# Advanced Statistics
# Part one: Parametrical Statictics

M. Pérez-Casany

Dept. of Statistics and Operations Research and DAMA-UPC
Technical University of Catalunya

First Semester

# Program: Parametrical Modelling

**Session 1: Introduction**
1. The objective of Statistics. 2. Descriptive Statistics. 3. Statistic. 4. Some Special Probability distributions. 5. Normal and related distributions. 6. Central Limit theorem.

**Session 2: Parameter Estimation and Hypothesis Testing**
1. Point estimation: moment method, maximum likelihood and least squares (for the mean value). 2. Confidence interval estimation. 3. Measures of quality of an estimator. 4. Definition of Normal linear Model. 5. Examples

**Session 3: Normal Linear Model: ANOVA**
1. One way ANOVA. 2. Parameter Estimation. 3. Confidence interval for a mean and a difference of means. 4. Multiple comparaisons. 5. Model checking. 6. Random Blocks model. 7. Two way ANOVA with crossed and nested factors. 8. Examples

### Session 4: Normal Linear Model: Regresion

1. Simple linear regresion models: parameter estimation, coefficient of determination, Mean square error, confidence intervals for parameteres and predictions, model checking. 2. Multiple linear regresion models. 3. Collinearity and causality. 4. Robust fit and outlier detection. 5. Common mistakes in regresion. 6. Examples

### Session 5: Generalized Linear Models I

1. Exponential families of probabilities: definition, main properties and examples. 2. Generalized Linear model: canonical link, variance function, dispersion parameter. 3. Models for binary responses. 4. Examples

### Session 6: Generalized Linear Models II

1. Models for count data. 2. Overdispersion. 3. Inference. 4. Model cheking. 5. Contingence tables. 6. Examples

# Session 1: Introduction

# 1. Objective of Statist

**Objective of statistics**: to study the behaviour of a random variable (r.v) in a given *population*

Examples:

To study the performance of a *system* one can choose between different *merics* as for instance: *system throughput, network bandwidth, response time*.

The metrics are usually the r.v.'s and It is crucial to choose the correct ones and to measure them appropiately.

The systems are going to be the population

# 1. Objective of Statist

Types of r.v's: **numerical** and **cathegorical**

The numerical ones are divided into:

- **continuous**: values in a given interval
- **discrete**: integer values

# 1. Objective of Statist. Descriptive

Examples:

- *time of a given workload*
- *number of jobs pocessed in an hour*
- *waiting time in a queue*
- *visits to a given site*
- *number of bits transmited per unit of time*
- *number of bits correctly transmited in a set of 24 bits*
- *time to the first machine collapse*

# 2. Descriptive Statistics

- $X$ denotes a r.v.
- $x_1, x_2, x_3, \cdots x_n$ denotes a *sample* from $X$

  That is a set of realizations (observations) of $X$ obtained **independently**
- $n$ is the *sample size*

**Descriptive Statistics** objective: to describe what is observed in the sample by means of some *measures* and/or *plots*

It has **no error**

# 2. Descriptive Statistics

- Position measures:

1) **sample mean** or **arithmetic mean**

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

Observation: It minimize the function $f(a) = \sum_{i=1}^{n}(x_i - a)^2$

2) **mode**, value that appears more often in the sample
3) **median**, central value once the sample has been ordered

Observation: it minimize the function $g(a) = \sum_{i=1}^{n} |x_i - a|$.

# 2. Descriptive Statistics

- Dispesion measures:

  1) **variance**

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

  2) **standard deviation** $S = \sqrt{S^2}$,

  3) **variation coefficient**

  $$CV = 100 \times \frac{S}{\overline{x}}$$

  larger than 50 indicates to much variability in the sample.

  4) **interquartilic range** $Q_3 - Q_1$
  where $Q_i$ $i = 1..4$ are the quartiles.

# 2. Descriptive Statistics

## 2. Descriptive Statistics

Denoting by $\mu_k^* = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^k$

1) **skewness**

$$skew = \frac{\mu_3^*}{(\mu_2^*)^{3/2}}$$

Measuare of the lack of symmetry . Simmetric data $skew = 0$

2) **kurtosi**

$$kurt = \frac{\mu_4^*}{(\mu_2^*)^2}$$

Measures the intensity of the peak for unimodal distribution.

