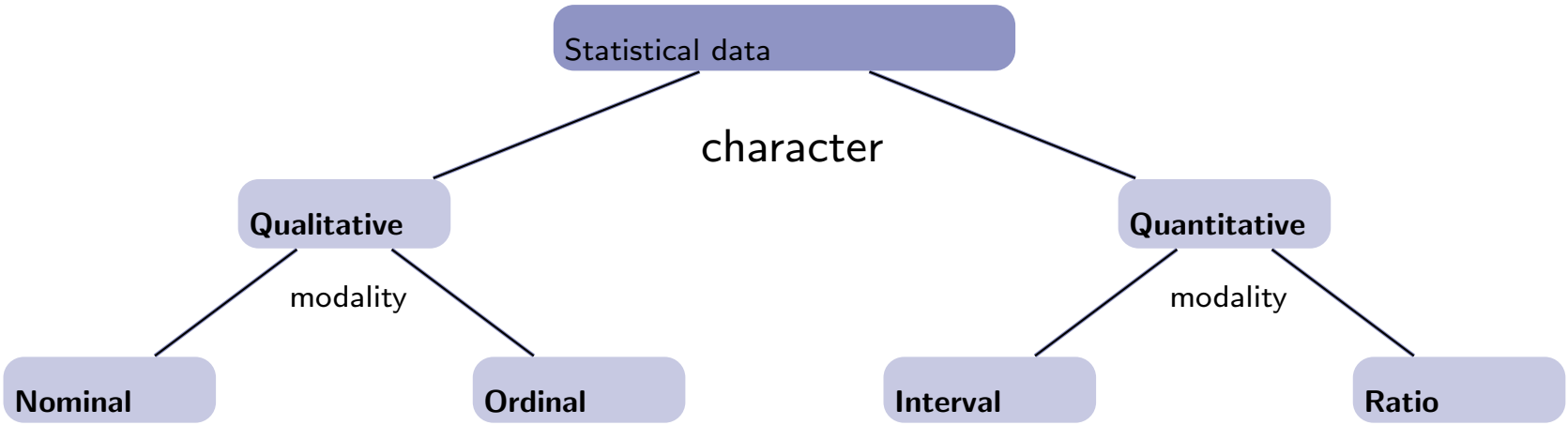
Statistics for Nominal Data

During the statistical analysis, data features have to be considered carefully.

The experimental phase in which information is collected is a key point: both description methods and applicable tests rely on experimental phase.

Depending on the kind of information that can be collected, **Data types** or **Data scales** can be classified into four classes.

Statistical universe



Then, four types of scale exist: **Qualitative Nominal**, **Qualitative Ordinal**, **Quantitative Interval**, **Quantitative Ratio**

This course pertains only to the analysis of **Qualitative** data.

Depending on the kind of data, different statistical options are available.

In general, on Qualitative data fewer statistical options are available than on quantitative data.

For example, on Quantitative data one might calculate the average, whereas, on qualitative nominal data this couldn't be done.

Again, on Nominal data fewer statistical options are available than on Ordinal data.

Ordinal data can be treated as Nominal; the converse is not true.

Quantitative data can be always considered as Ordinal; the converse is not true.

A **random variable** (or **stochastic variable**) is a variable whose values are associated to a probability law. When a variable is not associated to a probability law it is called **deterministic variable**.

The **random variables** are split into two main categories that are treated in different ways:

1. **Discrete random variables**
   that are usually used to model qualitative (nominal or ordinal) data frequencies.
2. **Continuous random variables**
   expressed on a continuous scale of values. Values are often associated to a proper **unit of measurement**.

The use of **discrete** or **continuous** random variables to model random phenomena is not always clearly defined, and it can depend from the number of **categories** associated to the variable.

Discrete variables have less explicative power than continuous ones:

- ⋆ statistics computed on continuous variables contains more information;
- ⋆ to get good quality statistics using discrete variables a greater **number of observations** is required.

Data should be collected using the more accurate scale. When possible, data should be collected on a continuous scale.

Anyway, qualitative data (or their frequencies) are almost more often modeled by discrete random variables.

For the aim of this course, the data will be considered modelable as discrete random variables only.

Values of **frequencies** produced on the categories of a data series can be:

- ⋆ **absolute**: the counts of observed data values in each category.
- ⋆ **relative**: the absolute frequency divided by the total count of observations
- ⋆ **cumulative (absolute or relative)** : the (absolute or relative) frequency of each category added to the sum of frequencies of lower categories. It requires ordered data.

The following example shows an excerpt of 50 evaluations of defective items produced by the three machines: M01, M02 e M03.

Five possibile defect levels (categories) for each article are taken: A, B, C, D and E.

The defect levels are sorted in ascending order: level A is less serious than level B, etc..

M01	A	B	B	D	A	A	A	E	A	D	A	B	A	....	A
M02	A	E	A	A	B	A	B	D	A	B	A	A	A	....	A
M03	A	A	D	A	B	A	A	B	A	D	A	B	A	....	B

Considering the aggregate data on all machines, the table below reports the number of defective articles per defect type:

Level category	**A**	**B**	**C**	**D**	**E**
Absolute Freq	69	43	18	9	11
Relative Freq (%)	0.46	0.29	0.12	0.06	0.07
Cumulative Abs Freq*	69	112	130	139	150
Cumulative Rel Freq *	0.46	0.75	0.87	0.93	1.00

* only when the categories are ordered.

The **mode** is the value (category) that occurs most frequently in a data series.

When a phenomenon is qualitative, the value of the mode is almost the only statistics (other than frequency table) always available to summarize data.

Also, the Mode is the only statistics that may be calculated either on Nominal or on Ordinal data.

Frequency distributions with one mode only are called **unimodal distributions** while distributions with two or more modes are called **bimodal or multimodal distributions**.

The **median** is the category that separates the "higher half" of a data series from the "lower half".

It can be found by arranging all the observations from lowest value to highest value and picking the middle one.

For its use, at least **ordinal (or quantitative) data** are required.

To **calculate the median of a data series** it needs:

- ⋆ arranging all the observations from lowest value to highest value (or from the highest value to lowest value) and counting the total number n of observations;
- ⋆ if n is odd, the median is the category that correnspond to the middle observation, i.e. the modality of observation in the position $\boldsymbol{\frac{n+1}{2}}$;
- ⋆ if n is even, and the $\frac{n}{2}$-th and $\frac{n+1}{2}$-th observations belong to the same category, then this category is the median;
- ⋆ if n is even, and the $\frac{n}{2}$-th and $\frac{n+1}{2}$-th observations do not belong to the same category, then the median is undefined (or is between the two adjacent categories).

When the data are of ordinal type, one might calculate the **Rank** values on data series.

The ranks are numerical sequences associated to categories, such that the ordering on numerical values is the same of the ordering on categories.

For example: if a *Weight* ordinal character is used, with three categories (*Low*, *Medium*, *High*), one might associate the numerical values (ranks) 1, 2, 3, respectively, to *Low*, *Medium*, *High* categories.

In these cases, sometimes the average of ranks is calculated as a summary statistics.

Note that this use of Average is questionable and makes sense only in specific cases.

A **contingency table** is essentially a display format used to analyse and record the relationship between frequencies of two or more categorical variables.

Character A	Character B		
	B1	B2	Margin
A1	10%	1%	11%
A2	7%	81%	89%
Margin	17%	83%	100%

Character A	Character B		
	B1	B2	Margin
A1	15	2	17
A2	11	123	134
Margin	26	125	151

The contingency table, along with the frequency table, can also be used to check if two or more phenomena can be considered independent.

The data of above example on Machines and Defect levels can be represented in a two-ways contingency table:

Machine	Defect level					Tot
	A	B	C	D	E	
M01	20	15	5	7	3	50
M02	24	16	5	1	4	50
M03	25	12	8	1	4	50
Tot	69	43	18	9	11	150

As a table of column percents

Machine	Defect level					Tot
	A	B	C	D	E	
M01	20.0%	34.9%	27.8%	77.8%	27.2%	33.3%
M02	34.8%	37.2%	27.8%	11.1%	36.4%	33.3%
M03	36.2%	27.9%	44.4%	11.1%	36.4%	33.3%
Tot	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

As a table of row percents

Machine	Defect level					Tot
	A	B	C	D	E	
M01	40.0%	30.0%	10.0%	14.0%	6.0%	100.0%
M02	48.0%	32.0%	10.0%	2.0%	8.0%	100.0%
M03	50.0%	24.0%	16.0%	2.0%	8.0%	100.0%
Tot	46.0%	28.7%	12.00%	6.00%	7.3%	100.00%

As a table of total percents

Machine	Defect level					Tot
	A	B	C	D	E	
M01	13.3%	10.0%	3.3%	4.7%	2.0%	33.3%
M02	16.0%	10.7%	3.3%	0.7%	2.7%	33.3%
M03	16.7%	8.0%	5.3%	0.7%	2.7%	33.3%
Tot	46.0%	28.7%	12.0%	6.0%	7.3%	100.0%

Question: do all the tables make sense in this specific example?

The **Bernoulli distribution** is perhaps the simplest discrete probability distribution.

The Bernoulli distribution describes the distribution of a random phenomenon that may show only two possible outcomes (events).

The two events are "labeled" with numeric values **0** and **1**.

Usually, the event "1" is also named "success".

This distribution states that the probability of having 1 ("success") is p $(0 \leq p \leq 1)$, and the probability of having 0 is $1 - p$.

The probability function of Bernoulli distribution is then stated as:

$$f(x;p) = \begin{cases} p & \text{if } x = 1 \\ 1-p & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases}$$

That can also be expressed as

$$f(x;p) = p^x(1-p)^{1-x} \text{ for } x \in 0, 1$$

The Expected value ("mean") and Variance of a Bernoulli distribution are, respectively:

$$E\{X\} = 0 \cdot (1-p) + 1 \cdot p = p$$

$$V\{X\} = E\{(X - E\{X\})^2\} = (0-p)^2 \cdot (1-p) + (1-p)^2 \cdot p = p \cdot (1-p)$$

Example:

100 balls are contained in a bag:

- ⋆ 80 balls are red
- ⋆ 20 balls are white

Question: What is the probability function about the extraction of a white ball ("success") from the bag in only one trial?

Answer: Since it is known that the white balls are 20 on a total count of 100, the probability of randomly drawing 1 white ball is $p = \frac{20}{100} = 0.2$, and then the probability function may be stated as:

$$f(x; p = 0.2) = 0.2^x 0.8^{1-x} \text{ for } x \in 0, 1$$

where $x = 1$ means "Draw a white ball"

□

Example:

A production machine has yielded 1300 defective units on a total of 200000 units produced in the past.

Question: Hypothesizing that the production machine is working always in same manner, what is the probability function that describes the probability of obtaining defective/safe unit drawing randomly from the process?

Answer: If the production process is "stable", we could expect that the defective rate for the machine is 1300/200000; then, the probability of having a defective unit is $p = 0.0065$, and then the probability function may be stated as:

$$f(x; p = 0.0065) = 0.0065^x 0.9935^{1-x} \text{ for } x \in 0, 1$$

Where $x = 1$ means "Draw a defective unit"

□

Now, next paragraphs will show some relevant distributions that may be thought deriving from the Bernoulli distribution.

The **binomial distribution** is the discrete probability distribution of the number of "successes" in a sequence of n independent "success"/"failure" (or yes/no) experiments, each of which yields "success" with probability p.

$$f(x; n, p) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \text{ for } x \in 0, \cdots, n$$

where:

$x =$ number of successes

$p =$ probability of success in a trial

$n =$ number of trials

The Expected value ("mean") and Variance of a Binomial distribution are, respectively:

$$E\{X\} = \sum_{x=0}^{n} x \cdot \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} = n \cdot p$$

$$V\{X\} = \sum_{x=0}^{n} (x - np)^2 \cdot \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} = n \cdot p \cdot (1-p)$$

Example:

A coin is perfectly balanced, and we want to know the probability of having exactly 8 "heads" on 10 coin flips

Question: What is the probability of having exactly 8 "heads" on 10 coin flips?

Answer: Since the coin is perfectly balanced, $p = 1/2$, and then the probability is:

$$f(x = 8; n = 10, p = 1/2) = \frac{10!}{8! \cdot 2!} \cdot \left(\frac{1}{2}\right)^8 \cdot \left(\frac{1}{2}\right)^2 = 0.04$$

□

Example:

A production process yields matchboxes with 100 matches each. The probability that a match does not work (defective) is 1%. Each match is independent from the others.

Question: What is the probability that a box contains no defective matches?

Answer: The defective rate for the process is 1/100; then, the probability of having a defective unit is $p = 0.01$, and then the probability of having 0 defectives in a box is:

$$f(x = 0; n = 100, p = 0.01) = \frac{100!}{100! \cdot 0!} \cdot 0.01^0 \cdot 0.99^{100} = 0.366$$

Where $x =$ *Number of defective units* (or number of successes)

□

Example:

A production process yields matchboxes with 100 matches each. The probability that a match does not work (defective) is 1%. Each match is independent from the others. Customers require that each box contains at most 2 defective matches

Question: Is the process compliant to customer requirements?

Answer: The defective rate for the process is 1/100; then, the probability of having a defective unit is $p = 0.01$, and then the probability of having at most 2 defective matches is:

$$f(x = 0; n = 100, p = 0.01) + f(x = 1; n = 100, p = 0.01) +$$

$$+ f(x = 2; n = 100, p = 0.01) \simeq 0.366 + 0.370 + 0.185 = 0.921$$

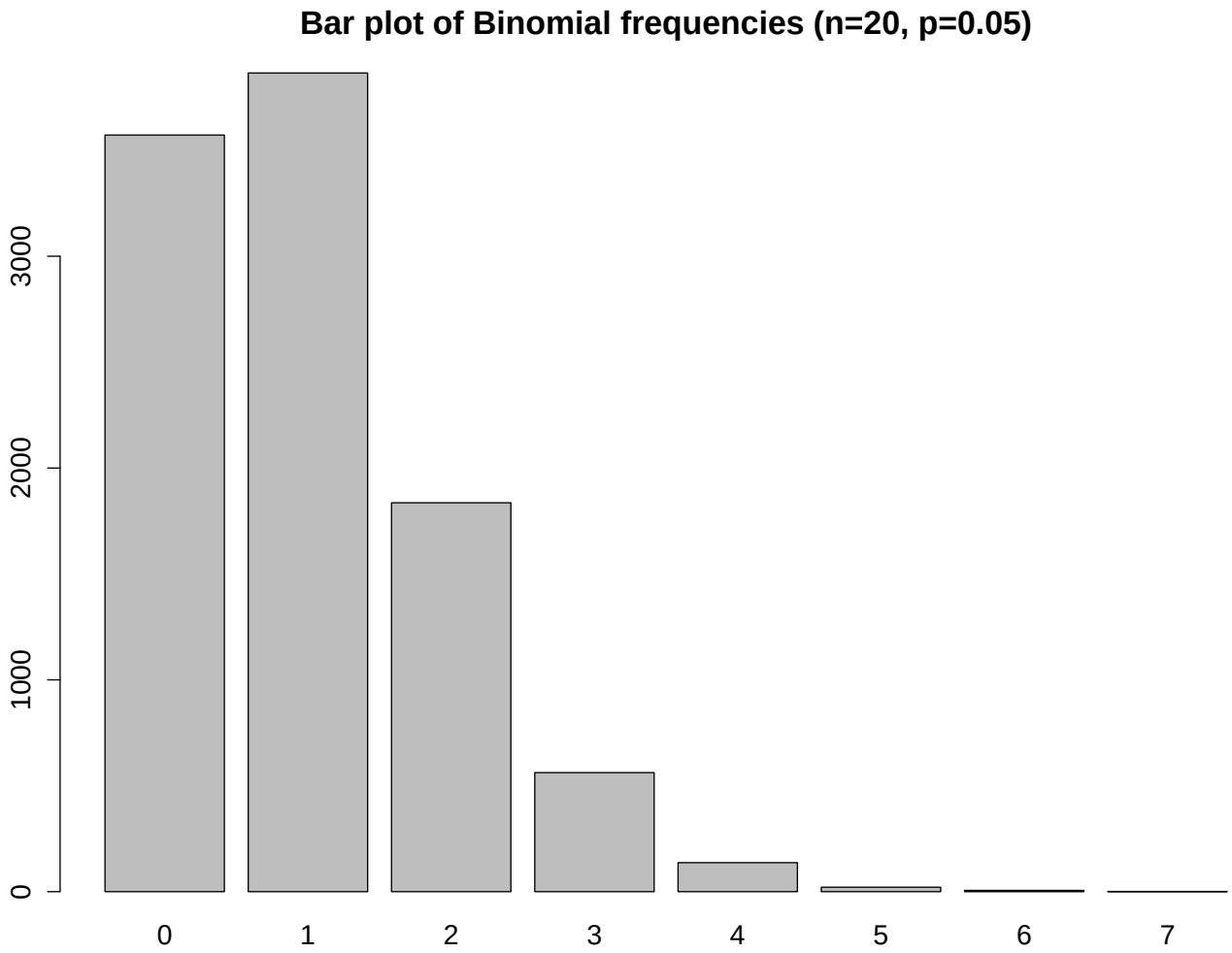
Where $x =$ *Number of defective units* (or number of successes).

