Control de calidad en R: Conceptos y controles básicos en calidad de la IG

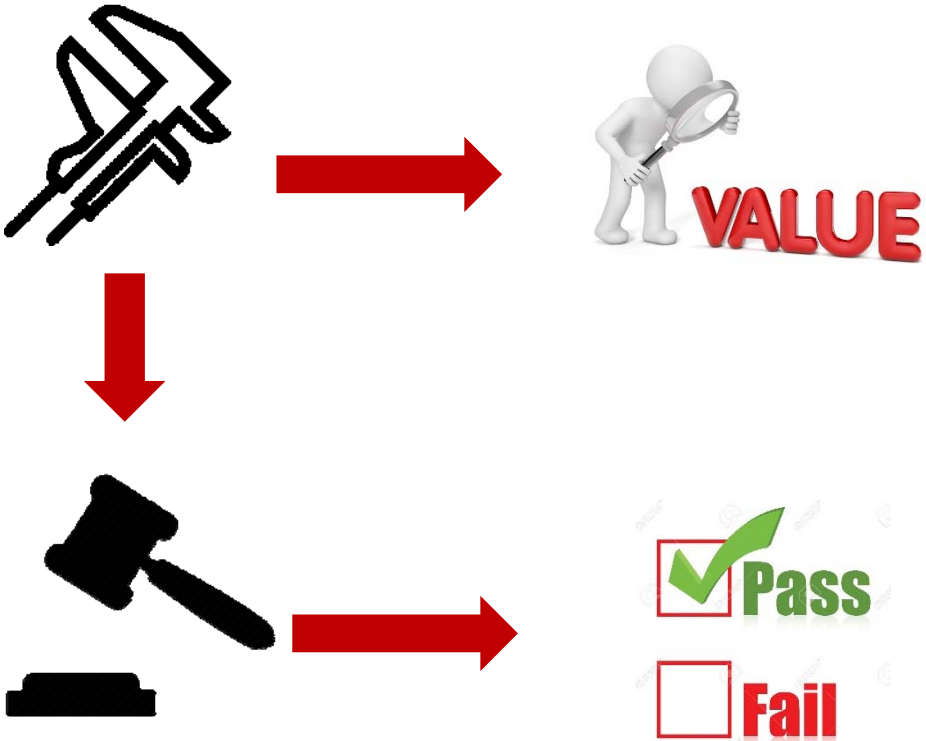
Francisco Javier Ariza López
fjariza@ujaen.es
Universidad de Jaén (España)

Quito, Noviembre de 2016

Índice

- **Estimation and control**
- **Errors**
- **Statistical hypothesis**

Estimación y control



μ, σ, ξ, α

Valor medio, desviación, precisión, significación, tamaño muestra

Estimación

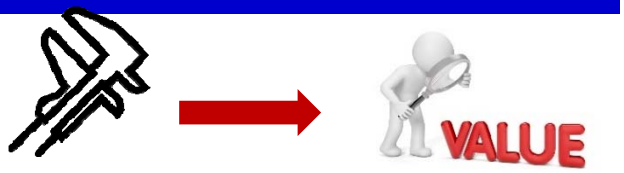
Tol, α, β

Tolerancia, significación, potencia (riesgos tipo I y II),

Control

Estimación y control

Differences



Estimation:

Estimation means the desire to know the true value of a parameter (e.g. mean, deviation, proportion, etc.) with a limited uncertainty.

Estimation is the reliable determination of a value → **Quality estimation.**

In quality assessment → The quality of a parameter of a product is determined by means of a sample. An outcome (estimation), without trial for acceptance or rejection, occurs.

Sample size is related to the dispersion in the population (deviation)

$$n = \frac{N\sigma^2}{(N-1)D + \sigma^2}$$

Where:

$D = B^2/\mathrm{k}^2$

B	is the limit for the estimation error,
k	is a value related to the confidence level (usually $k = 2$ for a 95,5% confidence level)
σ^2	is the error's population variance (that has to be known in some sense)

$$n = \frac{\sigma^2}{D}$$

Estimación y control

Differences

Control:

Control means the desire to know if the product, or an attribute of the product, meets or not a specific requirement, in general if it is good or bad for a specified use.