# 2. Descriptive Statistics

Important plots:

- **Histogram** (numerical var.)
- **Box-Plot** (numerical var. Useful to identify outliers)
- **Bar plot** (caegorical var.)
- **pie-chart** (categorical var.)
- dispersion plot or **scatterplot** ( for two numverical var.)

**Outlier**: An observation that appears to deviate markedly from the other members of the sample in which it occurs (*The Cambridge Dictionary of Statistis: B.S. Everitt*)

# 2. Descriptive Statistics

Typical sentences from Descriptive Statistics:

- 4% of the works are prosessed with less than 5 msec
- 60% of the works last between 10 msec and 1 sec
- just one work takes more than half a minute
- the mean time is 457 m sec

# 2. Descriptive Statistics

Limitations of Descriptive Statistics

It doesn't allow:

- to predict the performance of future works
- to compare two systems
- to determine the optimal value of a parameter

Question: Given two performance techniques and assuming that the *time* is the metric, not always the one with minimum sample mean is the best. Why?

# 3. Some Special Prob. distributions

- Discrete case: **probability mass function** (pmf)

$$p_k, \quad \forall k \in \mathcal{S}, \quad \text{such that } 0 \le p_k \le 1 \;\; \text{and} \; \sum p_k = 1$$

- Continuous case: **density function** $f(x)$ (pdf)

$$0 \le f(x) \;\; \forall x \in \mathcal{S}, \quad \text{such that} \int_S f(x)\,dx = 1$$

# 3. Some Special Prob. distributions

Let $Y$ be a r.v.

- mean

$$\mu = E(Y) = \sum_{k \in S} k\, p_k, \qquad \mu = E(Y) = \int_S x\, f(x)\, dx$$

- variance

$$\sigma^2 = Var(Y) = \sum_{k \in S} (k-\mu)^2\, p_k, \quad \sigma^2 = Var(Y) = \int_S (x-\mu)^2\, f(x)\, dx$$

Observation: $Var(Y) = E(Y^2) - (E(Y))^2$

# 3. Some Special Prob. distributions

If $Y$ is a r.v. with a given prob. distribution,

Also important:

- **cumulative probability function** (cpf) or **distribution function**

$$F(m) = P(Y \leq m) = \sum_{k \in S,\, k \leq m} p_k \quad F(m) = \int_l^m f(x)\, dx$$

- **survival function** (sf)

$$S(m) = P(Y > m) = 1 - F(m)$$

Observation: In the continuous case, $F'(x) = f(x)$

# 3. Some Special Prob. distributions

The most important parameters of a distribution are the mean ($\mu$) and the variance ($\sigma^2$).

It is very important to clearly distinguish the theoretical or *population parameters* and the *sample parameters*.

By *statistic* one understants a numerical characteristic of a sample (The Cambridge Dictionary of Statistics. B.S. Everitt)

For example, the sample mean and sample variance. Which are good estimations of the population mean and variance respectively.

$$\hat{\mu} = \overline{x}, \qquad \hat{\sigma}^2 = S^2$$

# 3. Some Special Prob. distributions

### Discrete Uniform

A r.v. $Y \sim Uniform$ in $\{m, m+1, \cdots, n\}$ if and only if

$$\Pr\{Y = x\} = \frac{1}{n - m + 1}$$

The cdf is equal to:

$$F(x) = \begin{cases} 0 & \text{if } x < m \\ \frac{x-m+1}{n-m+1} & \text{if } m \leq x < n \\ 1 & \text{if } n \leq x \end{cases}$$

# 3. Some Special Prob. distributions

$$\mu = \frac{n+m}{2}, \quad \sigma^2 = \frac{(n-m+1)^2 - 1}{12}$$

EXAMPLES:

- *The track numbers for seeks on a disk*
- *The I/O device number selected for the next I/O*
- *The source and destination node for the next packet on a network*

# 3. Some Special Prob. distributions

**Continuous Uniform**

$$f(y) = \frac{1}{b-a} \quad \forall y \in (a,b)$$

The cdf is equal to:

$$F(x) = \begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b-a} & \text{if } a \leq x < b \\ 1 & \text{if } b \leq x \end{cases}$$