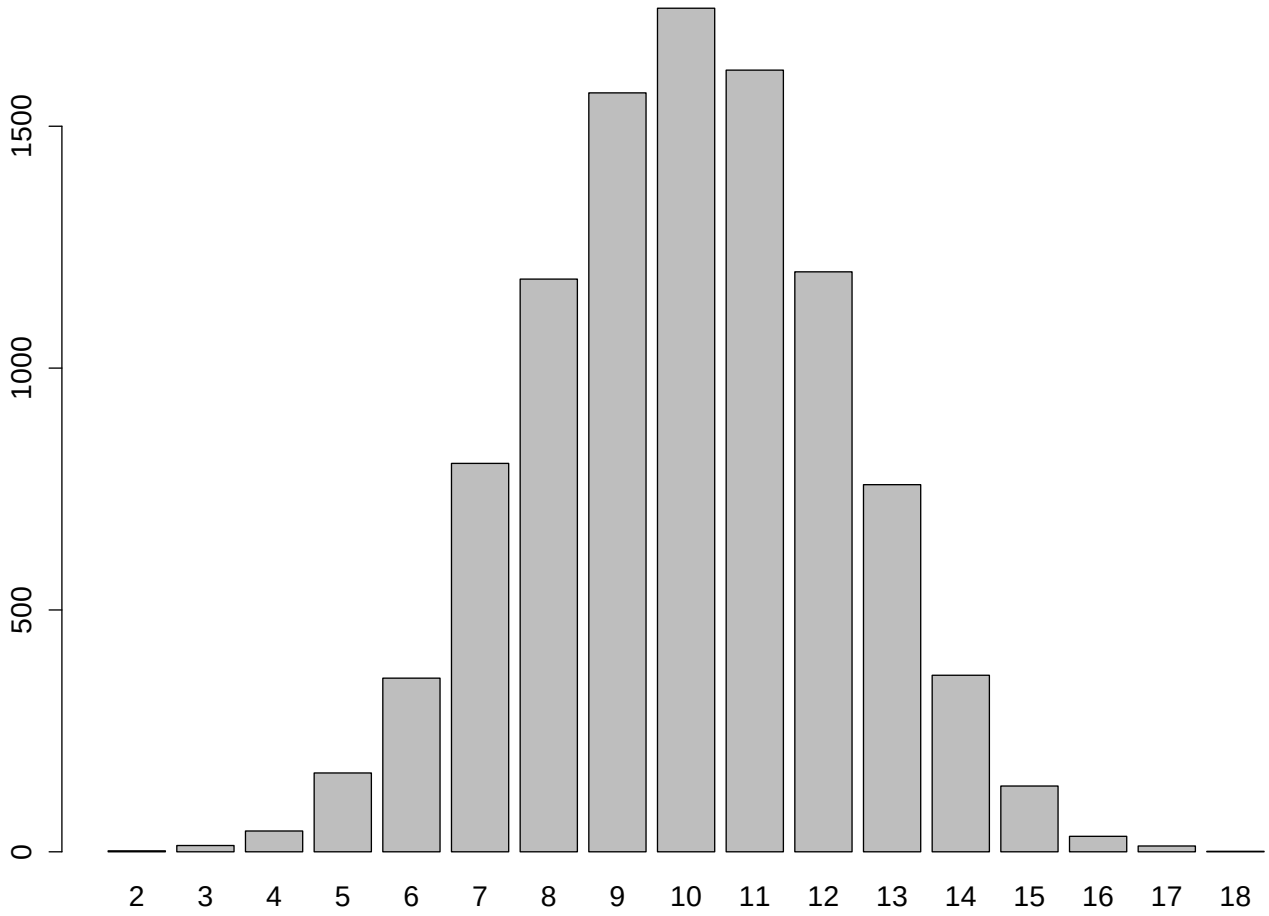
Then the expected fraction of compliant boxes is about 0.921 (or 92.1%).

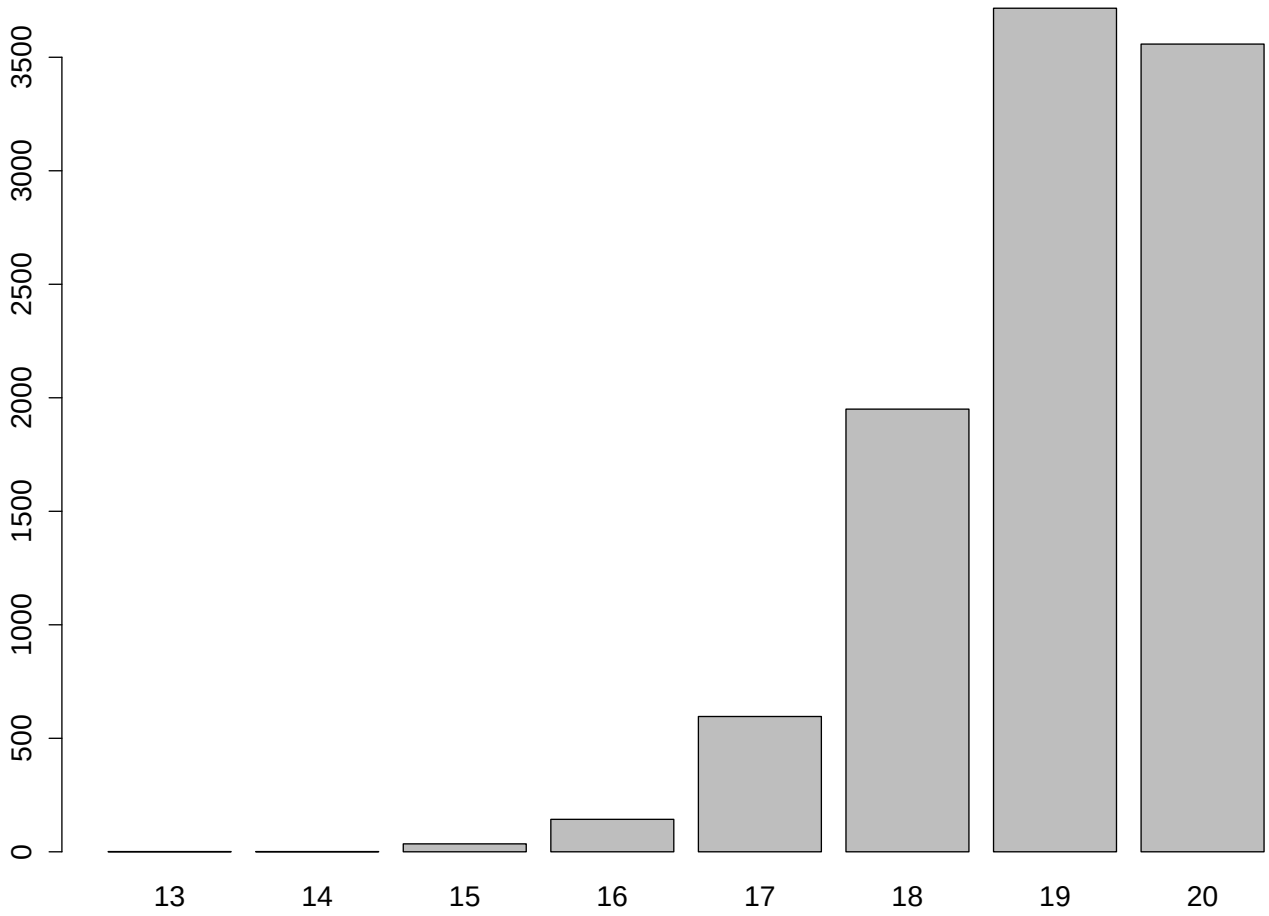
□

- The Bernoulli distribution is a special case of the Binomial distribution, where $n = 1$.
  Conversely, any binomial distribution, $B(n, p)$, is the sum of n independent Bernoulli trials each with the same "success" probability p.

- If p and $\boldsymbol{q = 1 - p}$ are equal to 0.5, the distribution is always symmetrical. If p is much larger or much smaller than q (then p and q are very different), then the distribution is asymmetric. But the asymmetry decreases with the growth of n.


Bar plot of Binomial frequencies (n=20, p=0.05)
0
1000
2000
3000
0
1
2
3
4
5
6
7


Bar plot of Binomial frequencies (n=20, p=0.5)
0
500
1000
1500
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18


Bar plot of Binomial frequencies (n=20, p=0.95)
0
500
1000
1500
2000
2500
3000
3500
13
14
15
16
17
18
19
20

The Poisson is a theoretical discrete distribution, and it is completely defined by only one parameter, μ.

The Poisson distribution is usually (but NOT always) used to model the counts of rare random events, and its probability distribution function is:

$$f(x;\mu) = \frac{\mu^x}{x!}\, e^{-\mu} \text{ for } x \in 0, 1, \cdots, \infty$$

The main "conceptual" difference between Poisson and Binomial distribution is that the count values (x) in Poisson distribution are virtually unlimited, whereas in Binomial distributions are limited to the number of trials (n).

It may be shown that the Expected value ("mean") and Variance of a Poisson distribution are, respectively:

$$E\{X\} = \sum_{x=0}^{\infty} x \cdot \frac{\mu^x}{x!} e^{-\mu} = \mu$$

$$V\{X\} = \sum_{x=0}^{\infty} (x-\mu)^2 \cdot \frac{\mu^x}{x!} e^{-\mu} = \mu$$

Thus, for the Poisson distribution, expected value (mean) and variance are identical, and they equal the μ parameter.

Example:

In a LCD display production process, the count of defects for each display is recorded. The historical average of defects counts for all the displays is 0.45. It is known that the count of defects for each display is modelable by a Poisson distribution.

Question: What is the probability of having at most 1 defect in a randomly chosen display?

Answer: The historical mean is 0.45, and then $\mu = 0.45$. The probability of having at most 1 defect in one display, then, is:

$$f(x = 0; \mu = 0.45) + f(x = 1; \mu = 0.45) \simeq 0.638 + 0.287 = 0.925$$

Where $x =$ *Number of defects*

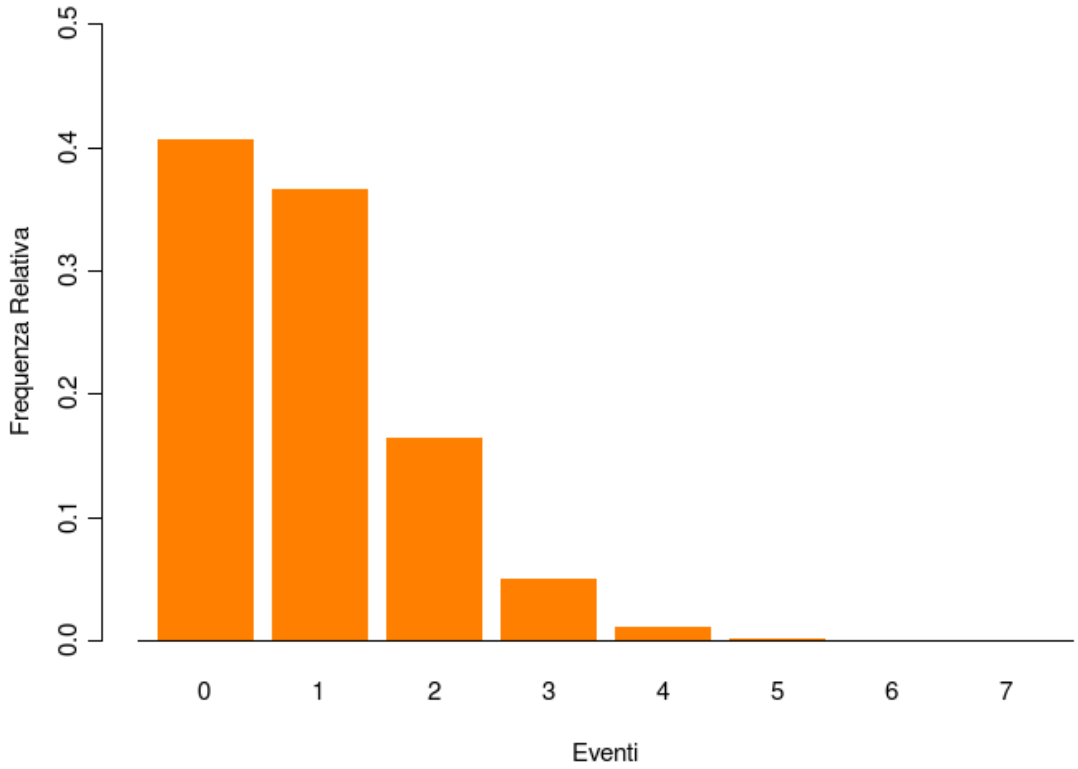
□

Note: $1 - 0.925 = 0.075$ in this case may be seen as the fraction of displays with more than one defect.

