TOPIC 1: MAIN PROBABILITY DISTRIBUTIONS

A) DISCRETE DISTRIBUTIONS

A.1. BINOMIAL

a) ***Binomial (1;p)***: $\xi \sim B(1; p)$

Characteristics:

- A random phenomenon occurring just once
- Two possible outcomes (dichotomy): A (yes) y Ā (no)
- Incompatible events
- One of those events will take place necessarily
- Formulation:

$$\mathrm{A} \to 1 \qquad P(\xi = 1) = p$$

$$\overline{\mathrm{A}} \to 0 \qquad P(\xi = 0) = q$$

$$p + q = 1$$

Probability distribution:

$$P(\xi = x) = p^x \cdot q^{1-x} \quad ; \; x = 0;1$$

Cumulative distribution function:

$$F(x) = P(\xi \leq x)$$

$$x < 0 \; ; \quad F(x) = 0$$

$$0 \leq x < 1; \quad F(x) = q$$

$$x \geq 1 \qquad F(x) = 1$$

Expected value: $E(\xi) = p$

Variance: $V(\xi) = p \cdot q$

Standard deviation: $\sigma = +\sqrt{p \cdot q}$

b) *Binomial (n;p)*: $\xi \sim B(n;p)$

Characteristics:

- Random phenomenon based on the realization "*n*" times of a dichotomous phenomenon.
- The trials are ***independent***, meaning that an outcome does not affect subsequent outcomes
- Interpretation: the number "*x*" of successes obtained from *n* possible dichotomic outcomes

Formulation of the DRV:

$\xi = \xi_1 + \xi_2 + \ldots + \xi_i + \ldots + \xi_n$

where $\xi \sim B(1;p)$ independent

Then $\xi \sim B(n;p)$

Probability distribution:

$$P(\xi = x) = \frac{n!}{x!\,(n-x)!} \cdot p^x \cdot q^{n-x}; \qquad x = 0,1,2,3 \ldots n$$

$$\sum_{x=0}^{n} P(\xi = x) = 1$$

Cumulative distribution function: $F(x) = P(\xi \leq x)$

$x < 0$; $F(x) = 0$

$0 \leq x < 1;$ $F(x) = P(\xi = 0)$

$1 \leq x < 2;$ $F(x) = P(\xi = 0) + P(\xi = 1)$

$2 \leq x < 3;$ $F(x) = P(\xi = 0) + P(\xi = 1) + P(\xi = 2)$

....

$x \geq n$ $F(x) = \sum_{x=0}^{n} P(\xi = x) = 1$

Expected value: $E(\xi) = n \cdot p$

Variance: $V(\xi) = n \cdot p \cdot q$

Standard deviation: $\sigma = +\sqrt{n \cdot p \cdot q}$

- Additive property:

• Let $\xi_1 \sim B(n_1;p)$ and $\xi_2 \sim B(n_2;p)$ independent

• If $\gamma = \xi_1 + \xi_2$ then $\gamma \sim B(n_1+n_2;p)$

A.2. POISSON $\xi \sim P(\lambda)$

Features:

Let $\xi \sim B(n;p)$

where $E(\xi) = np = \lambda$

$n \to \infty$

$p \to 0$

then ξ ~ P(λ) (Poisson of lambda parameter)

Approximation rule: $n \geq 30 \quad p \leq 0{,}1$

Another perspective: this distribution is useful to determine probabilities for random events occurring in continuous fixed intervals (of time and space)

- Those random events are dichotomic
- The process is stable meaning that, on the long term, an average number of events per unit of time or space occur
- Occurrences are independent. Therefore, the number of occurrences in a given unit is independent of the number of occurrences in any other nonoverlapping unit
- In this context, two situations can be considered:
  - The time elapsed between the occurrence of two consecutive events (exponential distribution)
  - The number of events happening in an interval of time (Poisson's distribution)

- Examples:

  The number of breakdowns, during a month, in certain machine

  The number of clients entering a branch from 10 am to 11 am

  The number of faulty articles in a shipment of 1000 units

  The number of claims in a life insurance company for a given day

  The number of phone calls received, per second, in an office

The number of landings in an airport, every 15 minutes

- It is also called "*the law of rare events*" because the interval of time, in which the occurrence of the events is studied, can be divided in small subintervals where the happening of such event is B(1;p) with p $\rightarrow$ 0

Probability distribution:

$$P(\xi = x) = \frac{\lambda^x}{x!} \cdot e^{-\lambda} \;\; ; \quad x = 0,1,2....$$

Cumulative distribution function: $F(x) = P(\xi \le x)$

$x < 0$; $F(x) = 0$

$0 \le x < 1$; $F(x) = P(\xi = 0)$

$1 \le x < 2$; $F(x) = P(\xi = 0) + P(\xi = 1)$

$2 \le x < 3$; $F(x) = P(\xi = 0) + P(\xi = 1) + P(\xi = 2)$

....