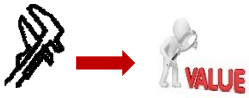
Control is the reliable determination of whether a condition is met → Quality control.

In quality assessment → The control of a parameter of a product is performed by means of a sample and some rules in order to determine whether the quality of the product reaches specifications. The outcome is judgment on whether to accept or reject the product.

Sample size is related to the power of the result

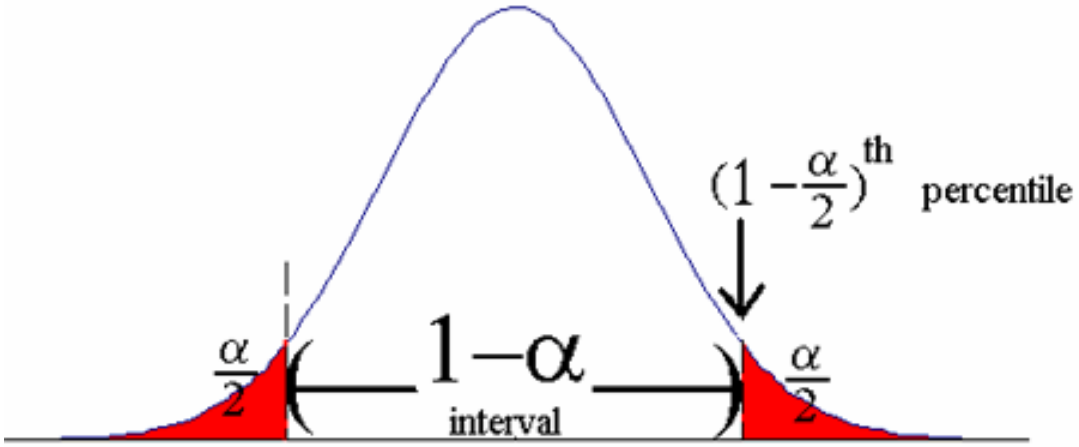
Estimación y control

Quality estimation / parameters estimation



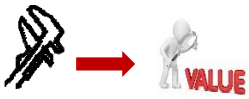
The confidence interval for the mean μ, with a $(1-\alpha)\%$ confidence level in an unidimensional space is:

$$\mu \in \left(\bar{x} - k_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \; ; \; \bar{x} + k_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$