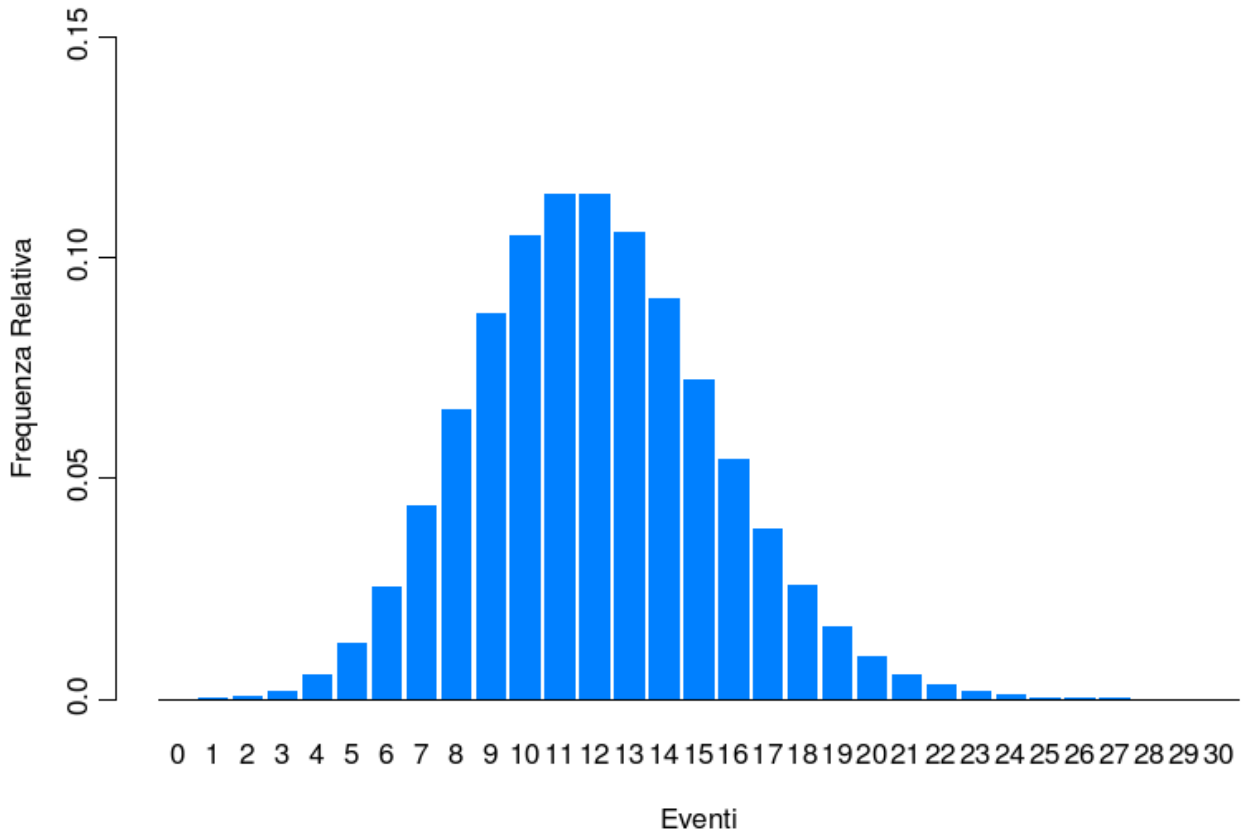
Notes:

- The Poisson distribution is a limiting case of a Binomial distribution, when $n \to \infty$, and the mean of Binomial distribution $(n \cdot p)$ remains finite $(n \cdot p = \mu)$.
- That's because the Poisson distribution may be used as an approximation of Binomial distribution when $p < 0.05$ and $n > 100$.
- The Poisson distribution may be used for events that occur either in space or in time; for example: the number of particles in a given space, or the number of arrivals in a given time span.
- Sometimes, the parameter is named λ instead of μ.
- The Poisson distribution is very (right) skewed when $\mu < 1$, is still asymmetric for $\mu < 3$, and is nearly symmetric and may be approximated by the normal distribution for μ greater than 6.

This graph shows the Poisson distribution with $\mu = 0.9$. It is right skewed.



This graph shows the Poisson distribution with $\mu = 12$. It is "almost Normal".



The **Multinomial** distribution is a generalization of Binomial distribution. The Multinomial distribution, instead of represent phenomena with only a couple of possible outcomes (TRUE/FALSE, YES/NO, ...), is used to model *m*-uple of mutually exclusive categories (HIGH/MEDIUM/LOW, Green/Blue/Brown/Grey Eyes, ...). The Multinomial distribution has the following probability density function:

$$f(x_1, ..., x_m; n, p_1, ...p_m)) = \frac{n!}{x_1!x_2!...x_m!} {p_1}^{x_1} {p_2}^{x_2} ...{p_m}^{x_m}$$

where:

x_i = number of events ("successes") for the $i-$th $(i = 1, \cdots, m)$ category $(x_1 + x_2 + \cdots + x_m = n)$

p_i = probability of success in only one trial for the $i-$th category

m = total number of possible categories

n = total number of trials

Example:

A perfectly balanced dice is thrown 8 times.

Question: What is the probability of having 2 times "1", 2 times "3", 1 times "4", and 3 times "6" ?

Answer: The probability may be calculated as

$$f(x_1 = 2, x_2 = 0, x_3 = 2, x_4 = 1, x_5 = 0, x_6 = 3;$$
$$n = 8, p_1 = \tfrac{1}{6}, p_2 = \tfrac{1}{6}, p_3 = \tfrac{1}{6}, p_4 = \tfrac{1}{6}, p_5 = \tfrac{1}{6}, p_6 = \tfrac{1}{6}) =$$
$$= \tfrac{8!}{2!0!2!1!0!3!} (\tfrac{1}{6})^2 (\tfrac{1}{6})^0 (\tfrac{1}{6})^2 (\tfrac{1}{6})^1 (\tfrac{1}{6})^0 (\tfrac{1}{6})^3 = 0.001$$

□

The problem of **statistical inference** arises when the researcher obtained only a subset of overall available data (the statistical population), and he/she wants to draw decisions about the overall population.

In production set-up, these needs arise several times and in very different fields.

Example.

- ⋆ A factory produces switches. One researcher wants to test if the switches produced in his plant meet well some requirements on some destructive tests. The target of study is to know the fraction of switches that don't meet the requirements. Of course, in that case testing all switches is impossible.
- ⋆ The researcher will collect a sample of switches to assess with a given level of precision which percentage of switches will be defective in all the production process.
- ⋆ After that, the researcher would also test if the fraction of defective units found on the sample is compatible with the hypothesis that the true (population) percentage of defective units fall within pre-specified parameters.

The statistical inference techniques give guidelines and formulas to obtain this kind of information.

The statistical inference techniques can be splitted in three main subgroups:

1. Point estimation techniques;
2. Confidence intervals evaluation;
3. Statistical hypothesis tests.

The techniques at the first point are usually specific of the specific problem and give intuitive enough results. The other two points are less intuitive, but they use conceptually similar techniques.

The point estimation problem is usually a trivial one, when categorical data are analyzed.

Examples:

- ⋆ In a Bernoully or Binomial framework, the probability of success (p) may be estimated with the relative frequency of success.
- ⋆ In case of Multinomial data, the probability of having each of the categories is estimable by the relative frequency too.
- ⋆ In case of Poisson model, the value of μ parameter may be estimated simply with the average of counts.

The problem of statistical tests and confidence interval, on the contrary, is more complex and requires more in-depth examination.

Often the researcher wants to test if the analyzed phenomenon meets specific requirements.

Examples:

- ⋆ In a production process, the reasearcher could need to test if the percentage of defective units is no greater than a specific level.
- ⋆ In medical statistics, one might want to test if the fraction of medical patients that heal after the use of a drug is equal to some percentage.
- ⋆ In banking industry, one may want to check if two groups of customers have the same level of bankrupt percentage.

The researcher often can draw only a sub-sample of the overall population, and he/she have to draw conclusions about the overall population.

All the above example assertions could be stated in mathematical form, and they are examples of so-called **Null hypoteses** ($\boldsymbol{H_0}$).

The null hypothesis is the main hypothesis stated by the researcher about a phenomenon to be tested.

To conduct a statistical test, the Null hypothesis must be tested against an alternative hypothesis.

The alternative hypothesis is usually called $\boldsymbol{H_1}$ or $\boldsymbol{H_A}$.

$\boldsymbol{H_0}$ and $\boldsymbol{H_A}$ must be mutually exclusive and exhaustive.

Statistical tests give methods and criteria to assess if $\boldsymbol{H_0}$ or $\boldsymbol{H_A}$ is true, based on sample data, with a pre-assigned error probability.

Suppose that a sample of n observations from the population has been drawn to test the "success" probability.

If the number of observed "successes" is denoted with x, and each sample unit has a "success" (e.g., percentage of defective units) probability p and it is independent from the others, then, prior to sample extraction, x will be a Binomial random variable, with parameters n and p.

The researcher could want to test the hypothesis that the true (and unknown) proportion p of "successes" is equal to p_0.
This is the null hypothesis, and it is stated as: $\boldsymbol{H_0 : p = p_0}$

The alternative hypothesis $\boldsymbol{H_A}$ could be that the true (population) proportion $\boldsymbol{p}$ of "successes" is different from $\boldsymbol{p_0}$. In mathematical form: $\boldsymbol{H_A : p \neq p_0}$

This results to an **hypothesis system**:

$H_0 : p = p_0$,

$H_A : p \neq p_0$.

Denoting with the symbol $\hat{p} = \frac{x}{n}$ the observed proportion of "successes" within the sample, the simplest **test statistics** that allow to test H_0 against H_A is the following:

$$\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

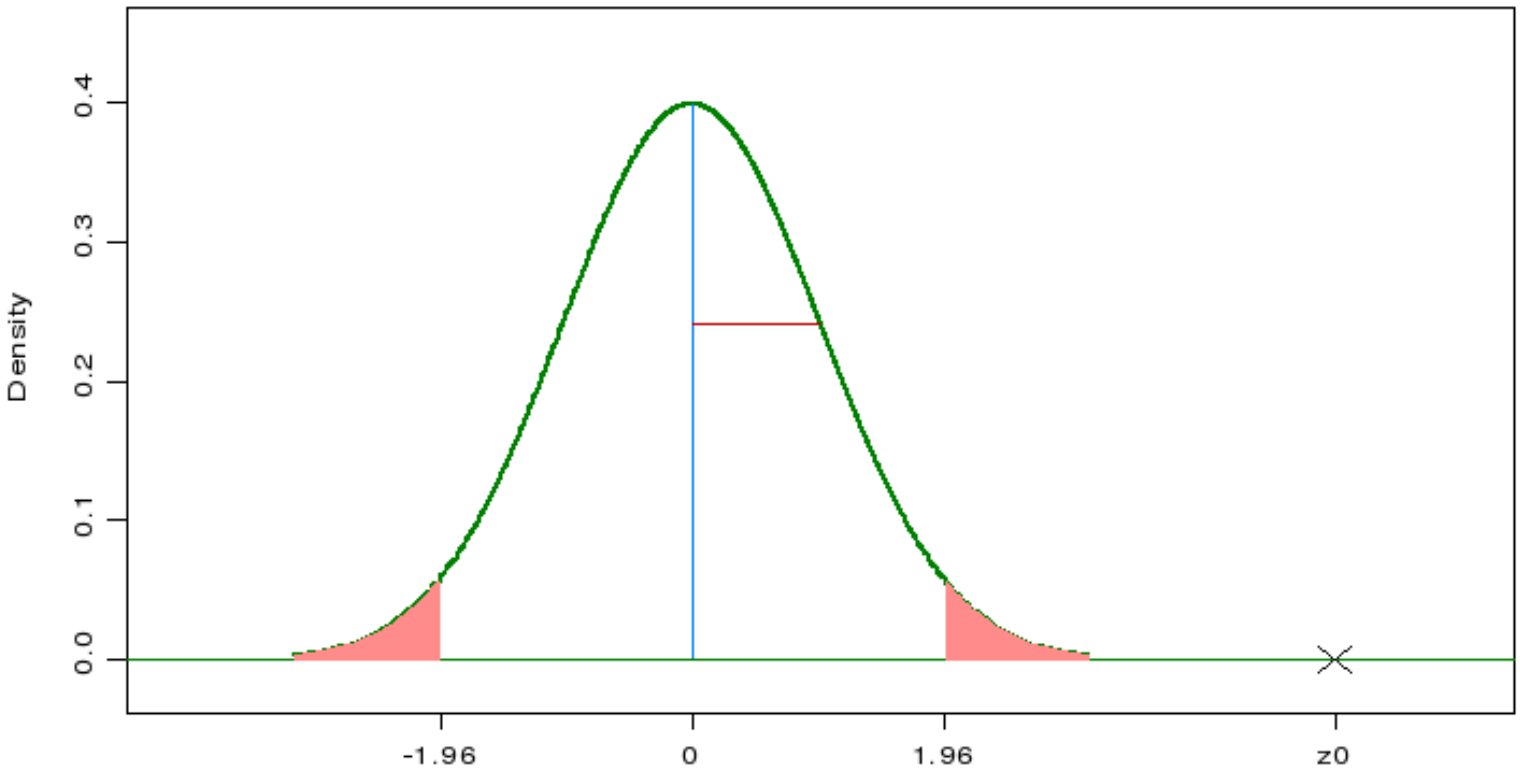
If H_0 is true, and n is sufficently large, then the above quantity (prior conducting the sample extraction) will be distributed as a standard Normal distribution.

If H_A is true, and n is sufficently large, then the above quantity will be distributed as a normal distribution centered "away from zero".

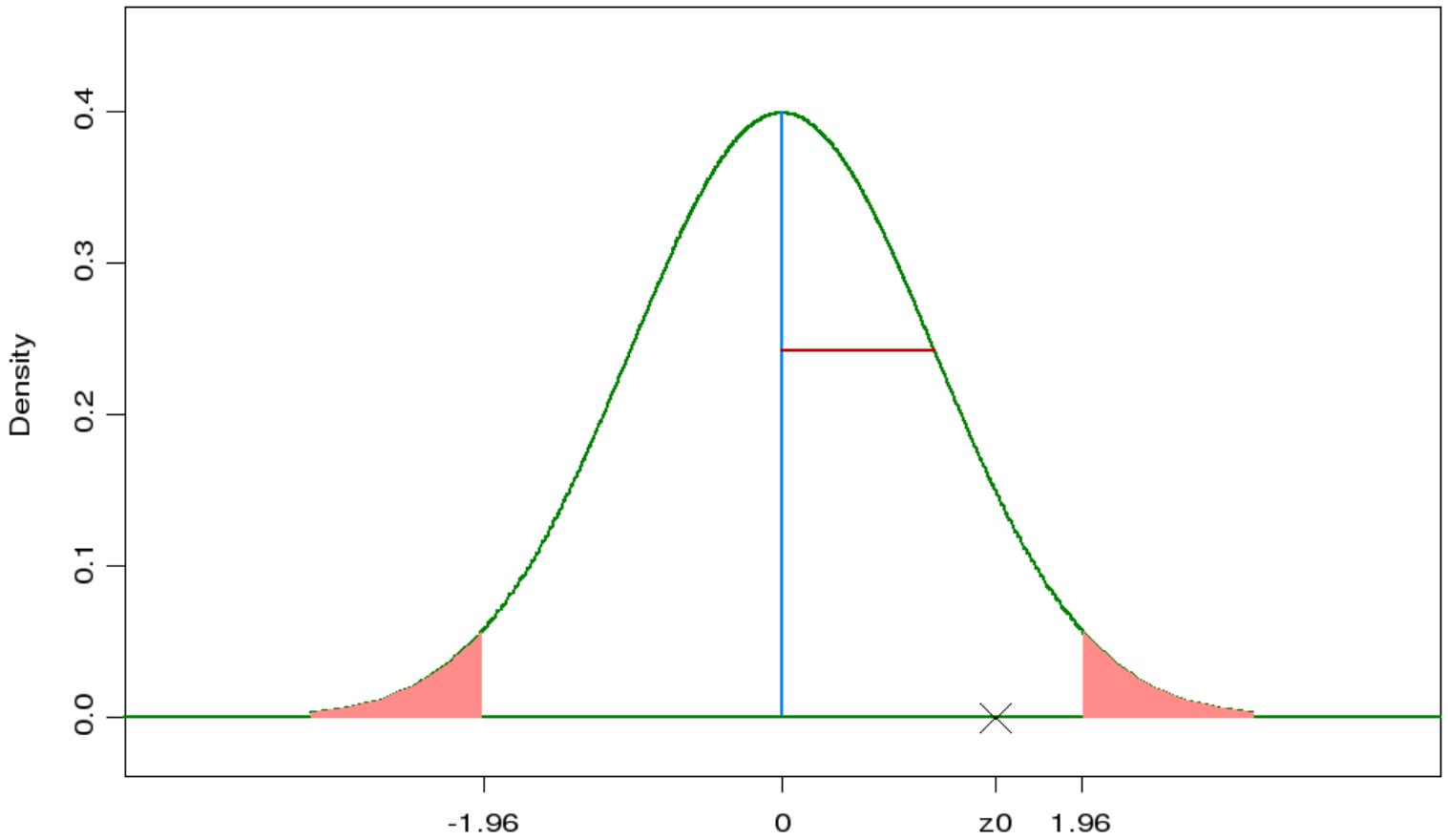
- If H_0 is true, then one expects that the value of above formula, after the experiment, will lie between the values -1.96 and 1.96 with probability $0.95 (= 1 - \alpha)$.