$x \ge n$ $\qquad F(x) = \sum_{x=0}^{n} P(\xi = x) = 1$

Expected value and variance:

$$E(\xi) = V(\xi) = \lambda$$

Standard deviation: $\sigma = +\sqrt{\lambda}$

Additive property:

Let $\xi_1 \sim P(\lambda_1)$ and $\xi_2 \sim P(\lambda_2)$ independent

If $\gamma = \xi_1 + \xi_2$ then $\gamma \sim P(\lambda_1+\lambda_2)$

B) CONTINUOUS DISTRIBUTIONS

B.1. *Uniform distribution*: $\xi \sim \mathrm{U}[a;b]$

- It resembles a random variable taking values *necessarily* inside a closed interval with extrems a and b, provided that the density function $f(x)$ is constant along such interval:

$$f(x) = \begin{Bmatrix} \dfrac{1}{b-a}; & a \le x \le b \\ 0\ ; & otherwise \end{Bmatrix}$$

Hence:

$$f(x) \ge 0;$$
$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

- The probability for the variable falling in a given subinterval [c;d] inside [a;b] depends on the length of such subinterval rather than on its position.

  Cumulative distribution function:

$$F(x) = \begin{Bmatrix} 0 & ; x < a \\ \dfrac{x-a}{b-a} & ; a \le x < b \\ 1 & ; x \ge b \end{Bmatrix}$$

Expected value: $$\mu = \frac{a+b}{2}$$

Variance:

$$\sigma^2 = \mu_2 - \mu^2 = \frac{(b-a)^2}{12};$$

$$\mu_2 = \frac{b^3 - a^3}{3\cdot(b-a)}$$

Standard deviation: $$\sigma = \sqrt{\frac{(b-a)^2}{12}}$$

B.2. *Normal distribution*

- It is the most important distribution in Statistics
- There is certain controversy in relation to the authorship of the discovery
- Some authors consider it was discovered by De Moivre in 1773 as an approximation to the B(n;p)
- But most concede this acknowledgement to Gauss, provided he was the first scientist in using the normal law to measure errors in experiments (1809)
- Laplace was also a key author, given that he presented among other things the central limit theorem (1812)
- The normal distribution approximates the probability distribution of many random variables, such as the B(n;p) and the Poisson
- Central Limit Theorem: if a rp is the result of a high number of independent random phenomenon, each of them quantitative small, then its probability distribution approximates a normal distribution

- Probability distributions of sample means approach a normal distribution when the sample size is high
- It is used in all the sciences requiring empirical experiments such as medicine, phisics, biology, economy, and so on.
- Some distributions arising from the normal distribuiton are Pearson's χ^2, Student's *t* and Snedecor's *F*
- We can distinguish the normal distribution and the standard normal distribution, the latter being a particular case of the former
- We will start with the standard normal distribution and follow with the general case

B.2.a) *Standard Normal distribution (0;1) or Z distribution:*

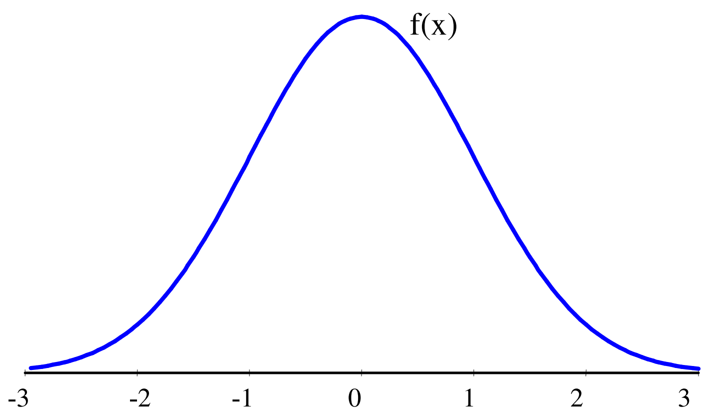
$$\xi \sim N[0;1]$$

- It is a normal distribution whose expected value equals cero and its standard deviation is one
- It has a bell-shaped density function whose formula is:

$$f(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2}}$$

- Most of the values fall around the mean
- It is simmetric about the mean
- Half of the values are below cero and half above cero
- Mean, median and mode are the same
- Probabilities are calculated by tables

- Its graph is called gaussian bell:

