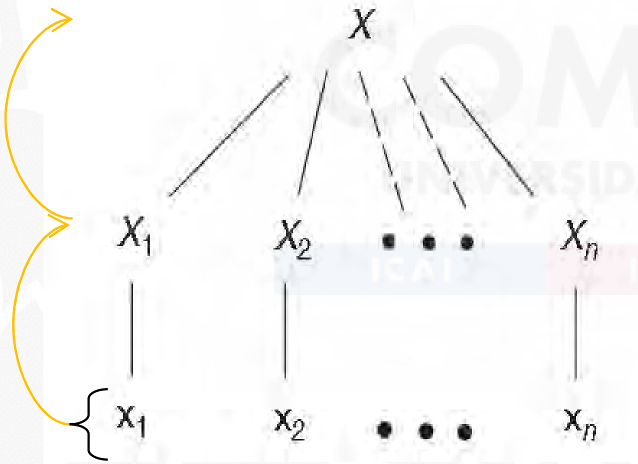


Distribution

Statistical inference

Random sample

- (Simple) random sample (s.r.s) of size $\boldsymbol{n}$
  - Set of n unidimensional independent and identically distributed random variables according to a probability law of the population
  - For a s.r.s. there exist theoretical developments of interest



Population (random variable)

Random sample of size n (random variables)

Observed sample (n values)

Statistical inference

Estimator vs. estimation

- Estimator
  - It is a random variable function of other random variables used to estimate the parameter of a population

$$\hat{\Theta} = h(X_1, X_2, \ldots, X_n) \longrightarrow \theta$$

  - Given that it is a random variable, it has an associated probability distribution (sample distribution)
  - The standard deviation of an estimator is called the standard error
- Estimation
  - It is the value obtained for an estimator from the specific observed sample data

$$\hat{\Theta} = h(X_1, X_2, \ldots, X_n) \searrow \text{Estimation}$$

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

Estimator. Sample mean

- Population (unknown)

$$\left.\begin{aligned} E\{X\} &= m, \\ \operatorname{var}\{X\} &= \sigma^2. \end{aligned}\right\}$$

- Sample mean is a (good) estimator of the population mean

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

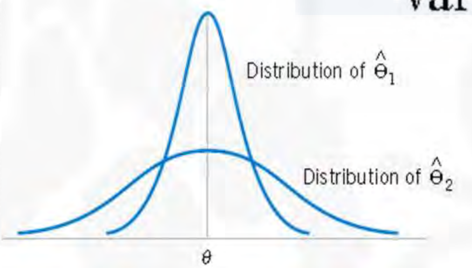
Estimator. Sample mean

- Expected value of the sample mean
  - Coincides with the population mean

$$E\{\overline{X}\} = \frac{1}{n}\sum_{i=1}^{n} E\{X_i\} = \frac{1}{n}(nm) = m,$$

- Variance of the sample mean
  - Proportional to the population variance
  - Decreases linearly with the sample size

$$\operatorname{var}\{\overline{X}\} = E\{(\overline{X}-m)^2\} = E\left\{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - m)\right]^2\right\}$$