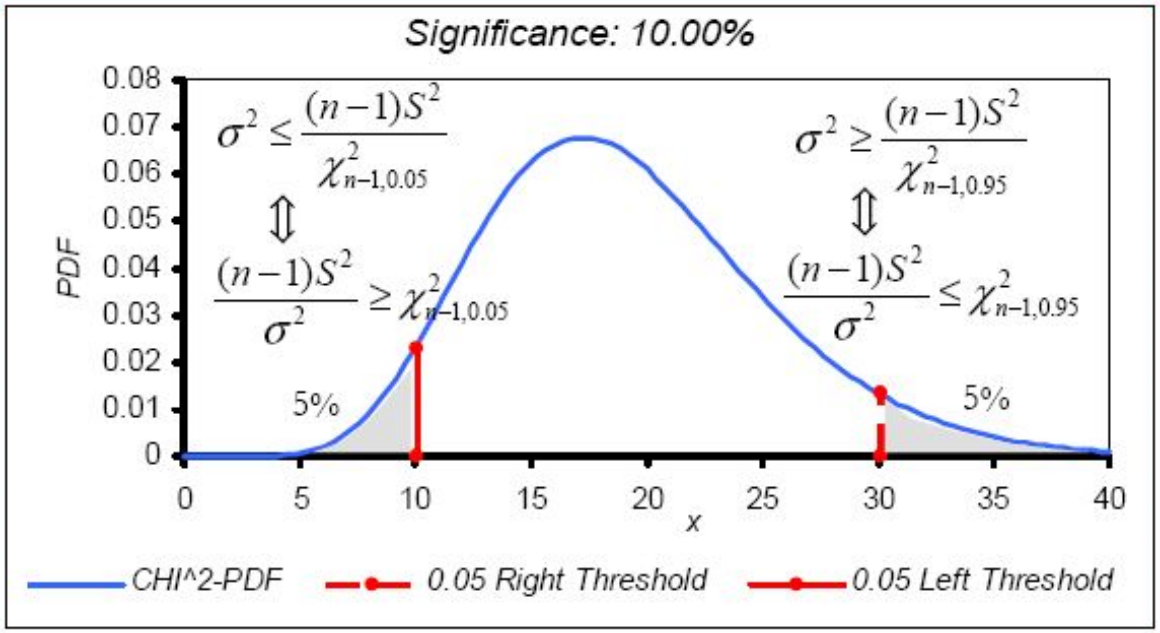
Estimación y control

Quality estimation / parameters estimation



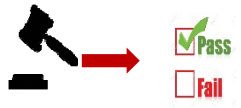
The confidence interval for the variance, with a $(1-\alpha)\%$ confidence level in an unidimensional space is:

$$\sigma^2 \in \left[\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}; \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}\right]$$



Estimación y control

Quality Control / Contrast of hypothesis

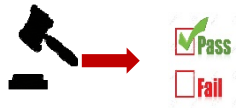


- In QC we use an statistical tool to make a decision in respect with a population hypothesis
- The hypothesis that would be contrasted is fixed in advance and it is called "null hypothesis", ($\mathbb{H}_0$). The decision we adopt if we consider that $\mathbb{H}_0$ is false is called "Alternative hypothesis" ($\mathbb{H}_1$)
- $\mathbb{H}_0$ can be true or not, and we can accept or reject it. So, four situations may be allowable:

HYPOTHESIS TESTING OUTCOMES		Reality: The Null Hypothesis Is True	Reality: The Alternative Hypothesis is True
Research	The Null Hypothesis Is True	Accurate $1 - \alpha$	Type II Error β
Research	The Alternative Hypothesis is True	Type I Error α	Accurate $1 - \beta$