- If, conversely, H_A is true, then one expects that the value of above formula, after the experiment, will lie away from 0.

- Then, a couple of thresholds could be selected, at -1.96 and 1.96:
  - if the value of test statistics falls between these limits, then H_0 is accepted (or, equivalently, H_A is rejected);
  - if the value of test statistics falls outside these limits, then H_0 is rejected (or, equivalently, H_A is accepted).



Reject the Null hypothesis (H_0), and accept H_A.



Accept the Null hypothesis (H_0), and reject H_A.

IMPORTANT NOTES:

- ⋆ Given the above thresholds, H_0 may be erroneously rejected, when H_0 is true, with probability 0.05.
- ⋆ Changing the threshold values, the probability (α) of erroneously rejecting H_0 changes too.
- ⋆ The α value (0.05 in above example) is called **significancy level**.
- ⋆ The $1 - \alpha$ value (0.95 in above example) is called **confidence level**.
- ⋆ The test presented is said a **two-sided test**, because H_A states that p is different from p_0. If H_A states that p is strictly greater (or less) than p_0, the test is said **one-sided test**, and the threshold to accept/reject H_0 changes, to obtain the same α level.

Example.

The management of a DVD manifacture company needs to test the hypothesis that the fraction of defectives units in the process is 0.01 (=1%).
The significancy level at which the test is performed is $\alpha = 0.05$.

With that aim, a sample of 100 units is randomly drawn from the process. After functional tests, only two units appear defective.

To check the null hypothesis $H_0 : p = 0.01$ against the alternative hypothesis $H_A : p \neq 0.01$ the above test statistics is used:

$$\frac{(2/100) - 0.01}{\sqrt{0.01(1-0.01)/100}} = 0.01/\sqrt{0.000099} = 1.005$$

Since the absolute value of 1.005 is less than the threshold value 1.96, then one can say that **there's no empirical evidence to reject the null hypothesis**.

The same conclusion may be drawn by reading the p-value: the probability (calculated on the standard Normal distribution) of having a value of test statistics outside the interval given by -1.005 and 1.005. This probability is 0.315.
Since the p-value is greater than the chosen α value, the null hypothesis is accepted.

IMPORTANT NOTE: the p-value is the usual manner of reading tests results.

An $(1-\alpha)$ **confidence interval** for the "true" fraction of "successes" may be generated, following the same guidelines used for the statistical test. The confidence interval is given by:

$$\left(\hat{p} - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}; \hat{p} + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

If $\alpha = 0.05$, then $z_{1-\frac{\alpha}{2}} = 1.96$.

Example (continue).

The "true" fraction of defective DVDs, at a 0.95 confidence level, with the above example data, is:

$$\left(\frac{2}{100} - 1.96 \cdot \sqrt{\frac{0.02(1-0.02)}{100}}; \frac{2}{100} + 1.96 \cdot \sqrt{\frac{0.02(1-0.02)}{100}}\right)$$

i.e., the interval $(-0.0074; 0.0474)$.

IMPORTANT NOTES:

- ⋆ The methods shown to construct statistical tests or confidence intervals for one proportion are the simplest ones, and rely on asymptotic behavior of statistcal test when n is "big".
- ⋆ An exact version of statistical test and a better approximation of confidence intervals are available.
- ⋆ The α value is a conventional one. A different value of α may be chosen, with different error probabilities. The most used α values are 0.1, 0.05, 0.01, 0.001, with the second and third ones more frequently used.
- ⋆ Confidence interval and statistical test usually give the same information: if the test accepts H_0, the confidence interval "encloses" the hypothesized value p_0, and viceversa.

If one wants to compare the fraction of "successes" (p_A and p_B) between two independent sub-populations starting from some sample data, the statistical test that he/she should use is the 2 Proportions Test.
The general form of 2 proportion test null hypothesis states that the difference of proportion of "successes" in two sub-populations are equal to a given value δ_0. The hypothesis system, then, may be stated as:

$H_0 : p_A - p_B = \delta_0$
$H_A : p_A - p_B \neq \delta_0$
And in its simplest form, the 2 proportions test is:

$$\frac{(\hat{p}_A - \hat{p}_B) - \delta_0}{\sqrt{\hat{p}_A(1-\hat{p}_A)/n_A + \hat{p}_B(1-\hat{p}_B)/n_B}}$$

If H_0 is true, and for n_A and n_B large enough, the distribution of test statistics is approximately a standard Normal.

Example:

Suppose that two production machines, A and B, produced, in one working day, respectively 200 and 500 units. Among the units, 3 units of machine A and 4 units of machine B were flawed.

Question: Do the two machines produces the same proportion of defective units ($H_0 : \delta_0 = 0$), or the two machines produce different proportions of defective units ($H_A : \delta_0 \neq 0$)? The requested error probability is $\alpha = 0.05$.

Answer: By using the above formula, the results are:

$$\frac{(3/200) - (4/500)}{\sqrt{(3/200)(1-(3/200))/200 + (4/500)(1-(4/500))/500}} = 0.73$$

Since the absolute value of above quantity (0.73) is less than the threshold value of a standard normal distribution for $\alpha = 0.05$, (1.96), then one may state that **there's no empirical evidence that allows to reject the Null hypothesis**.

The same conclusions may be drawn by reading the p-value (the probability of having a value "external" to the values -0.73 and 0.73 on a standard normal distribution). This probability is 0.465.
Since the p-value is greater than the fixed α value, the null hypothesis is accepted.

NOTES:

- ⋆ Several "forms" of two proportions test statistics are available; the proposed form is conceptually the simplest.

- ⋆ As for the one proportion test, for the two proportions test the one-sided alternative hypothesis is available. The threshold value, and the p-value, can be obtained in a similar manner than for one proportion test.

The preceding tests may be applied when the analyzed phenomenon has a Bernoulli or a Binomial distribution, or when the researcher is mostly interested to one specific category of analyzed dimension.

When the phenomenon can produce more than two outcome categories, and the researcher is interested in all (or more than only one) categories, then the one-proportion or the two-proportions test may be inadequate.

Suppose that the observed phenomenon may assume m distinct values (categories), and that the researcher needs to test, based on a observation sample, if the m individual categories have a "success" probability equal to specific values $p_{1_0}, \cdots, p_{m_0}$, where $\sum_{1=1}^{m} p_{i_0} = 1$.

This is a case of Multinomial distribution

The hypothesis test system, then, may be stated as:

$$H_0 : p_1 = p_{1_0}, \cdots, p_m = p_{m_0}$$
$$H_A : \exists i \ (i \in 1, \cdots, m) \mid p_i \neq p_{i_0}$$

In other words:

- H_0 states that the probability of success of each category is equal to its pre-specified value;
- H_A states that the probability of success for at least one category is different from its pre-specified value.

To test this hypothesis system, given a sample of n independent observations, one must:

- ⋆ Calculate the frequency table for the observed phenomenon, obtaining the **Observed frequency counts** $O_1, \cdots O_m$
- ⋆ Calculate the **Expected frequency values** as $E_1 = n \cdot p_{1_0}, \cdots, E_m = n \cdot p_{m_0}$
- ⋆ Calculate the following (Pearson) **Chi-square statistics**:

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

If H_0 is true, then one expects (prior the experiment) that the χ^2 statistics tends to be distributed (as n increases) as a well-known χ^2_{m-1} distribution, and that its value will be relatively small.

If H_0 is false (and then H_A is true), then one expects (prior the experiment) that the χ^2 statistics **shall not** be distributed as a χ^2_{m-1} distribution, and that its value will be relatively large.

An upper threshold on the χ^2_{m-1} distribution may then be chosen so that the probability of obtaining a χ^2 test value greater or equal than the threshold is α.

Alternatively, one may calculate the probability on the χ^2_{m-1} distribution to give a value greater or equal than the observed χ^2 test value, obtaining the p-value.

If, after the experiment, the observed χ^2 value is greater than the threshold value, or (equivalently) if the calculated p-value is less than the pre-assigned α level, then one must reject the null hypothesis (H_0), and accept the alternative hypotesis (H_A).

If, on the contrary, the observed χ^2 value is less than the threshold value, or (equivalently) if the calculated p-value is greater than the pre-assigned α level, then one must accept the null hypothesis (H_0), and reject the alternative hypotesis (H_A).

Example: Suppose, in the above example on defects per machine, that the researcher expects that the incidence of defects follows the percentages:

$$H_0 : p_{A_0} = 50\%, p_{B_0} = 25\%, p_{C_0} = 12\%, p_{D_0} = 8\%, p_{E_0} = 5\%.$$

The researcher needs to test (at a $\alpha = 0.05$ level) if the above sample confirm this rule or, conversely, if the sample pushes to reject this hypothesis. The observed table of defects frequency counts is:

A	B	C	D	E	Tot
69	43	18	9	11	150

In this example, m=5.

Example (continued): The expected frequencies, based on expected percentages are:

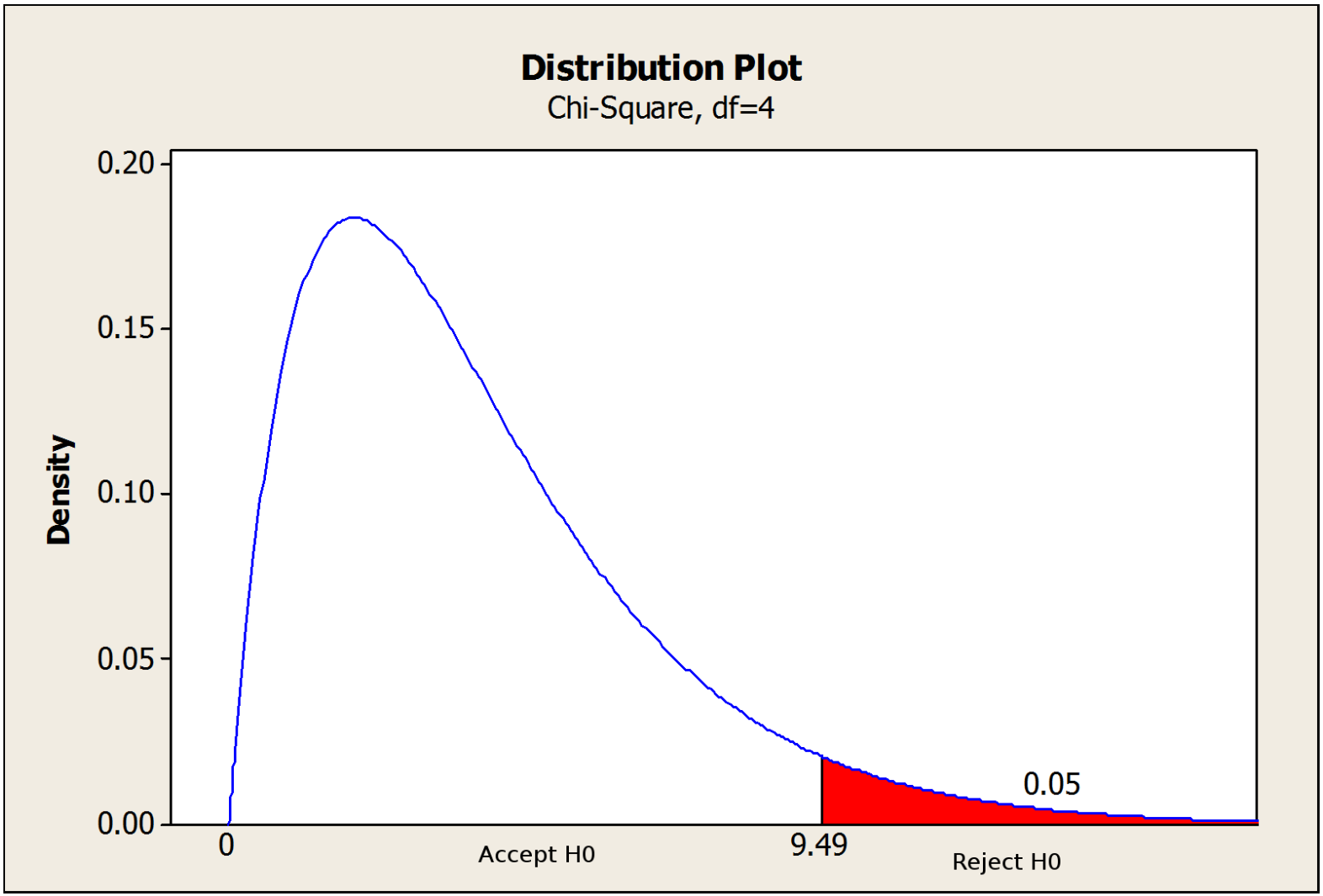
A	B	C	D	E	Tot
75	37.5	18	12	7.5	150

The calculation of χ^2 test statistics gives:

$$\chi^2 = \frac{(69-75)^2}{75} + \frac{(43-37.5)^2}{37.5} + \cdots + \frac{(11-7.5)^2}{7.5} \simeq 3.67$$

If H_0 is true, this quantity should come from a $\chi^2_{m-1} = \chi^2_4$ distribution.

Example (continued):



Example (continued):

Since the threshold to accept/reject H_0 for a χ^2_4 distribution is 9.49, and the test statistic is less than this threshold, the researcher must accept the H_0 hypothesis.

In other words, **there is no evidence in data to reject the null hypothesis** and to accept the alternative hypotesis.

Alternatively, the researcher could calculate the p-value, i.e., the probability on a χ^2_4 distribution of having values greater or equal to the obtained statistics (=3.67).

In this case, the p-value is 0.453, greater than α (=0.05).

□

IMPORTANT NOTES:

- The Chi-square Observed Vs Expected is an asymptotic test; that means that it needs sufficiently large sample sizes (n: total number of units tested) to be applicable.
- The expected cell count (i.e., the expected number of events for the individual categories) is also expected being greater than 5 to obtain "good" asymptotic results.
- When the sample size is too small, the test tends to be less "precise", i.e., it tends to reject or accept the null hypothesis with greater probability than expected.
- See also the indications for the test of next paragraph.
- **DO NOT confuse the χ^2 test statistics with the χ^2 distribution!**

This test may be seen as an "extension" of preceding two tests.