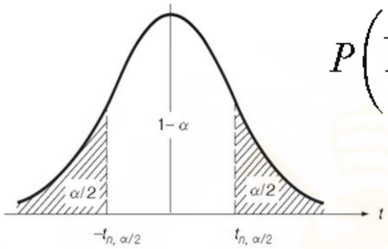


$$= \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n},$$

Statistical inference

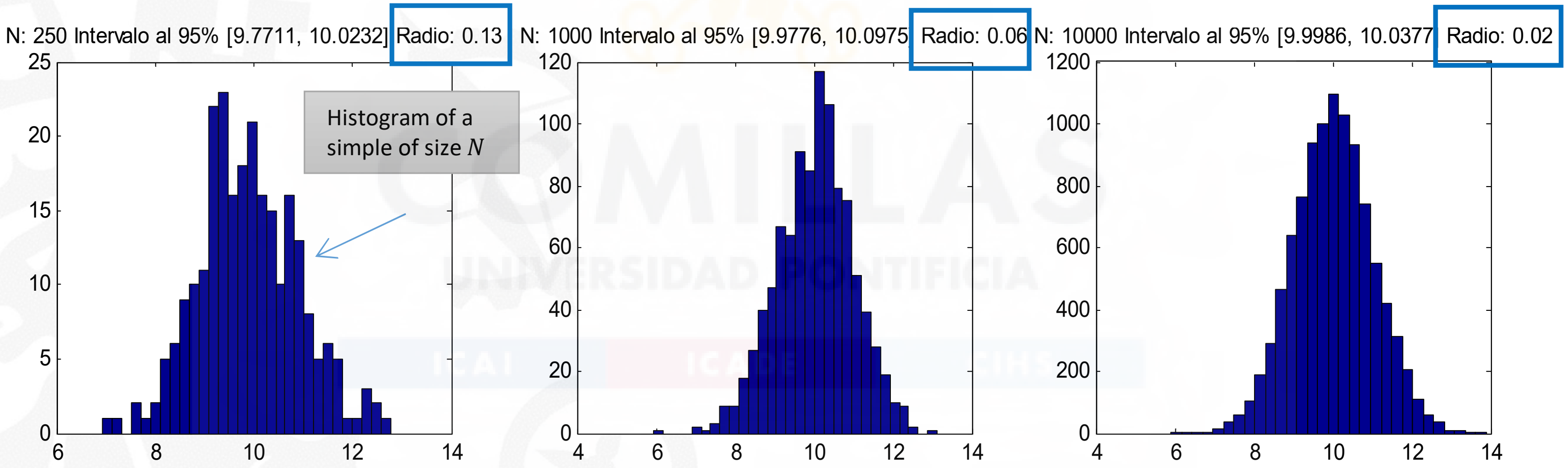
Example. Accuracy of sample mean estimator

$$X \approx N(10,1)$$



$$P\left(\overline{X} - \frac{t_{n-1,\alpha/2}S}{n^{1/2}} < m < \overline{X} + \frac{t_{n-1,\alpha/2}S}{n^{1/2}}\right) = 1-\alpha$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$



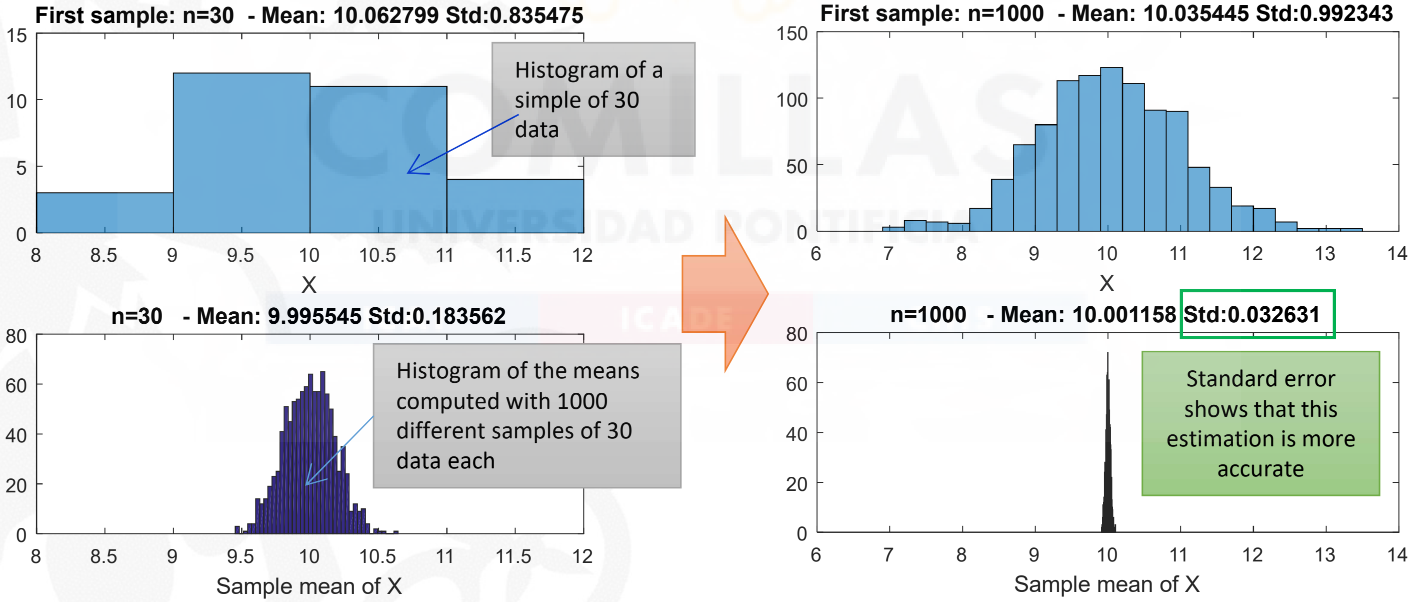
Increasing n (sample size) increases the accuracy

Statistical inference

Example. Accuracy of sample mean estimator

```
% SAMPLE MEAN DISTRIBUTION (EMPIRICAL VIEW)
n = 30; % sample size
m = 1000; % number of samples
X = normrnd(10,1,m,n); % each row is a sample
means = mean(X,2);
```