Estimación y control

Quality Control / Contrast of hypothesis

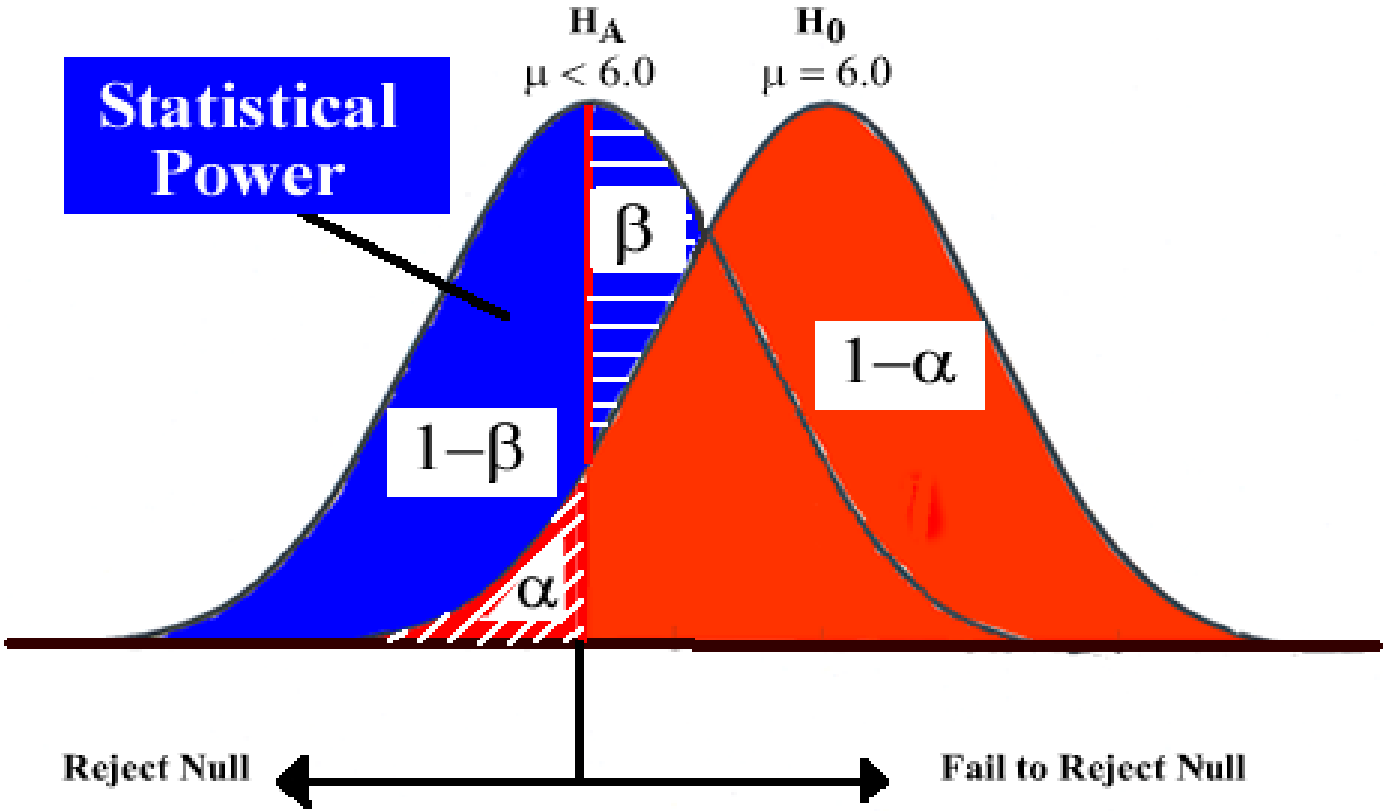


To verify the trueness of the hypothesis

- We select a sample of size n,
- We calculate a sampling function (test) whose distribution of probability is known,
- We calculate the probability that, if $\mathbb{H}_0$ is true, we have obtained a test functions so extreme so we have obtained. This probability is called "p-value", (p),
- The p-value is "the degree of credibility" about the trueness of $\mathbb{H}_0$. If $p < \alpha$, then the probability is not credible enough and we reject $\mathbb{H}_0$. In other case, we have to conclude that there are not evidence enough to reject $\mathbb{H}_0$

Estimación y control

Quality Control / Contrast of hypothesis



Errores

ERROR and UNCERTAINTY are key concepts.

Error: The difference between a measured value of a quantity and a reference value (conventional value or true value) [VIM 2007]

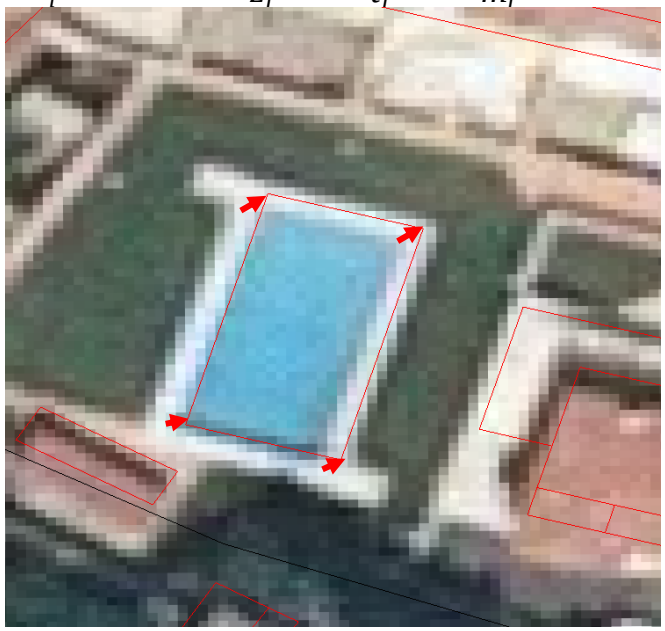
Error: The discrepancy between two values that are supposed to be equal.