It tests **if the categories of two qualitative dimensions measured on the same units are associated or not**.

To exemplify this test, a simplification of above example on production machines is used.

Suppose then that a sample of n units coming from 3 production machines (A, B, C) has been tested to check for the presence of defective units. The researcher needs to check if the number of (or the percentage of) defective changes with the production machine.

In other words, the researcher needs to test if there is an association between production Machine and Defects.

Suppose that a sample of 430 units were drawn from production process, and that the data were summarized as in table below.

		Units state	
		Safe	Defective
Machine	A	132	13
	B	99	18
	C	157	11

The table can be represented in percentage form, and then two alternative (but equivalent) hypotheses on these percentages could be made

Each machine has the same percentage of defective units.

		Units state	
		Safe	Defective
Machine	A	91.0%	9%
	B	84.6%	15.4%
	C	93.5%	6.5%

The percentage of machines within defective units and within safe units is constant.

		Units state	
		Safe	Defective
Machine	A	34.0%	31.0%
	B	25.5%	42.9%
	C	40.5%	26.2%

Each of two above statements can be equivalently expressed as:

H_0: No association exists between Machines and Defectives.

Against the alternative hypothesis:

H_A: An association exists between Machines and Defectives (or, "The defective percentage is different for at least one machine" or "At least one machine has a percentage within defective units and within safe units different from the others").

NOTE: If no association exists between Machine and Defectives, then one expects that the within row percentages (and within columns percentages) are constant and equal to the columns marginal percentages (and row marginal percentages).

In more general form: suppose that two qualitative characteristics where read on n sampled units. The first characteristic (Q1) has r categories, while the second one (Q2) has c categories. The researcher needs to test if the distribution of two characteristics **is not associated** within the overall population.

The following contingency table may be built:

	Q2			
Q1	n_{11}	$\cdots$	n_{1c}	$n_{1.}$
	$\cdots$	n_{ij}	$\cdots$	$\cdots$
	n_{r1}	$\cdots$	n_{rc}	$n_{r.}$
	$n_{.1}$	$\cdots$	$n_{.c}$	n

where n_{ij} is the number of units with i-th category of Q1 and j-th category of Q2; $n_{i.}$ is the i-th row total, and $n_{.j}$ is the j-th column total.

If we denote with p_{ij} the relative frequencies within columns $(p_{ij} = n_{ij}/n_{.j})$, then, if no association exists between characteristics, one expects that: $p_{i1} = p_{i2} = \cdots = p_{ic} = p_{i.}$, $(i = 1, \ldots, r)$, where $p_{i.} = n_{i.}/n$

In other words, the relative frequencies calculated within the columns are all equal, and are equal to the marginal relative frequencies.

The same sentence may be stated by calculating the relative frequencies within the rows: $p_{1j} = p_{2j} = \cdots = p_{rj} = p_{.c}$, $(j = 1, \ldots, c)$, where $p_{.j} = n_{.j}/n$

As stated above, the two forms are "equivalent".

Taking the first of two forms, if absolutely no association exists between the two qualitative characteristics, then one expects that:

$$p_{ij} = p_{i.} \Longleftrightarrow n_{ij}/n_{.j} = n_{i.}/n \Longleftrightarrow n_{ij} = (n_{i.} \cdot n_{.j})/n$$

The last equality is the mathematical condition because no association exists between the two characteristics.

Then, given a two-way contingency table, the above formula allows to assess what shall be the frequencies values within the cells if "perfect independence" (i.e. loss of association) exists between the two characteristics: $n^*_{ij} \, (= (n_{i.} \cdot n_{.j})/n)$

The (Pearson) **Chi-square test**, then, uses the following formula to test if no association exists:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - n^*_{ij})^2}{n^*_{ij}}$$

If H_0 is true, then one expects (prior the experiment) that the χ^2 statistics tends to be distributed (as n increases) as a $\chi^2_{(c-1)\cdot(r-1)}$ distribution, and that its value will be relatively small.

If H_0 is false (and then H_A is true), then one expects (prior the experiment) that the χ^2 statistics **shall not** be distributed as a $\chi^2_{(c-1)\cdot(r-1)}$ distribution, and that its value will be relatively large.

An upper threshold on the $\chi^2_{(c-1)\cdot(r-1)}$ distribution may then be chosen so that the probability of obtaining a χ^2 test value greater or equal than the threshold is α.

Alternatively, one may calculate the probability on the $\chi^2_{(c-1)\cdot(r-1)}$ distribution to give a value greater or equal than the observed χ^2 test value, obtaining the p-value.

If, after the experiment, the observed χ^2 value is greater than the threshold value, or (equivalently) if the calculated p-value is less than the pre-assigned α level, then one must reject the null hypothesis (H_0), and accept the alternative hypotesis (H_A).

If, on the contrary, the observed χ^2 value is less than the threshold value, or (equivalently) if the calculated p-value is greater than the pre-assigned α level, then one must accept the null hypothesis (H_0), and reject the alternative hypotesis (H_A)

NOTES:

- ⋆ The above result is asymptotic; this means that it needs large sample sizes ($n > 100$).
- ⋆ Particularly, statistics literature suggests to not having expected frequency counts (n^*_{ij}) less than 5.
- ⋆ Some authors say that at most 2 or 3 cells are allowed with expected count less than 5.

Let's come back to the above example. The observed counts are:

		Units state		
		Safe	Defective	Tot
Machine	A	132	13	145
	B	99	18	117
	C	157	11	168
	Tot	388	42	430

While the expected counts are:

		Units state	
		Safe	Defective
Machine	A	130.84	14.16
	B	105.57	11.43
	C	151.59	16.41

The test statistic value: 6.271 $Pr\{\chi^2_2 > 6.271\} = 0.043$

When the sample size (n) is small, or when many expected cell counts are less than 5, the Chi-square test statistics may produce unreliable results.

Some literature suggests to use the so-called Yate's correction, which reduces the absolute value of each difference between observed and expected frequencies by 0.5 before squaring.

But its use is a matter of discussion between statisticians.

When the the contingency table is a 2x2 table, then an exact Chi-square test is available: the Fisher Chi-square test.

Suppose to have a 2 by 2 contingency table as the following:

		Units state	
		Safe	Defective
Machine	A	a	b
	B	c	d

Where a, b, c, d are cell counts, and $a + b + c + d = n$.

Fisher showed that the probability of obtaining any such set of values, given marginal counts, is given by:

$$\frac{(a+b)!(c+d)!(a+d)!(b+d)!}{a!b!c!d!n!}$$

In order to test if association **does not exist** (H_0) between the two characteristics, the total probability of observing data 'as extreme' or more extreme when the null hypothesis is true must be calculated.

The above probability for all these tables must be calculated, and add them together.

This gives a one-tailed test.

For a two-tailed test we must also consider tables that are equally extreme but in the opposite direction.

Unfortunately, classification of the tables according to whether or not they are 'as extreme' is problematic, but almost all the software packages give an option to calculate these values, and then the Fisher test.

When the comparison of proportions with some reference values, or when two proportions should be compared, then the 1 proportion or the 2 proportions test (respectively) can be applied.

The above χ^2 tests might be seen as an alternative or a generalization of 1 proportion or 2 proportions tests, with two-sided alternative hypothesis .

Consider, for example, the 1 proportion test example:

$H_0 : p = 0.01$

$H_A : p \neq 0.01$

The above test may be "translated" in a "Chi-square Observed Vs Expected" problem, rewriting it in following contingency table:

	O_i	E_i
Defective	2	1
Safe	98	99

In that case, the Chi-square test gives:

$$\chi^2 = \frac{(2-1)^2}{1} + \frac{(98-99)^2}{99} \simeq 1.01010$$

With a p-value, calculated on a χ^2_1 distribution, of $0.315 > 0.05 (= \alpha)$, very close to the above (1 proportion) result.

If we consider the 2 proportions test example, the test is:

$H_0 : p_A = p_B$

$H_A : p_A \neq p_B$

If the sample data are rewritten in tabular form:

	Machine A	Machine B
Flawed	3	4
Safe	197	496

$$\chi^2 = \frac{(3-2)^2}{2} + \frac{(4-5)^2}{5} + \frac{(197-198)^2}{198} + \frac{(496-495)^2}{495} \simeq 0.707$$

With a p-value, calculated on a χ^2_1 distribution, of $0.400 > 0.05 (= \alpha)$, close to the above (2 proportions) result.

NOTES:

- ⋆ Chi-square tests may be seen as alternative tests to 1 proportion or 2 proportions tests.
- ⋆ Chi-square tests may be used only when the alternative hypotesis is two-sided.
- ⋆ Chi-square tests can manage also more than one or two proportions only; in that sense they are a generalization of "x proportions" tests.
- ⋆ In cases when the sample size is small, and the asymptotic approximation could not be applied, then the Yate's correction or (when possible) the Fisher exact test can be used. In above example, the Fisher exact test gives a p-value of 0.4140, very close to the above p-values.

Let's take again the last example, and suppose in this case that three samples were drawn: one sample for each of three raw material suppliers.

The aim of the test is always to check if **there's no association** between production Machine and Defectives:

H_0: No association exists between Machine and Flaws presence.
H_A: Some association exists between Machine and Flaws presence.

Since the three suppliers can produce materials with different quality levels, the researcher expects that aggregating all the sample data in only one sample, without considering differences between suppliers, could bias the results.

In this case, the hypothesis system must be rewritten in this form:

H_0: No association exists between Machines and Flaws presence within categories of Suppliers.

H_A: Some association exists between Machines and Flaws presence within categories of Suppliers.

The **Mantel-Haenszel-Cochran Chi-square test** allows to test this type of hypothesis system for a set of k 2 by 2 contingency tables.

In this example, $k = 3$ is the number of suppliers, and the 2x2 tables are the individual "Machines Vs Defectives" contingency tables.

If the four numbers in 2 by 2 tables are labeled like this:

$$\begin{matrix} a_j & b_j \\ c_j & d_j \end{matrix}$$

where $j = 1, \cdots, k$, and $(a_j + b_j + c_j + d_j) = n_j$, the equation for the Mantel-Haenszel-Cochran test statistic may be written as

$$\chi^2_{MHC} = \frac{\left(\left| \sum_{j=1}^{k} a_j - (a_j + b_j)(a_j + c_j)/n_j \right| - 0.5 \right)^2}{\sum_{j=1}^{k} (a_j + b_j)(a_j + c_j)(b_j + d_j)(c_j + d_j)/(n_j^3 - n_j^2)}$$

If H_0 is true, then one expects (prior the experiment) that the χ^2_{MHC} statistics tends to be distributed as a χ^2_1 distribution, and that its value will be relatively small.

If H_0 is false (and then H_A is true), then one expects (prior the experiment) that the χ^2_{MHC} statistics **shall not** be distributed as a χ^2_1 distribution, and that its value will be relatively large.

An upper threshold on the χ^2_1 distribution may then be chosen so that the probability of obtaining a χ^2_{MHC} test value greater or equal than the threshold is α.

Alternatively, one may calculate the probability on the χ^2_1 distribution to give a value greater or equal than the observed χ^2_{MHC} test value, obtaining the p-value.

If, after the experiment, the observed χ^2_{MHC} value is greater than the threshold value, or (equivalently) if the calculated p-value is less than the pre-assigned α level, then one must reject the null hypothesis (H_0), and accept the alternative hypotesis (H_A).

If, on the contrary, the observed χ^2_{MHC} value is less than the threshold value, or (equivalently) if the calculated p-value is greater than the pre-assigned α level, then one must accept the null hypothesis (H_0), and reject the alternative hypotesis (H_A).

Example:

Suppose that the collected sample data are summarized in following tables:

Suppl 1	Machine A	Machine B
Flawed	524	227
Safe	240	102

Suppl 2	Machine A	Machine B
Flawed	160	250
Safe	243	355

Suppl 3	Machine A	Machine B
Flawed	258	260
Safe	254	242

Example (continued):

If the three tables are aggregated in only one table, this is the result:

All Suppl	Machine A	Machine B
Flawed	942	737
Safe	737	699

If the researcher would test if no association exists (H_0) between Machine and Flaws using only this last table, the Fisher exact test would give a p-value of 0.008, quite less than 0.05, and then the researcher **should reject** H_0.

If the researcher calculates the χ^2_{MHC}, then the result is: $\chi^2_{MHC} = 0.366$ with p-value=0.545, and then the researcher **must accept** the null hypothesis.

The Poisson distribution may seldom be used to describe count data that have no prior upper limits.

For example, the number of flaws in LCD screens, or the number of arrivals at a front office in a given time span, are count data for which no upper limits can be prior established.

In that cases, tests could be performed on sample data to check if the overall population mean (or mean rate) follows pre-specified expected behaviors.

Suppose for example that 30 car hoods are inspected and 535 defects are found. The goal for the researcher could be to test if the mean defect count for all the population units is $= 15$ vs $\neq 15$.

In its simplest form, given a sample of size n of (for example, flaws) counts $(x_1, \cdots, x_n)$, the 1 sample Poisson Rate test try to decide between two alternative hypothesis as:

$H_0 : \mu = \mu_0$
$H_A : \mu \neq \mu_0$

Where μ is the parameter (the mean) of Poisson distribution, and μ_0 is an hypothesized value for that parameter.

Denoting with the symbol $\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$ the average number of flaws within the sample, $\hat{\mu}$ is an estimate of the true μ value.

The simplest **test statistics** that allows to test H_0 against H_A is then the following:

$$Z = \frac{\hat{\mu} - \mu_0}{\sqrt{\mu_0/n}}$$

If H_0 is true, and n is sufficently large (> 10), then the above quantity (prior conducting the sample extraction) will be distributed as a standard Normal distribution.

If H_A is true, and n is sufficently large (> 10), then the above quantity will be distributed as a Normal distribution centered "away from zero".

- ⋆ If H_0 is true, then one expects that the value of above formula, after the experiment, will be between the values -1.96 and 1.96 with probability $0.95 (= 1 - \alpha)$.
- ⋆ If, conversely, H_A is true, then one expects that the value of above formula, after the experiment, will be away from 0.
- ⋆ Then, a couple of thresholds could be selected, at -1.96 and 1.96:
  - if the value of test statistics falls between these limits, then H_0 is accepted (or, equivalently, H_A is rejected);
  - if the value of test statistics falls outside these limits, then H_0 is rejected (or, equivalently, H_A is accepted).

Example:

Let's go back to the car hood example. In that case, the hypothesis system states that:

$H_0 : \mu = 15$

$H_A : \mu \neq 15$

Also, $n = 30$, and the calculated average count of flaws is 17.83333. The value of test statistics, then, is:

$$Z = \frac{17.83333 - 15}{\sqrt{15/30}} \simeq 4.007$$

Since the test value "fall outside" the limits (or, equivalently, the p-value is very low ($\simeq 0.00006$) and less than $\alpha = 0.05$), the Null hypothesis must be rejected.

An $(1-\alpha)$ **confidence interval** for the "true" mean (rate) of counts may be generated, following the same guidelines used for the statistical test.