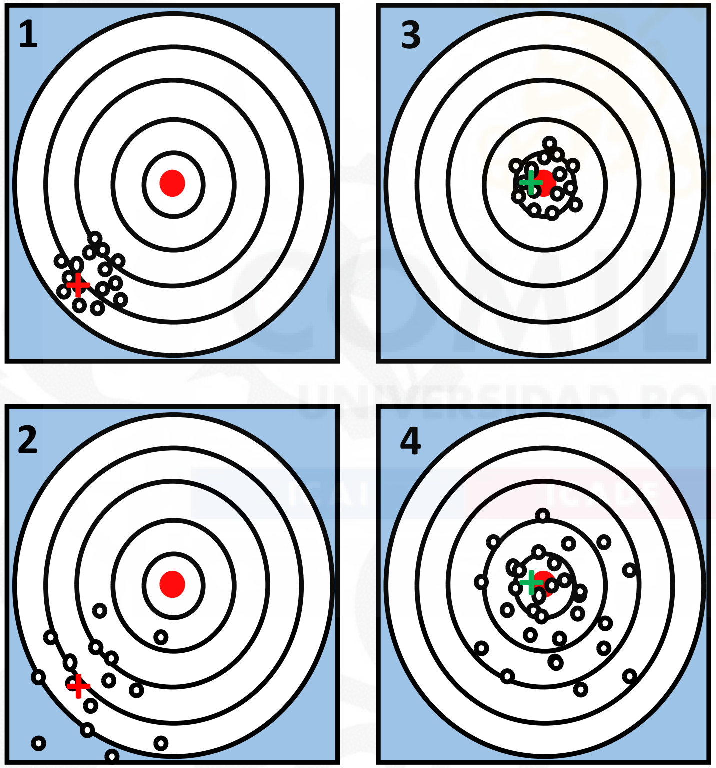
```
figure;
ax(1)=subplot(2,1,1);
histogram(X(1,:));
xlabel('X');
title (sprintf('First sample: n=%d - Mean: %f Std:%f', ...,
n, mean(X(1,:)), std(X(1,:))));
ax(2)=subplot(2,1,2);
hist(means,40);
title(sprintf('n=%d - Mean: %f Std:%f',n, mean(means),
std(means)));
xlabel ('Sample mean of X');
linkaxes(ax,'x');
```



Statistical inference

Estimator: desirable characteristics

- An estimator must give good estimations
- Example estimation of the center of the bullseye



- We want to estimate the center from the coordinated of the arrows thrown by a player
- Players 3 and 4 are clearly better than 1 and 2
- Expected value coincides with the center of the bullseye
- Player 3 is much better than player 4

Estimator. Unbiased sample variance

- Unbiased sample variance is a (good) estimator of the population variance (unknown)

$$\left.\begin{aligned} E\{X\} &= m, \\ \text{var}\{X\} &= \sigma^2. \end{aligned}\right\}$$

- Definition

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

  - If n is large, dividing by n or $n-1$ is quite similar
  - The expected value of the sample variance coincides with the population variance

$$E\{S^2\} = \sigma^2$$

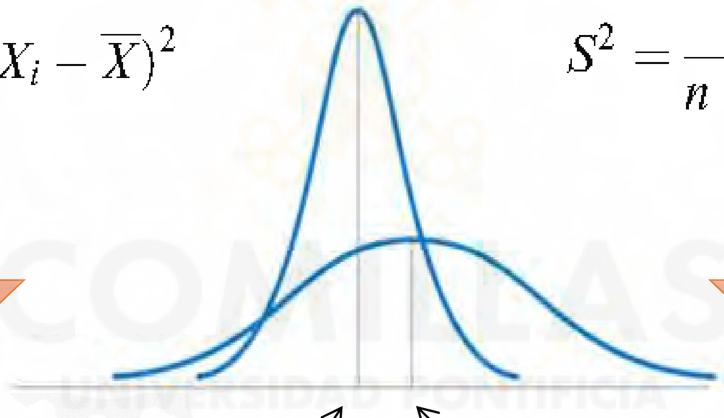
Statistical inference

Unbiased sample variance vs biased sample variance

- Sample variance is biased, the unbiased sample variance not

$$S^{2*} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$



$$E\{S^{2*}\} = \frac{n-1}{n}\sigma^2$$

Sample variance underestimated the population variance is n is small

$$E\{S^2\} = \sigma^2$$

Unbiased and consistent estimator of the population variance

Confidence interval of the population mean

- The natural point estimator is the sample mean

$$E\{\overline{X}\} = \frac{1}{n}\sum_{i=1}^{n} E\{X_i\} = \frac{1}{n}(nm) = m,$$