$$e_{x_i} = x_{t_i} - x_{m_i} \qquad e_{y_i} = y_{t_i} - y_{m_i} \qquad e_{z_i} = z_{t_i} - z_{m_i}$$

Possible errors:

- **Blunders or Mistakes.**
- **Systematic or bias (constant or variable)**
- **Random**
- ~~**Outliers**~~

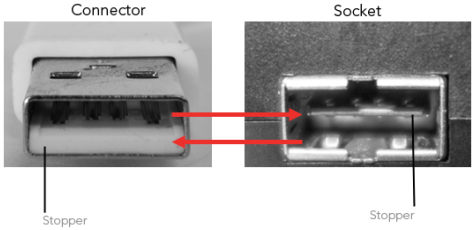
Outliers →signals, no errors

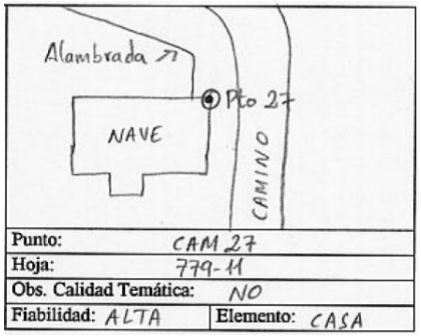


Errores

Blunders or Mistakes: A mistake typically caused by ignorance, carelessness or negligence.

- **Origin: human (confusion, mistake), mechanical, electronical, etc.**
- **Data: elimination of bad data.**
- **What can we do: Use methods that reduce the possibility of occurrence (poka-yoke), etc.**

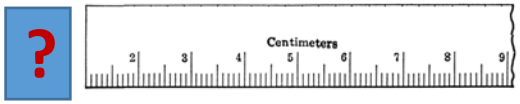


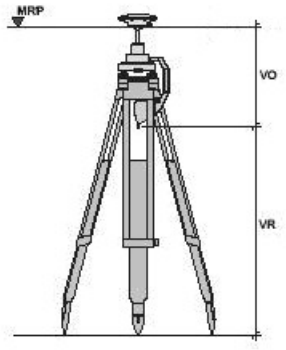


Errores

Systematic or bias (constant or variable): a particular tendency toward something. A systematic error inherent in measurements due to some deficiency in the measurement process or subsequent processing. It is the constant difference between a measure and its true value

- **Origin: Operators, instruments, methods, etc.**
- **Data: if correction of errors is possible → the use of data is possible.**
- **What can we do: detection system, calibration of instruments and methods, use of models (e.g. temperature, ionosphere, etc.).**



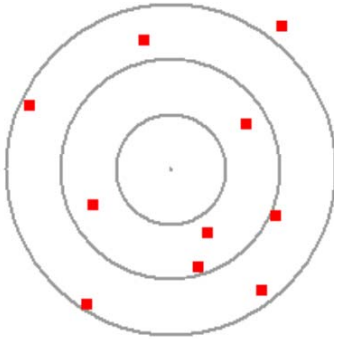


VO Vertical Offset
VR Vertical Height Reading
MRP Mechanical Reference Plane

Errores

Random errors: experimental variation in measurements that are caused by unknown and unpredictable reasons (pure chance).

- **Origin:** large number of parameters beyond the control of the process that may interfere with the results (e.g. temperature, age, wind, electronic noise, etc.).
- **Data: Are included into the data we use (have uncertainty)**
- **What can we do: Randomness is not removable. MODELS are used to model such errors. Change methods or technology.**



Random Error

Errores

UNCERTAINTY: Non-negative parameter characterizing the dispersion of the quantity values being attributed to a measurand, [VIM, 2007].