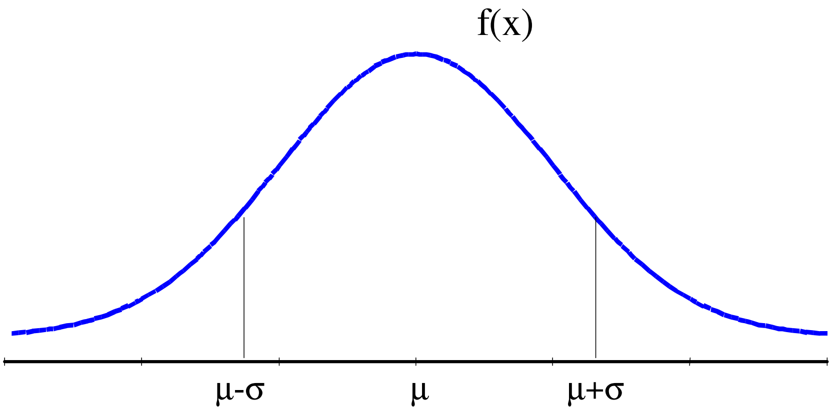


- Probabilities can be calculated using integrals, however tables are more convenient for that purpose.

B.2.b) ***Normal distribución (μ;σ):*** $\xi \sim N[\mu;\sigma]$

- It is a normal distribution whose expected value equals μ and its standard deviation is σ
- It has a bell-shaped density function whose formula is:

$$f(x)=\frac{1}{\sigma\cdot\sqrt{2\pi}}\cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



- The probability in [μ-σ;μ+σ] is about 68%
- The probability in [μ-2σ;μ+2σ] is about 95%
- The probability in [μ-3σ;μ+3σ] is almost 100%

- Any probability can be obtained by transforming the N(μ;σ) into the Z standard normal. This operation is called standardization:

$$\xi^* = \frac{\xi - \mu}{\sigma}$$

- Additive property:

  Let $\xi_i \sim N(\mu_i;\sigma_i)$ independent; $i = 1..n$

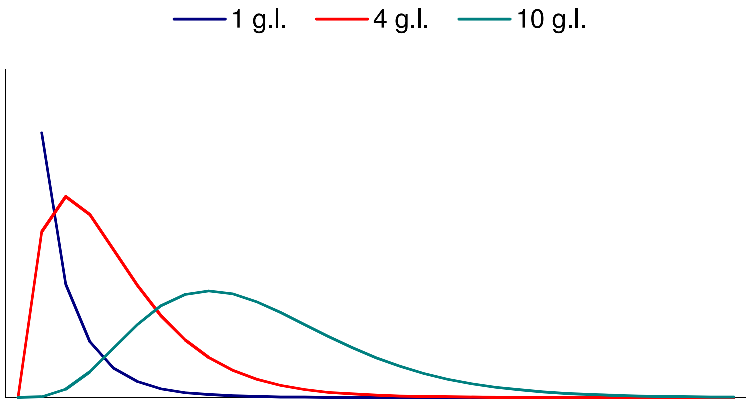
  $a, b_1, b_2, \ldots, b_n$ constants

  and $w = a + b_1\xi_1 + b_2\xi_2 + \ldots + b_n\xi_n$

Then:

$$w \sim N\left(a + \sum_{i=1}^{n} b_i\mu_i \; ; \; \sqrt{\sum_{i=1}^{n} b_i^2\sigma_i^2} \right)$$

B.3. *Pearson's Chi-square distribution*: χ^2_n



- It was developed by K. Pearson at the beginning of the 20th century
- There is not any event in the reality following this distribution

Definition:

$$Let\ \xi_i \sim N(0;1)\ independent\ with\ i = 1..n$$

Then:

$$\chi^2_n = \xi_1^2 + \xi_2^2 + \xi_3^2 + \cdots + \xi_n^2 = \sum_{i=1}^{n} \xi_i^2$$

- It only takes positive values
- Expected value and variance:

$$E(x) = n$$
$$V(x) = 2 \cdot n$$

- Its shape depends on the degrees of freedom (n).
- It presents a positive skewness, yet becoming simmetric when n increases.
- Additive property:

$$Let\ \ \xi_1 \sim \chi^2_n\ ;\ \ \xi_2 \sim \chi^2_m\ independent\ \ and\ \gamma = \xi_1 + \xi_2$$

Then:

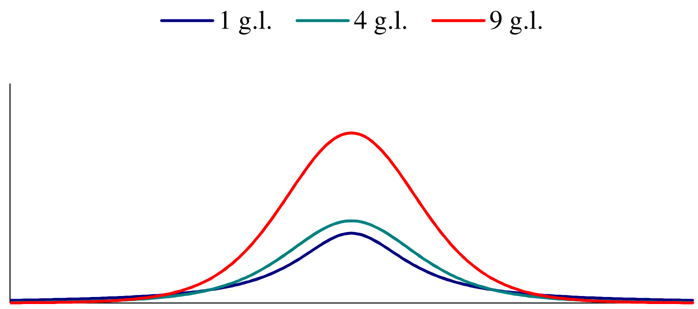
$$\gamma \sim \chi^2_r\ \ where\ r = n + m$$