- If the population follows a Normal distribution of known mean $N(m, \sigma^2)$

- Then

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \longrightarrow \overline{X} \approx N\left(m, \frac{\sigma^2}{n}\right) \longrightarrow U = \frac{\overline{X} - m}{\sigma / \sqrt{n}} \approx N(0,1)$$

- If n is large, it is not needed to follow a normal distribution (central limit theorem)

Statistical inference

Confidence interval of the population mean

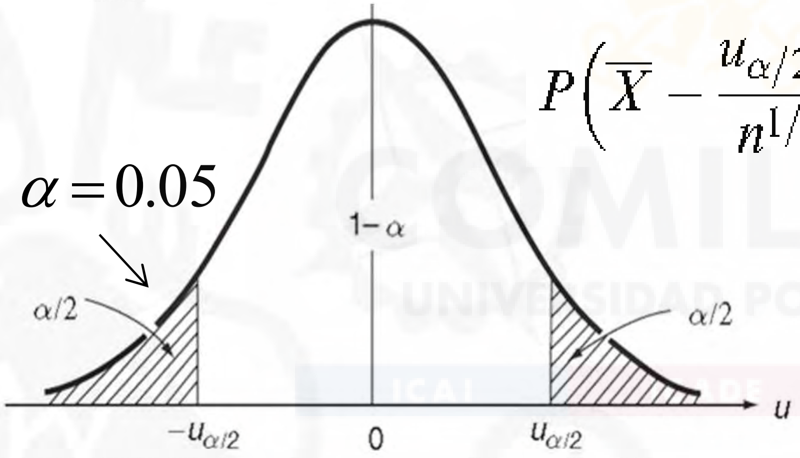
- Fixing a confidence level $1-\alpha$

$$U = \frac{\overline{X} - m}{\sigma/\sqrt{n}} \approx N(0,1)$$

$$P(-u_{\alpha/2} < U < u_{\alpha/2}) = 1 - \alpha$$

$$P\left(\overline{X} - \frac{u_{\alpha/2}\sigma}{n^{1/2}} < m < \overline{X} + \frac{u_{\alpha/2}\sigma}{n^{1/2}}\right) = 1 - \alpha$$

$\alpha = 0.05$



If repeated with infinite number of samples, in $(1-\alpha)100\%$ of the cases the CI will contain the true parameter

- Confidence interval obtained centered around the sample mean
- If the sample size increases, the confidence interval reduces (for the same confidence level)

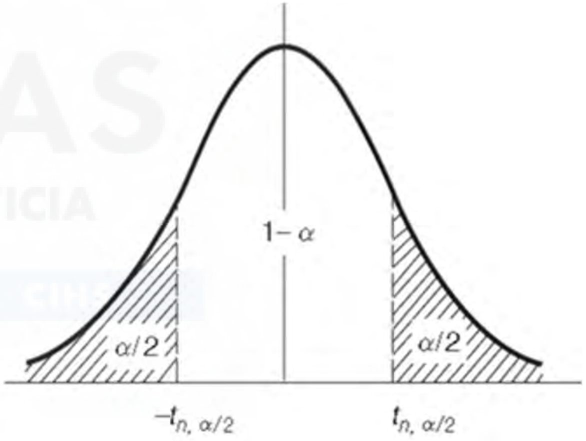
Confidence interval of the population mean

- Usually, the population variance is unknown
  - Unbiased sample variance used as an estimation of the population variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

$$T = \frac{\overline{X} - m}{S/\sqrt{n}} \approx t_{n-1}$$ (Student's t)

$$P\left(\overline{X} - \frac{t_{n-1,\alpha/2}S}{n^{1/2}} < m < \overline{X} + \frac{t_{n-1,\alpha/2}S}{n^{1/2}}\right) = 1 - \alpha$$



$$P(-t_{n-1,\alpha/2} < Y < t_{n-1,\alpha/2}) = 1 - \alpha$$

Point and interval estimation

- In both cases, the estimation is obtained from the (distribution) of an estimator and an observed sample

- Point estimation

$$\hat{\Theta} = h(X_1, X_2, \ldots, X_n)$$

$$E\{\hat{\Theta}\}$$

→ Estimation (value)