# 3. Some Special Prob. distributions

$$\mu = \frac{a+b}{2}, \qquad \sigma^2 = \frac{(b-a)^2}{12}$$

EXAMPLES:

- *Distance between source and destinations of messages in a network*
- *Seek time on a disk*

Comment: Specially interesting the Uniform in $(0,1)$. It alllows to generate observations from other prob. distributions using the **Inversion Method**

# 3. Some Special Prob. distributions

### Bernouïlli

A r.v. $Y \sim \mathrm{B}(p)$ (*Bernouïlli*), $0 \leq p \leq 1$ if and only if takes just values $0$ y $1$ with probabilities:

$$\Pr\{Y = 1\} = p \ y \ \Pr\{Y = 0\} = 1 - p.$$

$$\mu = p, \quad \sigma^2 = p(1-p)$$

EXAMPLES:

- *A computer system is up or down*
- *a packet in a computer network reaches or does not reach the destination*
- *a bit in the packet is afected by noise and arrives in error*

# 3. Some Special Prob. distributions

**Binomial**

A r.v $Y \sim \mathrm{Bin(n,p)}$ (*Binomial*) with parameters $n \in \mathbb{N}$ and $0 \leq p \leq 1$, if and only if takes values in $\{0, 1, 2, \cdots, n\}$ with probabilities:

$$\Pr\{Y = k\} = \binom{n}{k} p^k (1-p)^{n-k}, \quad \forall k \in \{0, 1, \cdots, n\}.$$

Is the distribution of the number of successes in $n$ **independent** trials with the **same probability** of succes.

# 3. Some Special Prob. distributions

$$\mu = n\,p \text{ and } \sigma^2 = n\,p\,(1-p).$$

EXAMPLES:

- *number of processors that are up in amultiprocessor system*
- *number of items in a batch that have certain characteristics*
- *number of bits in a packet that are not affected by noise*

Comment: If $n=1$ the Bernouilli distribution is obtained.

If $y$ is a realization of $Y$, $\hat{p} = y/n$.

# 3. Some Special Prob. distributions

## Multinomial

A r. vector $Y = (Y_1, Y_2, \cdots Y_N)$ follows a Multi$(N, p_1, p_2, \cdots, p_N)$ if and only if

$$P(Y = (y_1, y_2, \cdots y_N)) = \binom{N}{y_1, y_2, \cdots y_N} p_1^{y_1}\, p_2^{y_2} \;\cdots p_N^{y_N}$$

Important: $Y_i \sim Bin(N, p_i)$ and $corr(Y_i, Y_j) = -Np_ip_j$.

EXAMPLES:

- *Number of claims for each one of the $N$ types*
- *Type of web page visited for the $N$ considered types*

# 3. Some Special Prob. distributions

**Poisson**

A r.v. $Y \sim \mathrm{Po}(\lambda)$, $\lambda > 0$, if and only if takes values in $\mathbb{Z}^+$ with pmfs:

$$\text{Pr}\{Y = k\} = e^{-\lambda}\,\frac{\lambda^k}{k!}, \quad \forall k \in \mathbb{Z}^+.$$

$$\mu = \lambda, \quad \text{and} \quad \sigma^2 = \lambda$$

From where the **dispersion index**

$$I(Y) = \frac{Var(Y)}{E(Y)} = 1$$

# 3. Some Special Prob. distributions

The Poisson appears as limit of Binomial distributions

If $X_n \sim Bin(n, p_n)$ and it is verified that:

1) $n \to +\infty$
2) $p_n \to 0$ cuando $n \to +\infty$
3) $n \cdot p_n \to \lambda$ cuando $n \to +\infty$

then $X_n \to Y$, donde $Y \sim Po(\lambda)$.