The confidence interval is given by:

$$\left(\hat{\mu} - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{\mu}}{n}}; \hat{\mu} + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{\mu}}{n}}\right)$$

If $\alpha = 0.05$, then $z_{1-\frac{\alpha}{2}} = 1.96$

Example (continue):

The "true" mean count of flaws, at a 0.95 confidence level, with the above example data, is:

$$\left(17.83333 - 1.96 \cdot \sqrt{\frac{17.83333}{30}}; 17.83333 + 1.96 \cdot \sqrt{\frac{17.83333}{30}}\right)$$

i.e., the interval $(16.322; 19.344)$.

NOTES:

- The test presented is said a **two-sided test**, because H_A states that μ is different from μ_0. If H_A states that μ is strictly greater (or less) than μ_0, the test is said **one-sided test**, and the threshold to accept/reject H_0 changes, to obtain the same α level.

- Actually, the tests on Poisson distributions are often performed on rates (i.e., flaws per m^2 of car hood in example), rather than on mean counts; in such cases the results are (from a statistical point of view) unchanged. To obtain the results for rates one must simply multiply the results by the conversion factor that produces the rate.

- The methods shown to construct statistical tests or confidence intervals for one poisson rate are the simplest ones, and rely on asymptotic behavior of statistcal test when n grows.

- An exact version of statistical test and a better approximation of confidence intervals are available.

- Confidence interval and statistical test usually give the same information: if the test accepts H_0, the confidence interval "encloses" the hypothesized value μ_0, and viceversa.

If one wants to compare the mean counts (μ_A and μ_B) of two independent sub-populations (called **A** and **B**), starting from (respectively) n_A and n_B Poisson sample counts $(x_{A1}, \cdots, x_{An_A})$ and $(x_{B1}, \cdots, x_{Bn_B})$, the statistical test that he/she should use is the 2 Samples Poisson Test.

The simplest form of 2 Samples Poisson test states that difference of means of two sub-populations are equal to a given value δ_0. The hypothesis system, then, may be stated as:

$H_0 : \mu_A - \mu_B = \delta_0,$
$H_A : \mu_A - \mu_B \neq \delta_0.$

Denoting then with the symbols $\hat{\mu}_A = \frac{\sum_{i=1}^{n_A} x_{Ai}}{n_A}$ and $\hat{\mu}_B = \frac{\sum_{i=1}^{n_A} x_{Bi}}{n_B}$ the average counts, respectively, within the sub-sample A and within the sub-sample B, $\hat{\mu}_A$ and $\hat{\mu}_B$ are estimates of the true μ_A and μ_B values.

In its simplest form, the 2 Samples Poisson test is:

$$Z = \frac{(\hat{\mu}_A - \hat{\mu}_B) - \delta_0}{\sqrt{\hat{\mu}_A/n_A + \hat{\mu}_B/n_B}}$$

If H_0 is true, and for n_A and n_B large enough, the distribution of test statistics is approximately a standard Normal.

Example:

72 hoods of Type A car, and 80 hoods of Type B car were sampled and examined to find differences in mean flaws counts. The total count of flaws found in Type A car was 1080, while the total count of flaws found in Type B car was 720.

Question: Have the two car types the same flaws mean counts?

Answer: The average flaws count for Type A car is $\hat{\mu}_A = 15$, whereas the average flaws count for Type B car is $\hat{\mu}_B = 9$; consequently:

$$\frac{15 - 9}{\sqrt{(15/72) + (9/80)}} = 10.59$$

with a p-value $\simeq 0$.

Since the absolute value of above quantity (10.59) is greater than the threshold value of a standard normal distribution for $\alpha = 0.05$, (1.96), then one may state that **there's empirical evidence that allows to reject the Null hypothesis of no difference between means**.

The same conclusions may be drawn by reading the p-value (the probability of having a value "external" to the values -10.59 and 10.59 on a standard normal distribution). This probability is nearly 0.
Since the p-value is less than the fixed α value, the null hypothesis is rejected.

NOTES:

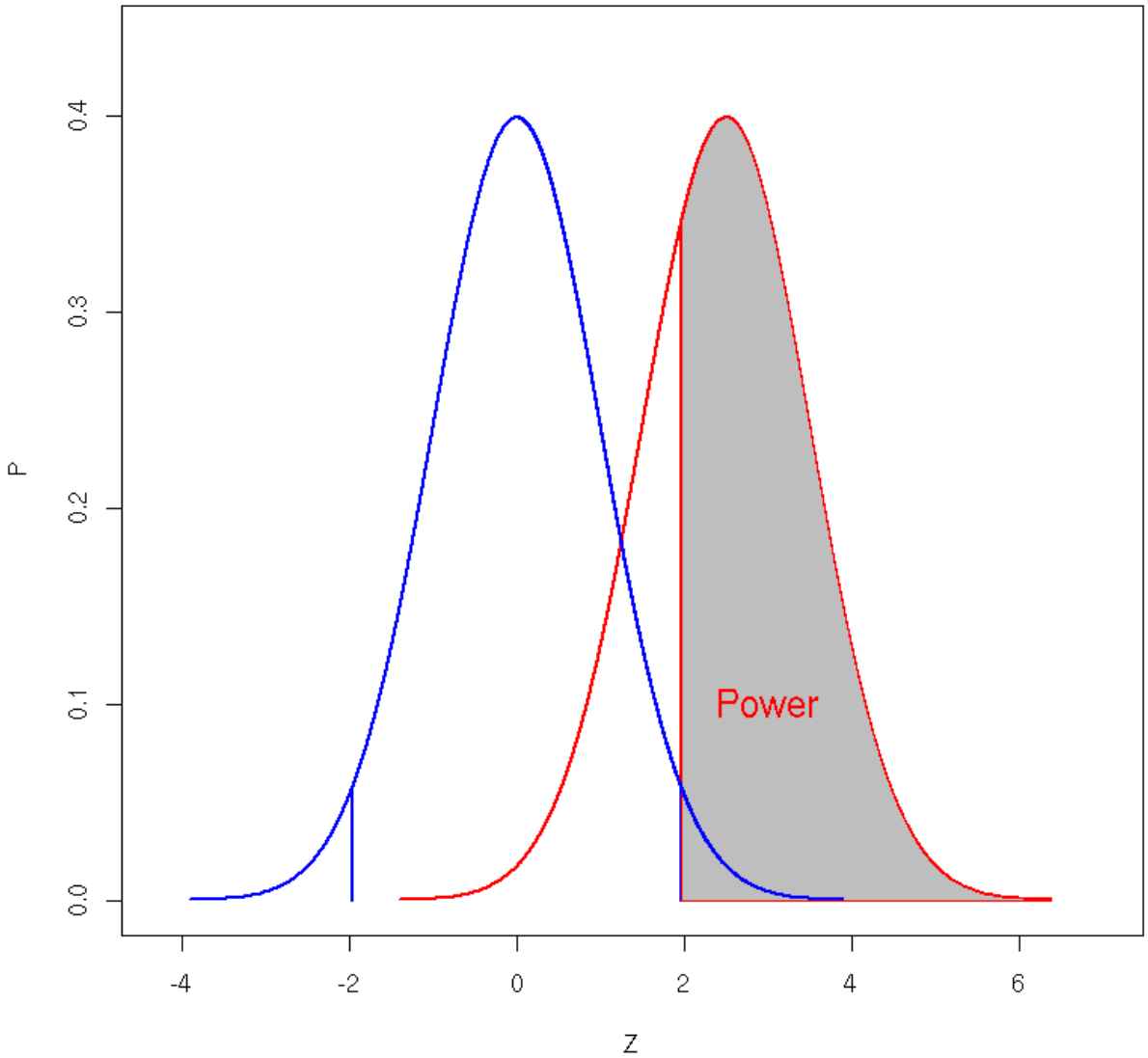
Several "forms" of two Samples Poisson test statistics are available; the proposed form is conceptually the simplest.

Finally, as for the 1 sample test, for the two samples test the one-sided alternative hypothesis is available. The threshold value, and the p-value, can be obtained in a similar manner than for the one proportion test.

When planning a statistical experiment, the first question one should almost always reply is: "**How many units must I collect?**" or "Which sample size should I choose?"

After a test, if the result is not statistically significant (i.e., H_0 is not rejected) an important question could be: "**Given the sample, what is the probability of rejecting H_0 under specific hypotheses?**"

The first question is called **prior power**, while the second is said **posterior power**.


P
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
6
Z
Power

The **power** and **sample size** analysis is an important tool that allows:

- ⋆ to check the ability of a statistical test to detect when the null hypothesis is false,
- ⋆ and to find the needed sample size to reach a specified probability of rejecting the null hypothesis when it is false.

In a generic statistical test, when the hypothesis system (H_0 against H_A) has been set up, the statistical procedure follows these steps:

✓ **BEFORE** collecting the data (the sample) from the population:

1. look for a suitable **statistics**, called **test statistics**, i.e., a function of data which summarizes specific characteristics of sample and gives information about the hypothesis system. Since the sampling is made randomly, the test statistics value shall be a random value. Also, the test statistics should be chosen so that it "behave" in two distinct manners, depending from the actual trueness of H_0 or H_A; for example, if H_0 is true, then the test statistics value should be "small", whereas, if H_A is true, then the test statistics value should be "large";
2. evaluate the test statistics **distribution** when H_0 is true;

3. determine a **threshold** value on potential values of test statistics, such that it is able to discriminate between H_0 or H_A. Since the test statistics will be a random value, the threshold should be chosen so that the probability of rejecting H_0, when it is true, is "small". That probability is usually denoted by α.

✓ **AFTER** collecting the data (the sample) from the population:

1. calcuate the test statistics value based on sample data;
2. given this value of test statistics and the threshold value, draw a conclusion about the trueness of H_0 or H_A.

PRIOR to the experiment execution, four conceivable possibility may occur:

	The test "says" H_0	The test "says" H_A
The population "is" H_0	OK	**Type I Error (α)**
The population "is" H_A	**Type II Error (β)**	OK

If the test rejects H_0 when H_0 is actually true for the population, a **Type I error** is produced; the probability of this error type is denoted by α, and it can be controlled by the experimenter.

If the test accepts H_0, when H_A is the "true state" for the studied population, a **Type II error** is produced; the probability of this error type is denoted by β, and it is not directly controllable during the test "building". The $(1 - \beta)$ value is called test **power**, and it represents the probability of properly rejecting H_0 when it is false.

The main parameters resulting from the power analysis are:

- ⋆ the sample size (n) or
- ⋆ the test power measure ($1 - \beta$)

and when one of two changes, the other changes accordingly.

One of two above values can be estimated, for a specific issue, based on existing relationships between 5 quantities:

- ⋆ the Type I error **probability** α; for this measure the **"direction" of** H_A **hypothesis** must be specified too: **one-sided** or **two-sided**;

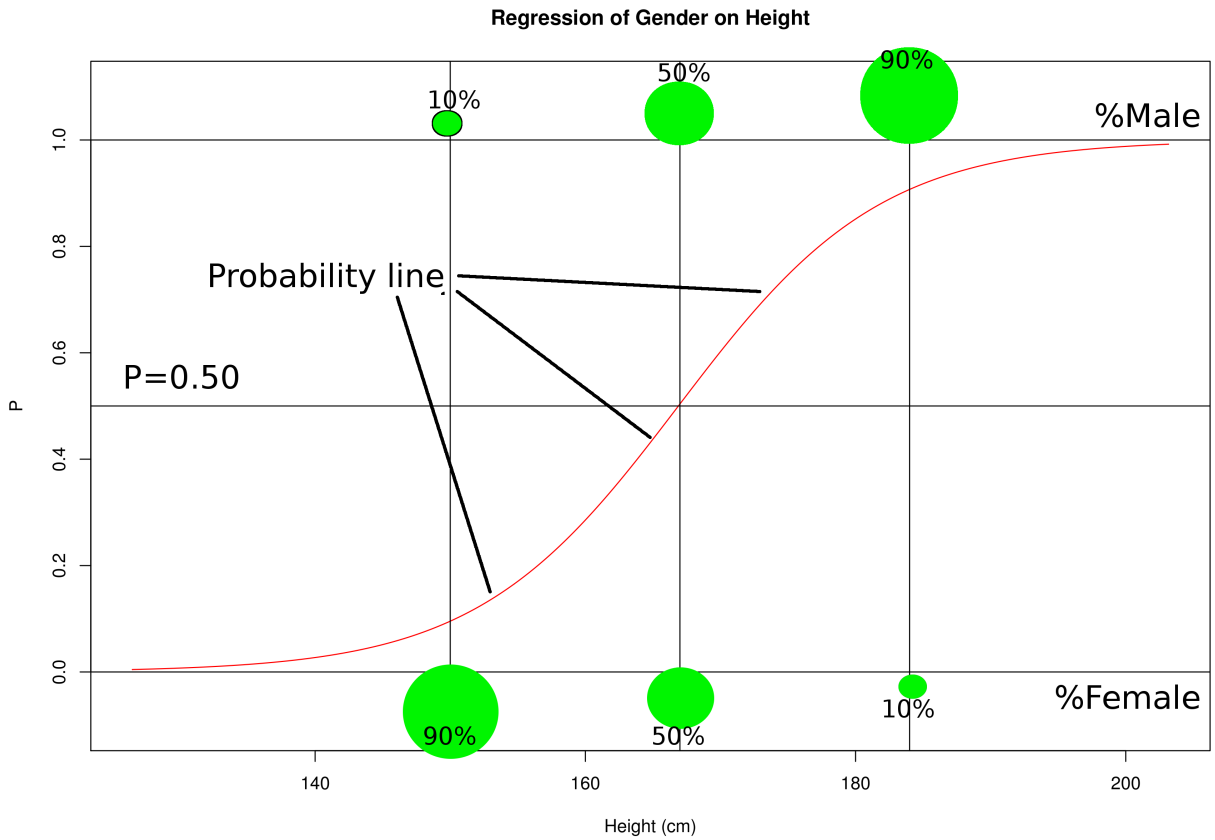
- **the β probability**, i.e., the probability of wrongly reject the alternative hypothesis H_A (i.e., to accept H_0) when it is true;

- the δ **difference** value, between the hypothesized and the true value of the studied quantity (or the difference **d** between the hypothesized and the true value of test statistics). For example, when considering an hypothesis test on one population mean , δ is the difference between the hypothesized mean μ_0, and the true population mean μ; when considering a test on two independent means, δ is the difference between μ_A and μ_B;

- ⋆ the population **variance σ^2**, if known, or an **estimate s^2** of population variance, when the **true variance σ^2 is unknown**;

- ⋆ **the sample size n**, or the sample size of each subgroup of whole sample, if more than only 1 sample shall be considered.

Suppose we want to predict whether someone is male or female (M=1, F=0) using height in centimeters. We could plot the relations between the two variables as we customarily do in regression. The plot might look something like this:



Some points to notice about the graph (data are fictional):

- ⋆ The regression line is a rolling average, just as in linear regression. The Y-axis is P, which indicates the proportion of 1s at any given value of height.
- ⋆ The regression line is nonlinear.
- ⋆ None of the observations –the raw data points– actually fall on the regression line. They all fall on zero or one.

The **logistic regression** is a "special type" of regression where the dependent variable has a Bernoulli(p) (or a Binomial(n,p)) distribution, and then it can take only 0 or 1 values.

Also, in the logistic regression the main parameter of interest (p: the "success" probability) must "fall" within the $(0, 1)$ interval.

As in "regular linear" regression, the logistic regression try to explain the variability of dependent variable based on some independent variables.

Independent variables, as in linear regression may be quantitative/continuous (as in graph above) as well as qualitative.