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

- Interval estimation

$$\hat{\Theta} = h(X_1, X_2, \ldots, X_n)$$

$$E\{\hat{\Theta}\} \quad \text{var}\{\hat{\Theta}\}$$

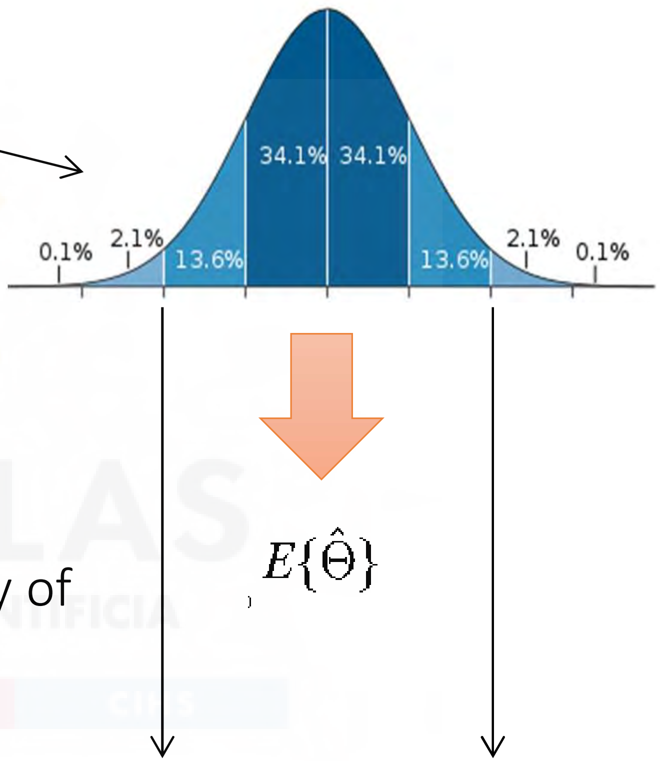
→ Estimation (interval)