Comment: That's why it is known as *law of rare events*

# 3. Some Special Prob. distributions

EXAMPLES:

- *Number of requests in a server for a given time interval*
- *Number of component failures per unit time*
- *Number of queries to a database system over t seconds*
- *Number of typing errors per form*

If $\overline{x}$ is the sample mean of a Poisson random sample, $\hat{\lambda} = \overline{x}$

# 3. Some Special Prob. distributions

**Exponential**

A r.v. $Y \sim \mathrm{Exp}(\lambda)$, $\lambda > 0$, if and only if takes values in $\mathbb{R}^+$ and

$$f(y) = \lambda e^{-\lambda\, y} \;\; \forall y \in (0, +\infty)$$

The cdf is equal to:

$$F(y) = 1 - e^{-\lambda\, y}$$

# 3. Some Special Prob. distributions

$$\mu = \frac{1}{\lambda}, \quad \sigma^2 = \frac{1}{\lambda^2}$$

EXAMPLES:

- *time between sccessive requests arrivals to a device*
- *time between failures of a device*

Comment: Very important in *queuing theory* since arrival times and waiting times are often assumed to be exponential.

# 3. Some Special Prob. distributions

## The Poisson related to other probability distributions

1) **Multinomial** If $Y = (Y_1, Y_2, \cdots, Y_n)$, $Y_i$ i.i.d $Y_i \sim Po(\lambda_i)$ then

$$Y | \sum_{i=1}^{n} Y_i = N \sim \text{Multinomial}(N, p_1, p_2, \cdots, p_n)$$

with $p_i = \frac{\lambda_i}{\sum_{i=1}^{n} \lambda_i}$.

2) **Exponencial** Fixed an interval of time $(0, t)$, if $T$ is the r.v. corresponding to the time between two consecutive independent successes, and $Y$ corresponds to the number of successes in $(0, t)$, one has

$$T \sim \text{Exponencial}(\lambda) \iff Y \sim \text{Po}(\lambda\, t)$$

# 3. Some Special Prob. distributions

If $X \sim Zipf(\alpha)$ its pmf is equal to:

$$P(X = x) = \begin{cases} \frac{x^{-\alpha}}{\zeta(\alpha)} & \text{if } x = 1, 2, 3, 4, \dots \\ 0 & \text{otherwise;} \end{cases}$$

where $\alpha > 1$ and $\zeta(\alpha)$ is the Riemann Zeta function,

$$\zeta(\alpha) = \sum_{k=1}^{\infty} k^{-\alpha}\,.$$

It is also known as: Discrete Pareto distribution, Riemann distribution or Power law distribution.

Zipf, G. K., (1932) "Selected studies of the principle of relative frequency in language", Cambridge: Harvard university Press

# 3. Some Special Prob. distributions

Main properties:

- Large probability at one.
- Skewed.
- Linear in the log-log scale: $\ln P(X = x) = -\alpha \ln x - \ln \zeta(\alpha)$.
- It is an exponential family.
- If $x_1, x_2, \cdots, x_m$ is a r. s. from $X$, the m.l.e. for $\alpha$ is obtained solving:

$$E(\log(X)) = m^{-1} \sum_{i=1}^{m} \log(x_i)$$

# 3. Some Special Prob. distributions

Applications

It is used to fit data corresponding to **frequencies of frequencies** or **ranked data**.

- In **Linguistics**: *number of occurrence of words in a given text*.
- In **Ecology**: *number os species per genus of mammals*.
- In **Bioinformatics**: *degree of proteins in a partially known protein-interaction network*.
- In **Social networking** *number of links in a given website* or the *number of visits in a given website*.

Observation: for many data sets, the m.l.e of $\alpha$, $\hat{\alpha}$ lies in $(0,1)$.