(Note: "multinomial" and "ordered multinomial" types of logistic regression exist too).

In one of its simplest forms, the mathematical formulation of logistic regression states that the dependent variable (Y) has a distribution whose expected value is such that:

$$E\{Y\} = p = \frac{e^{\alpha+\beta\cdot X}}{1+e^{\alpha+\beta\cdot X}}$$

Where X is the independent variable.

Since the expected value for a Bernoulli variable is p (the probability of "success"), the above formula states that the success probability is linked to the X independent variable by a logistic relation or, equivalently, that

$$ln\left(\frac{p}{1-p}\right) = \alpha + \beta \cdot X$$

The estimation of logistic model is a complex mathematical/statistical matter.

Usually, the more reliable method is the Maximum Likelihood estimation.

This method allows to estimate all the parameters included in model (remember that the model illustrated above is one of simplest models), with error and significancy evaluation too.

The exponential of estimated parameters, anyway, are estimates of odds-ratio for the specific effect involved by the parameter itself.

Example:

A sample of 40 height readings has been collected: 20 males and 20 females. The data are the following:

gender	F	F	F	F	F	F	F	F
height	158.2512	168.455	157.8914	164.0353	164.0082	147.6034	141.7605	155.4557
gender	F	F	F	F	F	F	F	F
height	165.6356	174.7999	165.8107	151.6972	174.2301	161.6354	177.6935	154.0353
gender	F	F	F	F	M	M	M	M
height	175.0083	160.0799	169.4225	183.6644	165.6596	185.3086	175.7386	168.63
gender	M	M	M	M	M	M	M	M
height	170.4932	179.8414	163.6995	148.1559	183.6873	169.3332	179.484	175.2373
gender	M	M	M	M	M	M	M	M
height	164.1659	172.8038	182.1371	160.7205	185.0418	169.9699	170.6346	166.1434

Estimating the probability of being "Male" via logistic model, gives the results:

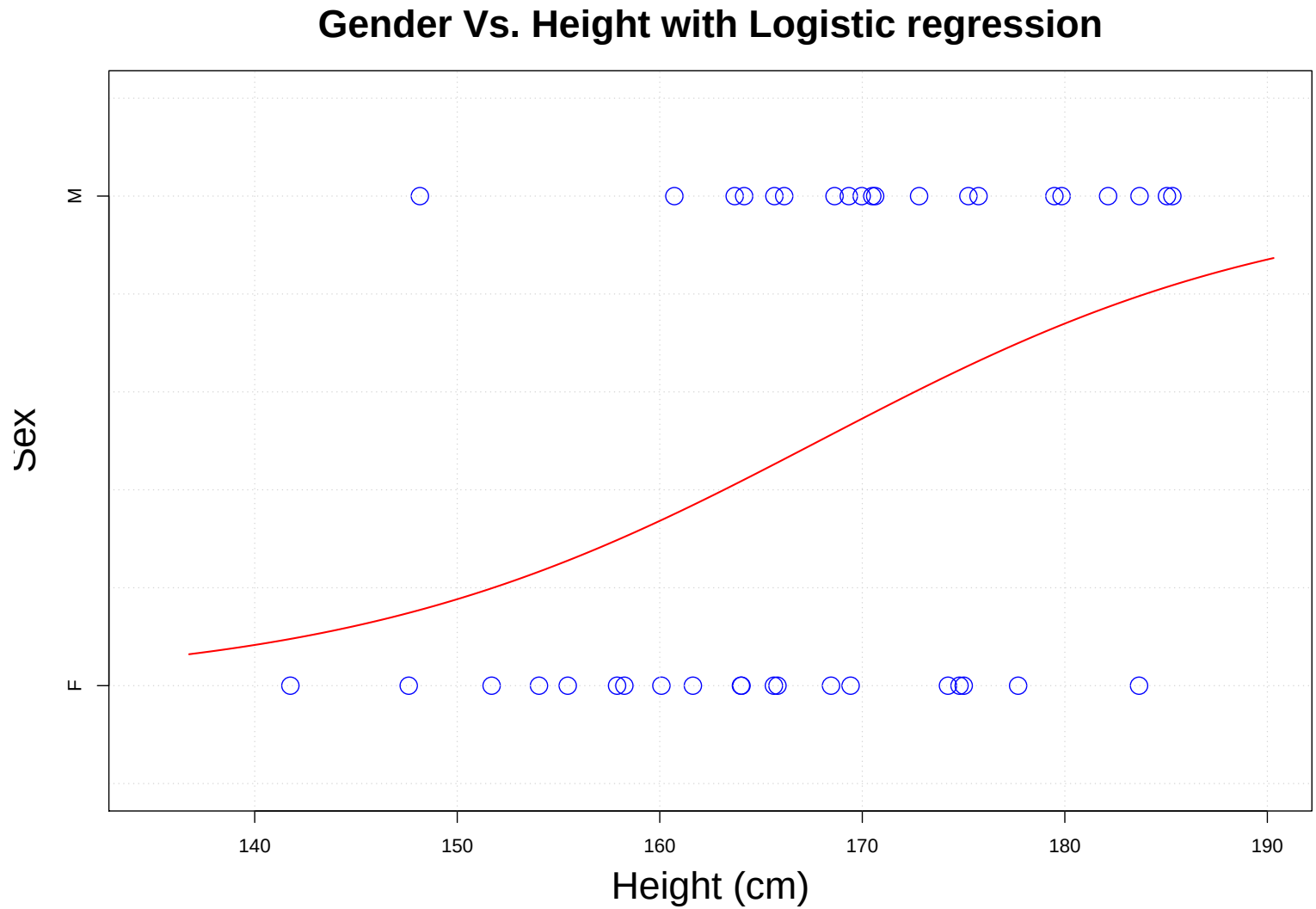
Coefficients:

	Estimate	Std. Error	z value	Pr($>\|z\|$)
(Intercept)	-14.45785	6.31375	-2.290	0.0220
height	0.08612	0.03750	2.297	0.0216

The table shows that the probability of having a Male is given by:

$$Pr\{Y = \text{“}Male''\} = \frac{e^{-14.45785+0.08612 \cdot height}}{1 + e^{-14.45785+0.08612 \cdot height}}$$

Note: the positive estimated coefficient for the `height` variable means that the probability of having a Male "grows" with the height of subject.


Gender Vs. Height with Logistic regression
Sex
M
F
140
150
160
170
180
190
Height (cm)

In above slides, some new quantities were used. The quantity:

$$\frac{p}{1-p}$$

is called **Odds**.

The quantity

$$ln\left(\frac{p}{1-p}\right)$$

is called **Logit**.

Note that, for example, in above case of Male/Female distinction, odds can also be found by counting the number of people in each group and dividing one number by the other.

The probability is not the odds.

In example, the odds of being male would be .90/.10 or 9 to one, and the odds of being female would be .10/.90 or 1/9 or .11. This asymmetry is unappealing.

The natural log of odds makes the logit values symmetrical around 0 (zero).

In a simpler "two groups" example, where male and female subjects should be predicted using only one qualitative variable with only two categories (e.g., foot size: Small or Big), the **odds-ratio** ϕ may be used to show "how much the independent variable is able to predict the dependent variable". Suppose that the relationship between Gender and Foot size can be expressed in this form:

	Big	Small
Male	a	b
Female	c	d

In this case, the odds-ratio is:

$$\phi = \text{odds-ratio} = \frac{\frac{a}{a+c} / \frac{c}{a+c}}{\frac{b}{b+d} / \frac{d}{b+d}} = \frac{ad}{bc}$$

Odds-ratio express the multiplicative change in probability rate between Male and Female subject when changing (in this case) the foot size from Small to Big. In above example, suppose that the table contains these data:

	Big	Small
Male	80	15
Female	10	90

Here, the Male Vs Female odds-ratio is 48.

That means that the relative probability of having a Male subject with respect to having a Female subject from Small foot to Big foot, grows 48 times.

Example:

If the above table is used to estimate the probability of being "Male" via logistic model, the model estimate result is:

Coefficients:

	Estimate	Std. Error	z value	Pr($>\|z\|$)
(Intercept)	2.0794	0.3354	6.200	5.65e-10
foot:Small	-3.8712	0.4362	-8.875	$<$ 2e-16

The table show that the probability of having a Male is given by:

$$Pr\{Y = \text{“}Male''\} = \frac{e^{2.0794-3.8712 \cdot foot}}{1 + e^{2.0794-3.8712 \cdot foot}}$$

Where foot=0 means "Big foot", and foot=1 means "Small foot"

The second line of table contents, show that the probability of having a Male changing the foot size from Big to Small reduces such that the $log(\text{odds-ratio}) = -3.8712$.

Consequently, the odds-ratio of change from Big to Small foot is

$$\phi = e^{-3.8712} \simeq 0.02083335$$

very close to 1/48=0.0208333333 obtained from above calculations.

For theoretical reasons (such as the central limit theorem), any variable that is the sum of a large number of independent factors is likely to be normally distributed. For this reason, the normal distribution is widely used in statistics, natural science, and social science as a simple model for complex phenomena.

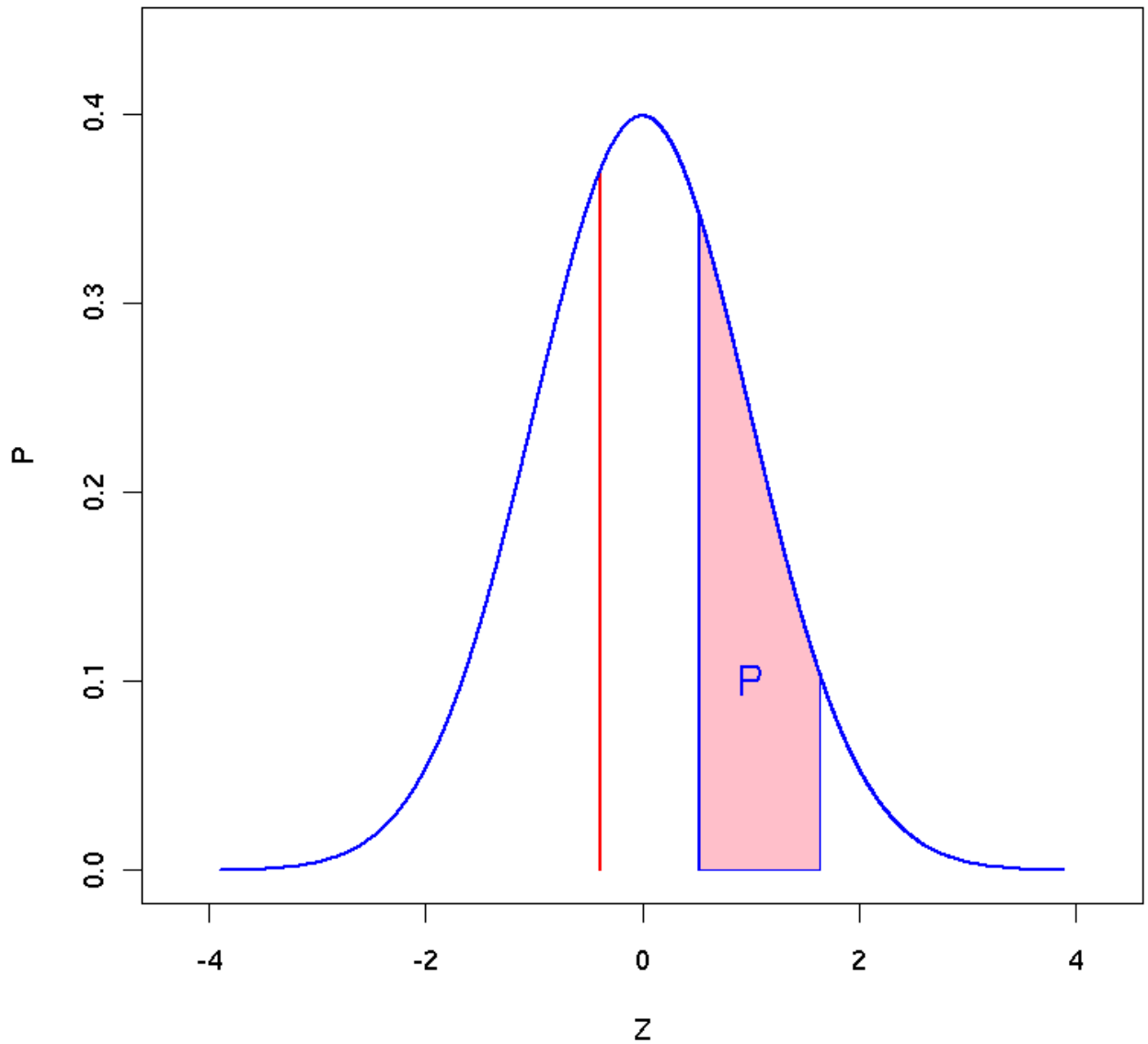
For example, the error in an experiment is usually assumed to follow a normal distribution, and the propagation of uncertainty is computed using this assumption.

The normal distribution ($N(\mu, \sigma)$) is defined by two parameters:

- μ, determines the position of the distribution;
- σ, determines the "variability" or "dispersion" of the distribution.

The normal distribution is defined between $-\infty$ and $+\infty$, and, being a probability distribution, his integral from $-\infty$ to $+\infty$ is equal to 1.

Asymptotically tends to zero, and the tails over the value 6σ, with good approximation, can be considered truncated.


P
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
Z
P

The mathematical formulation for the normal distribution is:

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

This is the expression of **probability density function (pdf)** of normal distribution.

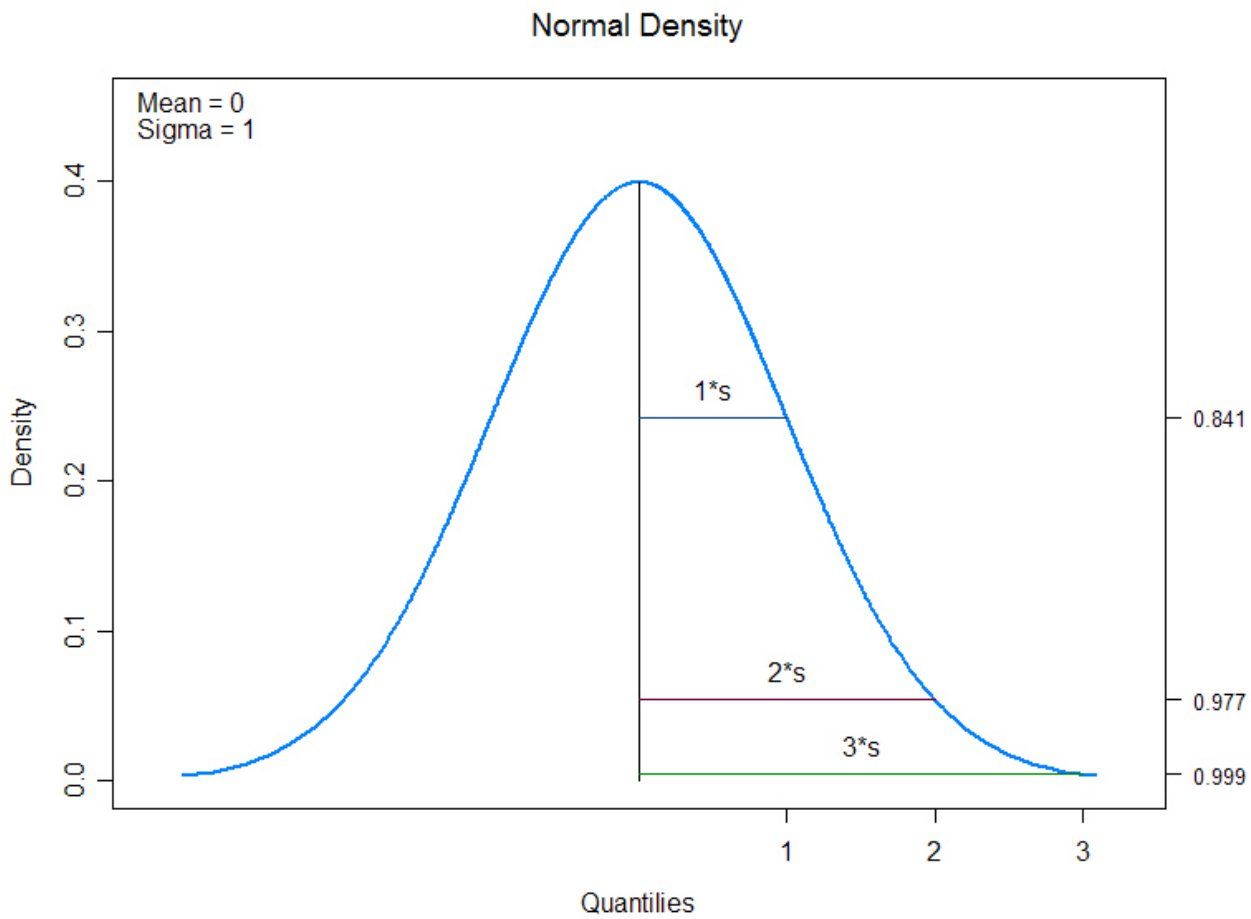
It allows to evaluate the value of Y (ordinate value) for all X values (abscissa value).