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

Statistical inference

Point and interval estimation

$$\hat{\Theta} = h(X_1, X_2, \ldots, X_n)$$



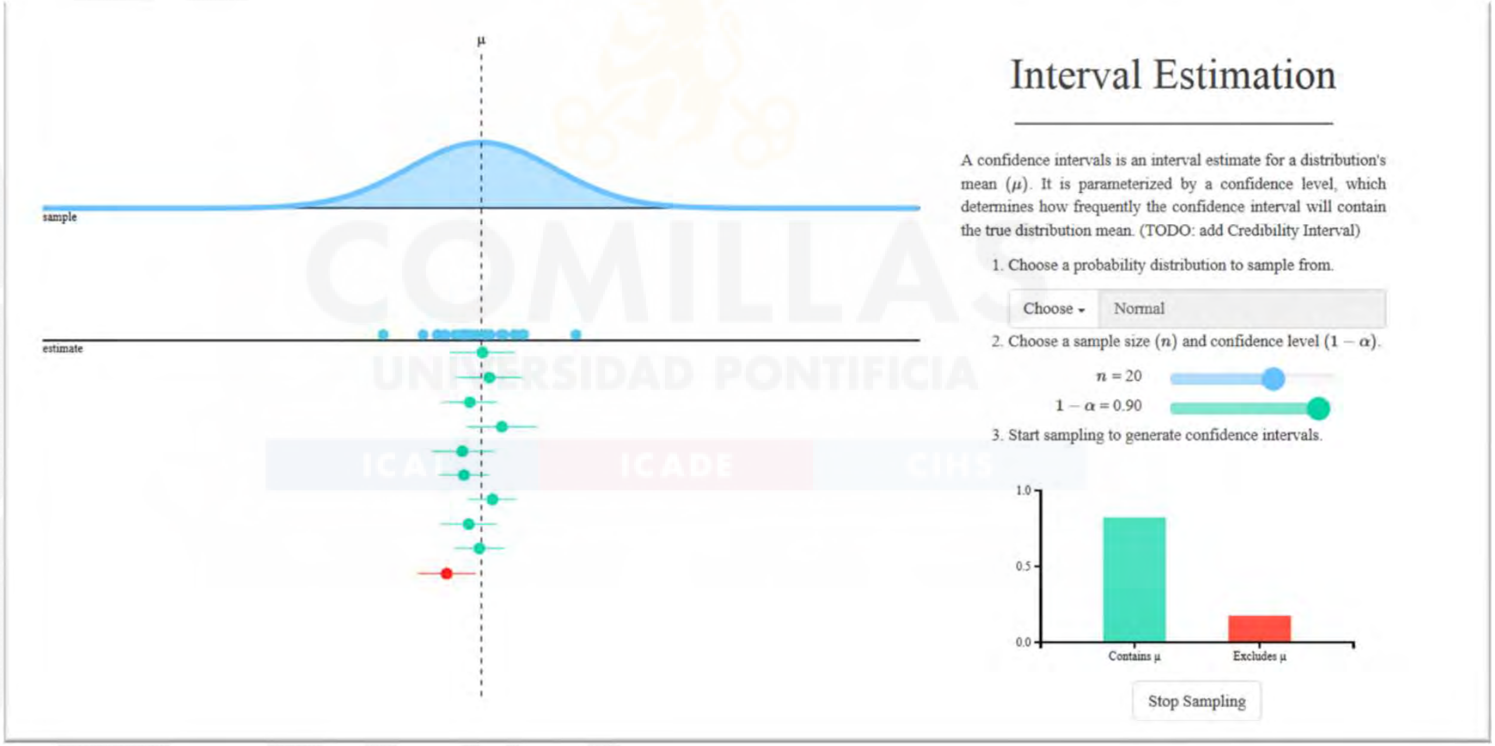
$$E\{\hat{\Theta}\}$$

$$L_1 < \theta < L_2$$

- In point estimation the expected value of the estimator is chosen, ignoring its standard error
- In interval estimation the accuracy of the estimator is not ignored (its variance). It is used to build a variance range with a certain confidence level

Confidence level for the mean

- https://seeing-theory.brown.edu/frequentist-inference/index.html#section2
  - Impact of n
  - Impact of the confidence level



Hypothesis test

- Hypothesis test
  - Allows to check if an assumed property of the population is compatible with the samples observed

- There are always two hypotheses
  - H0: Hypothesis NULL, the one we want to contrast
  - H1: Hypothesis ALTERNATIVE

- Types of hypotheses
  - Parametric
    - The download average speed is 300 Mbps
  - Nonparametric
    - The download average speed follows a normal distribution

Hypothesis test: type errors

- In general, H_0 is accepted unless the sample shows clear evidence against
- It is possible to make mistakes in both directions

	H_0 **is true**	H_1 **is true**
H_0 chosen	No error (true positive)	Type II error (β or false negative)
H_1 chosen	Type I error (α or false positive)	No error (true negative)

$$P(choose\ H_1 | H_0\ is\ true) = \alpha$$
$$P(choose\ H_0 | H_1\ is\ true) = \beta$$

Ideally probabilities must be as lower as possible

- Test power

$$P(choose\ H_1 | H_1\ is\ true) = 1 - \beta$$

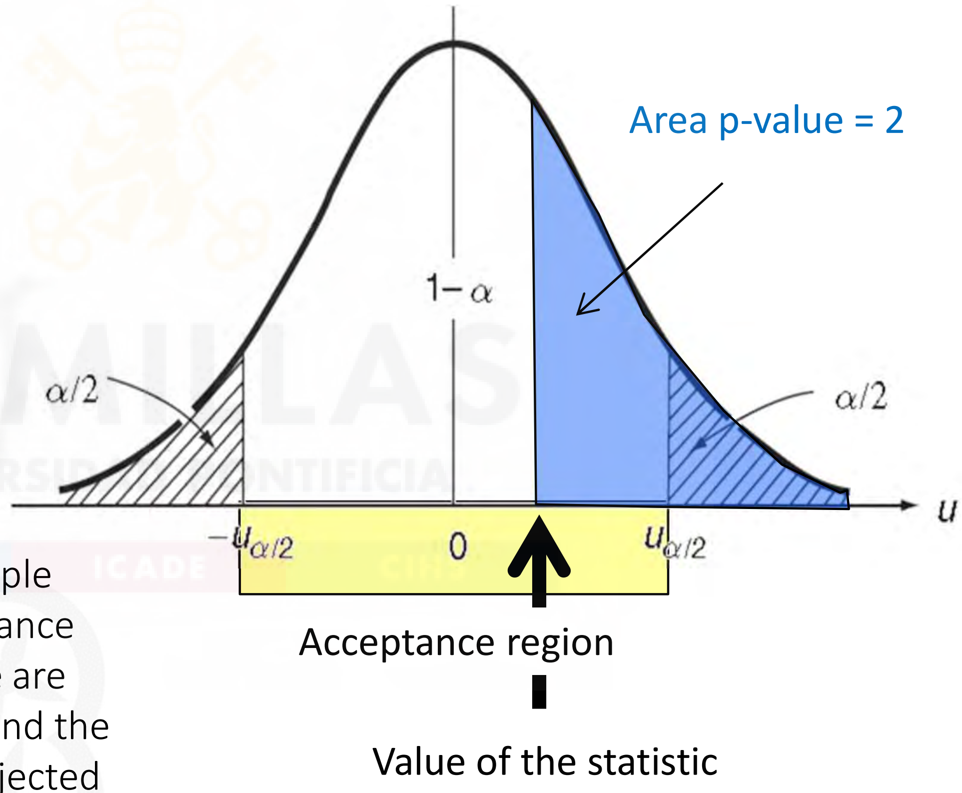
Parametric hypothesis test

- Bilateral test for the mean of a normal distribution with unknown variable

H_0: $\mu=\mu_0$

H_1: $\mu\neq\mu_0$

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$



If the estimator of the sample belongs to the non-acceptance region, it means that there are significant discrepancies, and the null hypothesis must be rejected