# 3. Some Special Prob. distributions

Suitable to fit **frequencies of frequencies** and **ranking data.**

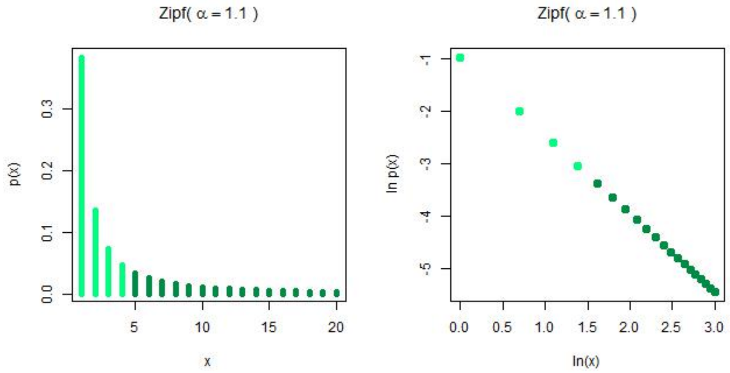
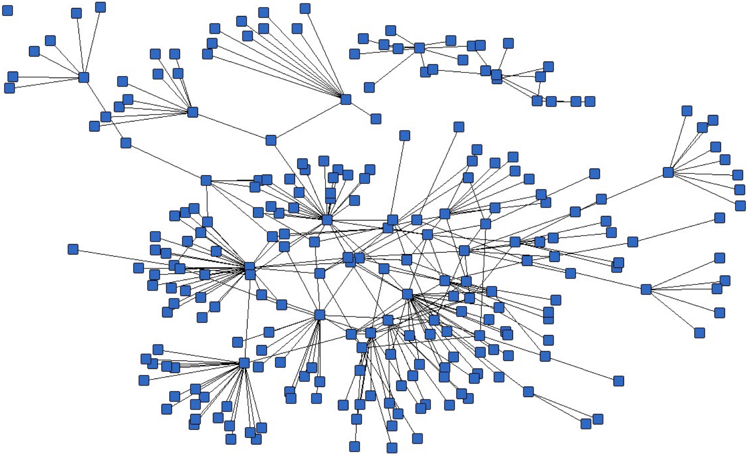


Figure : Zipf. distribution for $\alpha = 1.5$. On the right in the log-log scale.

# 3. Some Special Prob. distributions



# 3. Some Special Prob. distributions

## Weak points

From an **analytical** point of view:

- with just one parameter it is not flexible enough.

From a **practical** point of view, sometimes

- it is not able to adapt the empirical probability of the first values,
- the log-log linearity is only observed for values larger than $x_{min}$,
- the empirical distribution is not log-log linear in the tail.

# 4. Normal and related distrib.

**Normal o Gauss (N($\mu, \sigma^2$))**

$$f(y) = \frac{1}{\sigma\,\sqrt{2\,\Pi}} e^{\frac{-(y-\mu)^2}{2\,\sigma^2}} \quad y \in (-\infty, \infty)$$

$$\mu = \mu, \quad \sigma^2 = \sigma^2$$

When $\mu = 0$ and $\sigma^2 = 1$ one has the *standard normal distribution.*

# 4. Normal and related distrib.

EXAMPLES:

- *Errors in measurement in general*
- *Processing time*
- *Weight and Height of human beings*
- *Astronomical distances*

# 4. Normal and related distrib.

**Chi-square with $n$ degrees of fredom**

Defined as the sum of the square of $n$ independent $N(0,1)$ r.v's.

$$X^2 = X_1^2 + X_2^2 + \cdots X_n^2,$$

with $X_i \sim N(0,1)$.

$$\mu = n \quad \sigma^2 = 2\,n$$

EXAMPLE:

If $X \sim N(\mu, \sigma^2)$,then

$$\frac{(n-1)\,S^2}{\sigma^2} \sim \chi_n^2$$

# 4. Normal and related distrib.

**Students-t**

Defined as:

$$T = \frac{Z}{\sqrt{\chi^2/n}}$$

EXAMPLE: It is the distribution of the statistic used to compare two means of Normal distributed r.v.'s when the variances are equal and different

# 4. Normal and related distrib.

**F of Fisher-Snedecor**

Defined as:

$$F = \frac{X_1}{X_2}$$

where $X_i \sim \chi^2_{n_i}$ and $X_1$ and $X_2$ independent.

EXAMPLES: Useful two compare variances of two Normal distributed r.v's.

# 5. Central limit theorem

**Central limit Theorem**

If $x_i \; i = 1, \cdots, n$ is a random sample from a r.v. $X$ with mean$\mu$ and variance $\sigma^2$, then

$$\frac{\sqrt{n}(\overline{x} - \mu)}{\sigma} \sim_{aprox} N(0,1)$$

Comment: It is true for any $n$ if $X$ is Normal distributed. Otherwise, the $n$ value needs to be large enough.