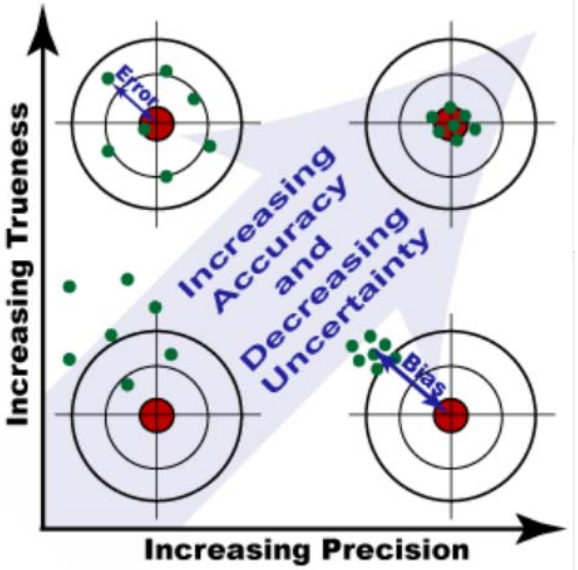
ACCURACY: The closeness of agreement between a test result and the accepted reference value. [ISO 3534-1]

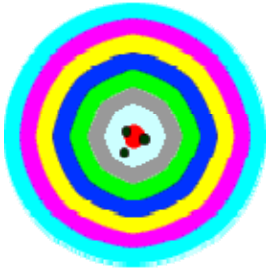
Accuracy = Trueness + precision

Bias component **Random component**

TRUENESS: The closeness of agreement between the average value obtained from a large series of test results and an accepted reference value.

PRECISION: The closeness of agreement between independent test results obtained under stipulated conditions.





Ideal situation

Errores

Working with positional errors we use models in order to simplify the work

For majority of cases: The Gaussian or Normal model is assumed

Many studies point that this is WRONG. This hypothesis is NOT true.

Other base models for positional uncertainty in SDS:

- LIDAR (Maune, 2007): Without a parametric model (distribution free)
- Manual digitizing (Bolstad et al 1990): Bimodal
- Digitizing (Tong & Liu, 2004): p-norm (Normal + Laplace)
- Geocodification (Cayo and Talbot 2003; Karimi and Durcik 2004, Whitsel et al. 2004): Log normal
- GNSS Observations (Wilson, 2006; Logsdon, 1995): Raleigh, Weibull
- Other models that are mentioned: Folded Normal, Half Normal, Gamma

Hipótesis básicas

Statistical prerequisites

Underlying hypothesis (randomness).

Underlying hypothesis (no outliers).

Underlying hypothesis (normality).

Underlying hypothesis (no correlated).

Underlying hypothesis (homocedasticity).

Hipótesis básicas

Statistical prerequisites: Randomness

Why is important data randomness?

- The hypothesis of randomness is very important in order to avoid bias in the sample. The absence of randomness means a data manipulation.
- Randomness guarantees sampling representation. It assumes that all sampling elements have the same probability to be token in the sample and that they have been independently selected.

Which is the origin of no-randomness?

- Data manipulation (intended or non-intended).

How no-randomness can affect my assessment?

- Results will be biased and representativeness is lost.

How can be verified?

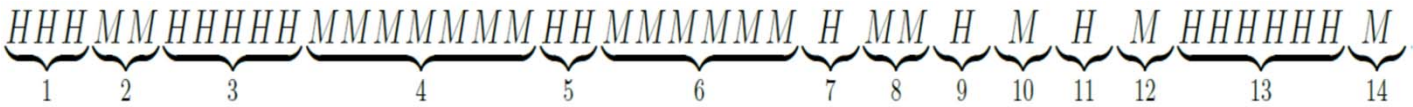
- They exist many statistical procedures to contrast this hypothesis.
- One of them is the Wald–Wolfowitz runs test.