Student's distribution

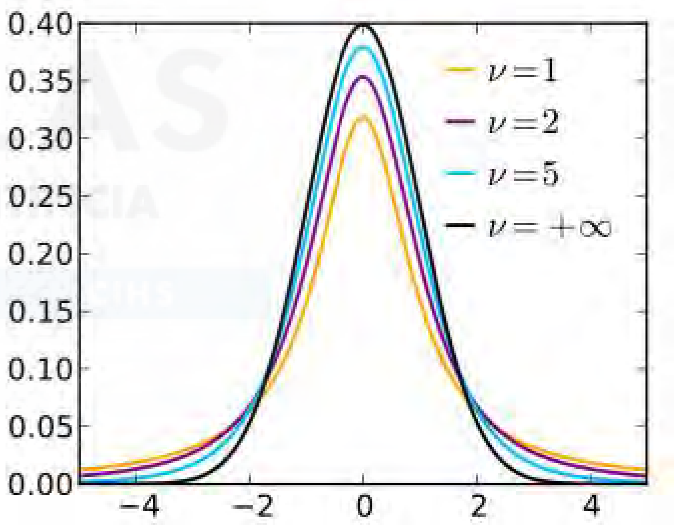
- Symmetric and centered around 0
- It has a parameter (degrees of freedom)
  - Doesn't change too much with the degrees of freedom

Sample of size n from a normally distributed population with expected value μ and variance σ^2

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$\frac{\bar{X}-\mu}{S/\sqrt{n}}$$