- It can be proved that:

$$If\ \ \xi \sim \chi^2_n \rightarrow Y = \sqrt{2\xi} \sim N\left(\sqrt{2n-1}; 1\right)$$

- It will be useful in estimating the variance of normal variables

B.4. *t Student distribution*: $\xi \sim t_n$



- It was obtained by W.S. Gosset in 1908
- It does not correspond to any situation coming from the real life

Definition:

$$Let\ \xi_i \sim N(0;\sigma)\ independent\ with\ i = 0..n$$

Then:

$$t_n = \frac{\xi_0}{\sqrt{\dfrac{\xi_1^2 + \xi_2^2 + \xi_3^2 + \cdots + \xi_n^2}{n}}} = \frac{Z}{\sqrt{\dfrac{\chi_n^2}{n}}}$$

Expected value and variance (being $n > 2$):

$$E(x) = 0$$

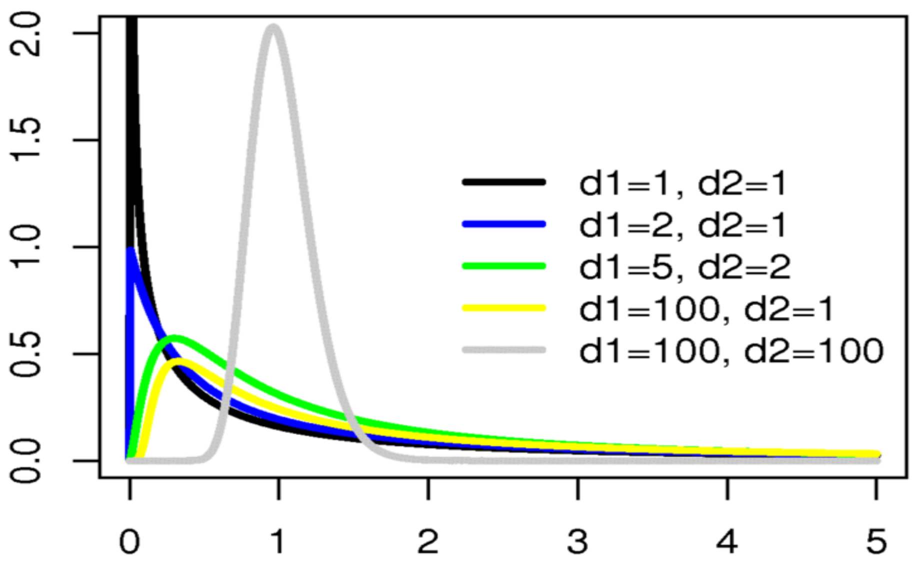
$$V(x) = \frac{n}{n-2}$$

- Its density function is simmetric and bell shaped
- It presents more dispersión than Z, but it lowers when n increases.

- It will be appropriate in estimating population means from normal variables with unknown variances
- It approaches Z when n increases:

$$Whenever\ \ n \to \infty\ \ then\ t_n \sim N(0;1)$$

B.5. *Fisher-Snedecor F distribution*: $\xi \sim F_{m,n}$



Definition:

$$Let\ \xi_i \sim N(0;\sigma)\ independent\ with\ i = 1..m$$

$$Let\ \gamma_j \sim N(0;\sigma)\ independent\ with\ j = 1..n$$

Then:

$$F_{m,n} = \frac{\dfrac{\xi_1^2 + \xi_2^2 + \xi_3^2 + \cdots + \xi_m^2}{m}}{\dfrac{\gamma_1^2 + \gamma_2^2 + \gamma_3^2 + \cdots + \gamma_n^2}{n}} = \frac{\dfrac{\chi_m^2}{m}}{\dfrac{\chi_n^2}{n}}$$

- It does not fit any phenomenon in the real life
- It does not depend on the variance of the variables involved in the definition
- It has got a tail on the right
- It can only take positive values
- If m = 1 then $F(1;n) = t^2(n)$
- If a variable X follows an $F(m;n)$ then $1/X$ follows an $F(n;m)$
- It will be useful in order to compare variances from two normal distributions

CENTRAL LIMIT THEOREM

Let ξ be a variable arisen from a set of n random variables not necessarily having the same probability distribution:

$$\xi = \xi_1 + \xi_2 + \xi_3 + \ldots + \xi_n$$

Verifying that:

1. The set of random variables are independent
2. Each of them with mean (μ_i) and finite variance (σ_i^2)
3. n being big enough (n > 30)

Then:

$$\xi \sim N\left(\sum_{i=1}^{n} \mu_i \;;\; \sqrt{\sum_{i=1}^{n} \sigma_i^2}\right)$$