Hipótesis básicas

Statistical prerequisites: Randomness

Wald–Wolfowitz runs test:

1. Write the sample in the same order it has been obtained and select the mean (or median) value. This is the "cut point". Let n_1, n_2 be the number of sampling elements below and under the cut point
2. Starting from the first value obtained, and counting the number of runs, R (number of times that we find a change between a value upper and below the cut point).



3. The test statistics,

$$Z = \frac{R - \mu_R}{\sigma_R} \sim N(0,1)$$

where

$$\mu_R = \frac{2n_1n_2}{n_1+n_2} + 1; \quad \sigma_R = \sqrt{\frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}$$

Hipótesis básicas

Statistical prerequisites: Randomness

It may be calculated in R using the function ***runs.test*** that belongs to the package ***randtests.*** The order is:

runs.test(x, alternative, threshold, pvalue, plot)

The arguments are:

- ***X***: a numeric vector containing the observations
- ***Alternative***. Must be “two sided” (default), “left.sided” or “right.sided”
- ***Threshold***: the cut-point to transform the data into a dichotomous vector
- ***p-value***: Must be “normal” (default) or “exact”
- ***plot***: If true, a plot should be created.

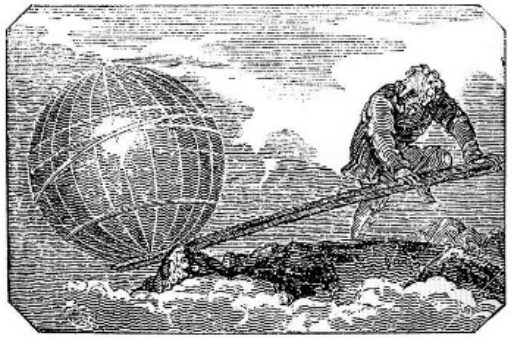
Hipótesis básicas

Statistical prerequisites: Outliers

Why is important to detect the presence of outliers?

- Outliers affect a lot the values computed for the mean and the standard deviation.
- They are far away from the mean and this has a lever effect on their computations.

Which is the origin of outliers?



- Overlap of two or more processes.
- Process with a long-tailed distribution.
- May be due to an obtaining error, variability in the measurement or it may indicate experimental error.

How can affect my assessment?

- We obtain bad estimations for the mean and standard deviation.

How can be detected?

- They exist many statistical procedures to contrast this hypothesis.
- One of them is the k-σ method.

Hipótesis básicas

Statistical prerequisites: Outliers

Advices:

- The distance for a point to be considered an outlier depends on the distribution. In consequence, when the number of outliers is high, it may indicate a heavier-tail distribution than the reference distribution (for instance, data are obtained from a mixture of distributions).
- Once a point has been classified as outlier, it must be carefully analyzed to determine why it is outlier and proceed in consequence.
- Usually, distance is referred in terms of the standard deviation, but this value is highly influenced by the presence of outliers, specially for small size samples. This problem can be avoided using robust estimation methods