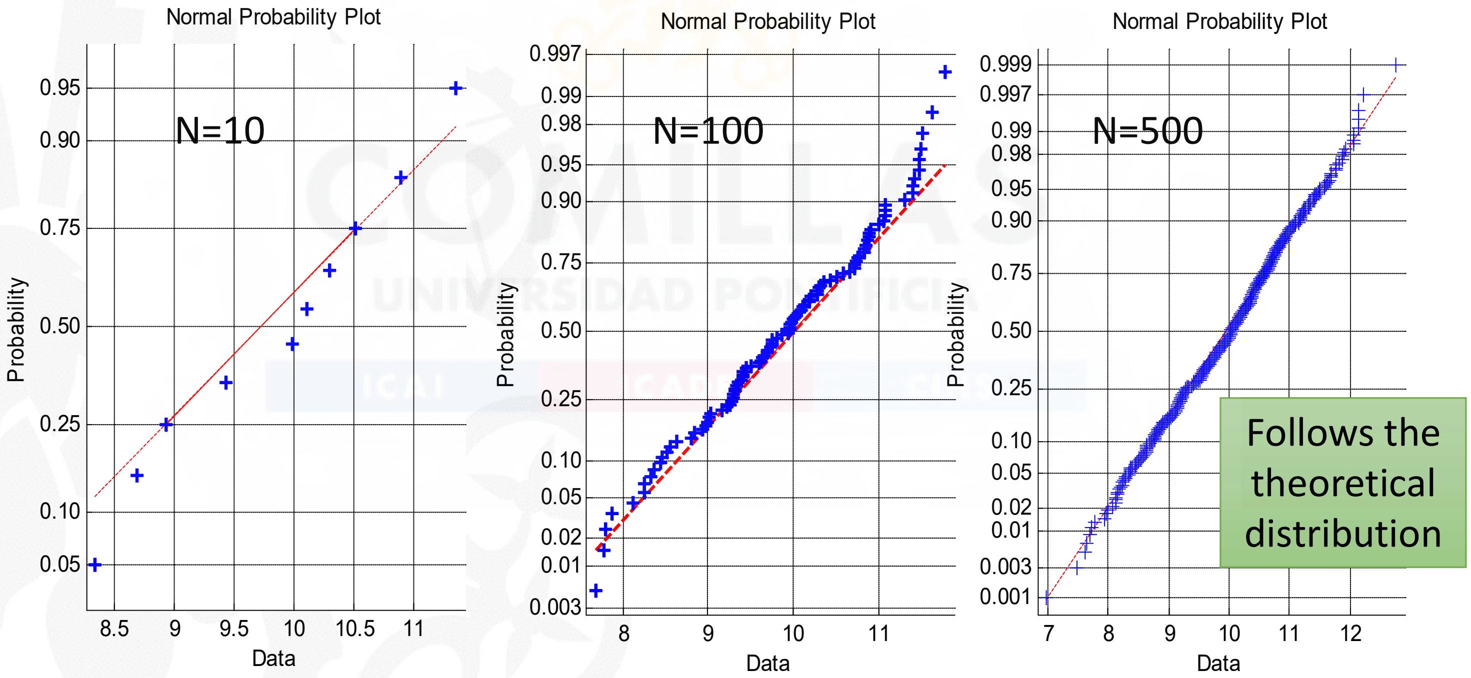
Probability plots

- Test by means of probability plots
  - Allows to compare visually the theoretical distribution (Normal) with an empirical function obtained from the sample
  - There are different types of plots, among them the Normal probability plot or qq-plot
  - Based on computing quantiles of the theoretical and empirical distributions and representing them together
  - If both distributions are equal, points concentrate along a straight line, being common that exists larger variability at the extremes

Normality plots

- Example
  - Random samples from a Normal of mean 10 and variance 1

```
x=normrnd(10,1,N,1);
figure; normplot(x);
```

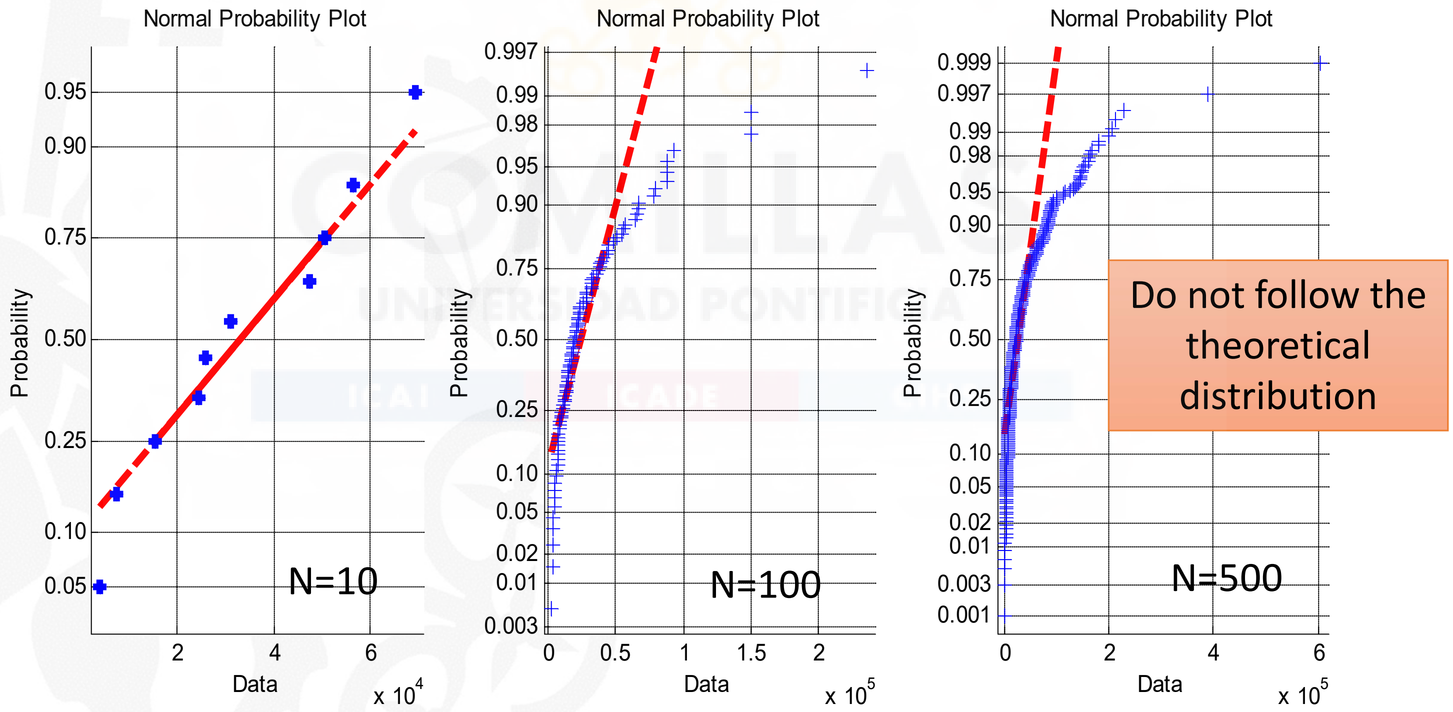


Statistical inference

Normality plots

- Example
  - Random samples from a Lognormal of mean 10 and variance 1

```
x=lognrnd(10,1,N,1);
figure; normplot(x);
```



Thank you for your attention

Prof. Eugenio Sánchez Úbeda