The μ and σ values define totally the probability density function.

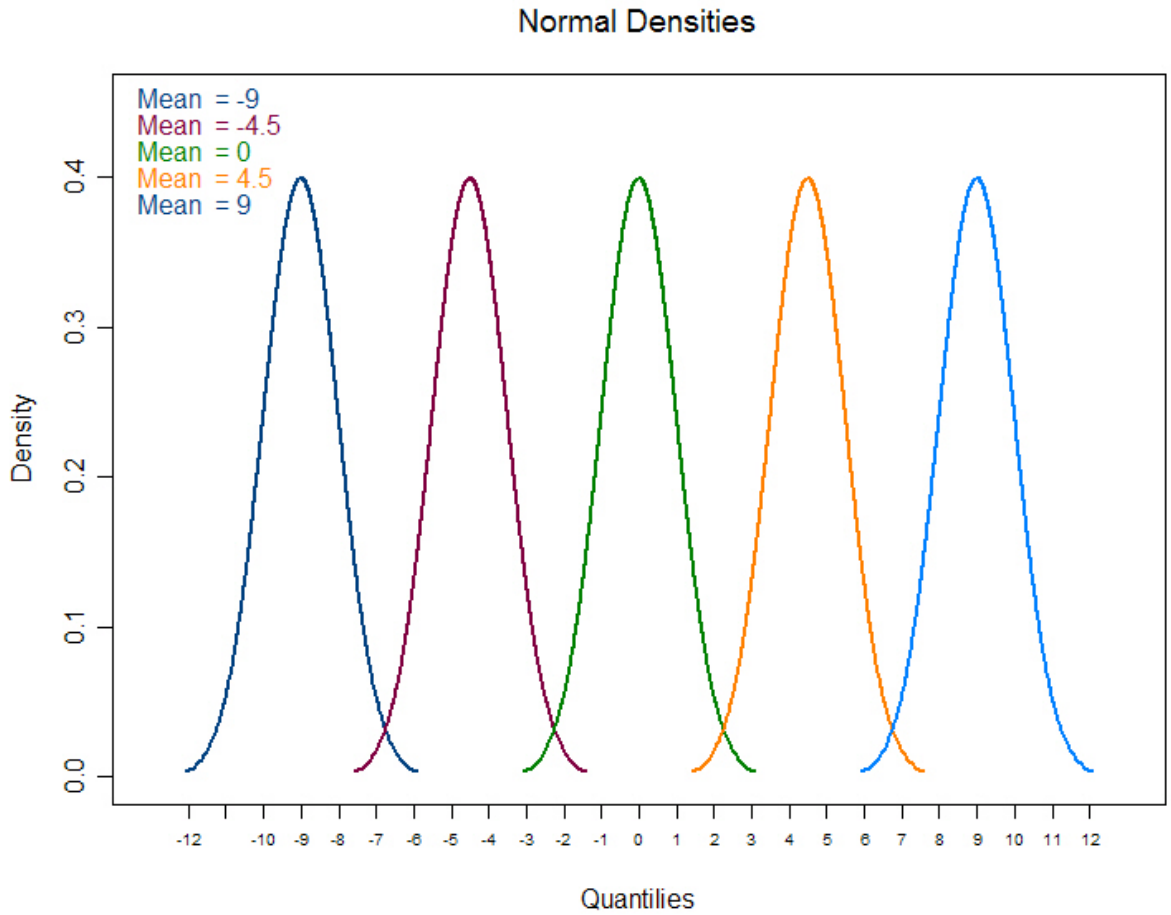
The density normal curves are infinite.

When $\mu = 0$ and $\sigma = 1$, then the Normal distribution is said a **standard normal distribution**.

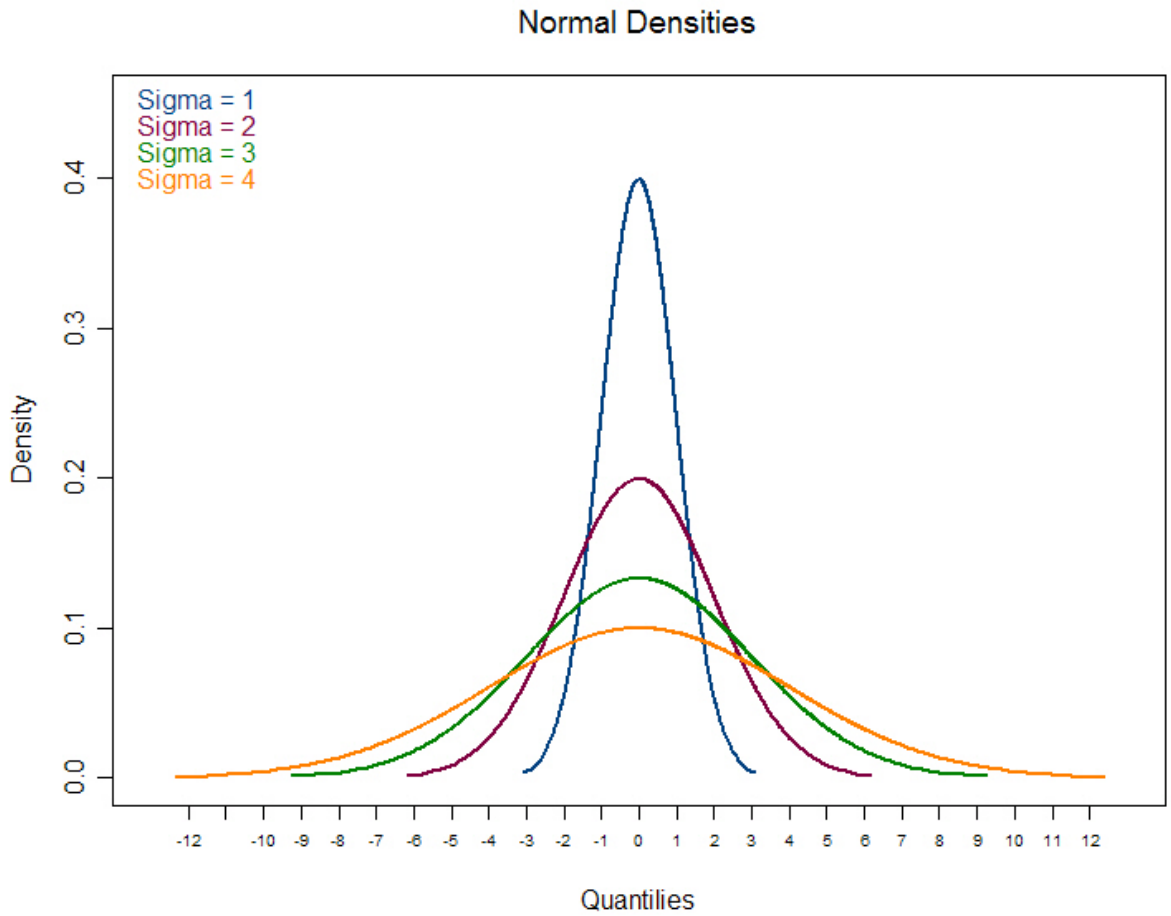
Graph for a Normal Distribution.



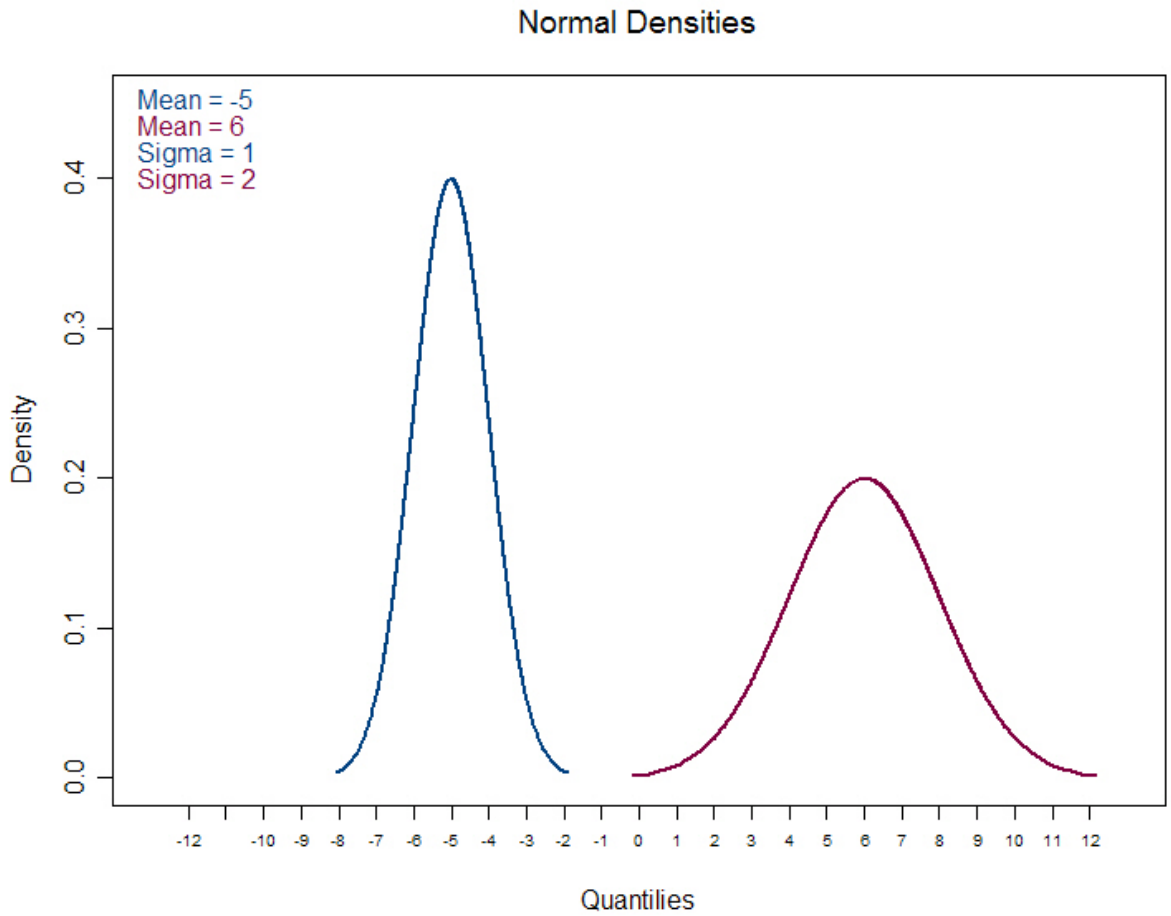
If μ changes and σ is constant, there are infinite normal curves with the same shape and size, but with a different axis of symmetry.



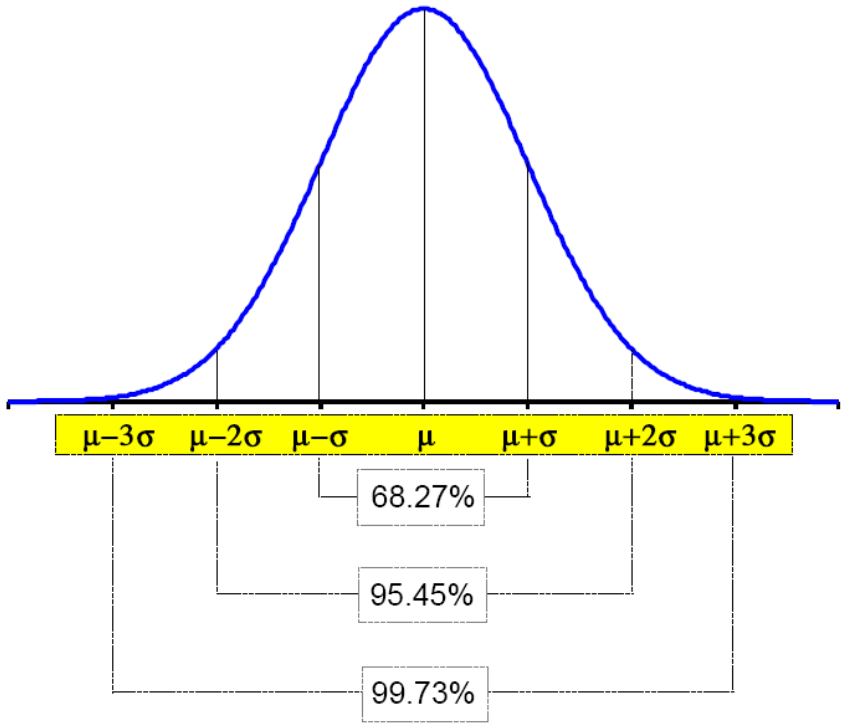
If μ is constant and σ changes, all the infinite curves have the same symmetry axis, but σ determines their shape.



In this case both μ and σ are different:



This graph shows some relevant probabilities for the Normal distribution:



The $\boldsymbol{\chi^2_\nu}$ random variable is a random variable obtained by summing up the square of ν independent standard normals ($\mu = 0$ and $\sigma = 1$) $Z_1,\ Z_2,\ \ldots,\ Z_\nu$:

$$\chi^2_\nu = Z_1^2 + Z_2^2 + \cdots + Z_\nu^2$$

The χ^2 probability density function (pdf) is determined **by only $\boldsymbol{\nu}$ parameter**, i.e., the number of degrees of freedom (df).

The pdf is defined between 0 and $+\infty$:

$$y = f(x) = \frac{1}{2^{\frac{\nu}{2}}\Gamma(\nu/2)} x^{\frac{\nu}{2}-1} e^{-\frac{x}{2}}$$

Where $\Gamma(\cdot)$ is the Gamma function.

When ν grows, the pdf tends to be similar to a Normal distribution.

The next graph shows the χ^2_{df} distribution for several df values.