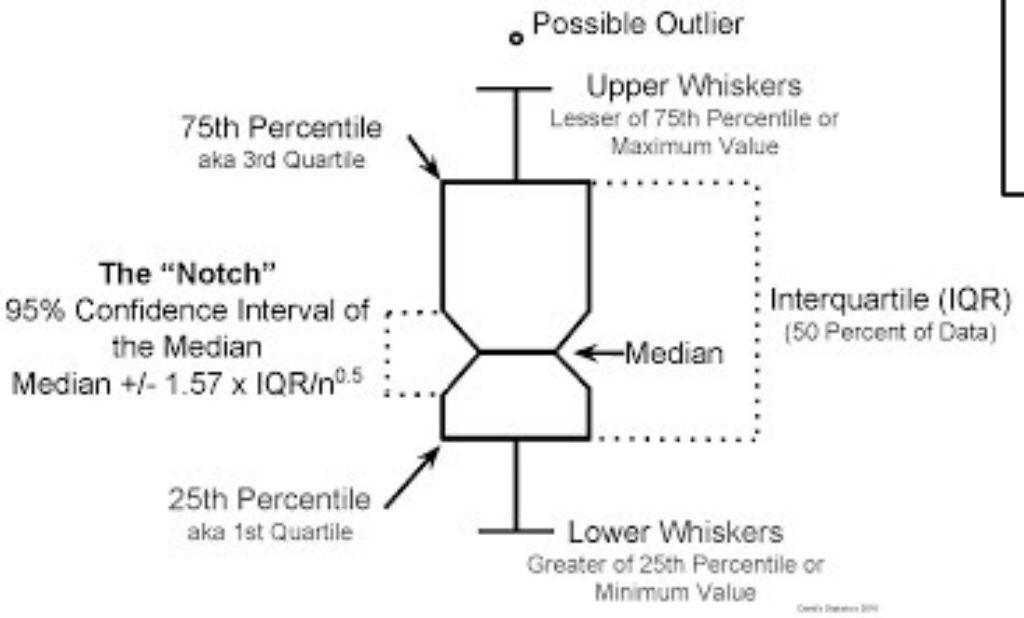
Hipótesis básicas

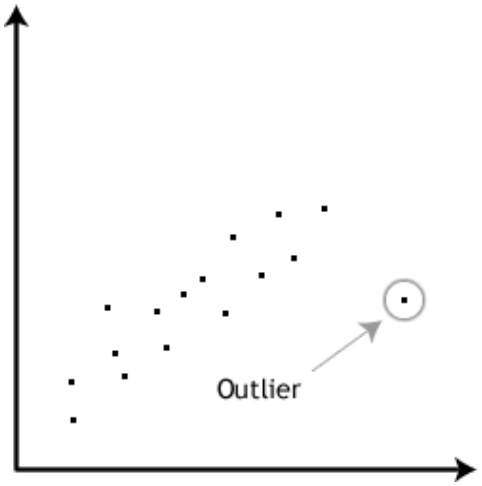
Statistical prerequisites: Outliers

Determination of outliers:

Graphical procedures: Box-Whisker plot.

- In R: ***boxplot()***





- It exists an modified version adapted to other distributions in the Standard ISO 16269-4.

Hipótesis básicas

Statistical prerequisites: Outliers

Determination of outliers:

Case normal: $k\sigma$ method

- The sample is typified and a point will be classified as outlier if its typified value (in absolute value) is greater than k

Outliers percentage for the $k\sigma$ method			
k	$P[Z < k]$	$P[\|Z\| > k]$	Percentage
4	0,99996	0.00006	0,01%
3	0,99865	0.00269	0,27%
2.5	0,99379	0.01242	1,24%
2.2	0,98609	0.02781	2,78%
2	0,97724	0.04550	4,55%
1.96	0,97500	0.05000	5,00%
1.69	0,95448	0.09103	9,10%

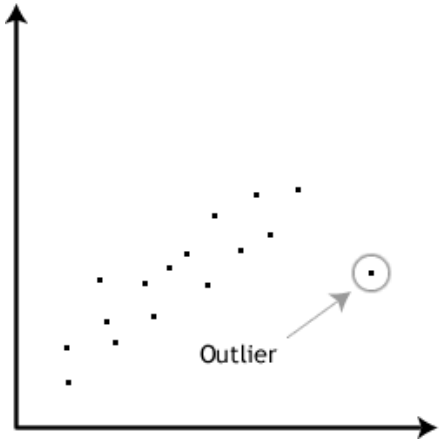
Hipótesis básicas

Statistical prerequisites: Outliers

Determination of outliers:

Case normal: $k\sigma$ method

1. The sample is typified and the point with the greatest (in absolute value) typified value will be classified as outlier if its value is greater than k
2. This point is removed and typified values are again calculated, and we proceed as in 1.
3. The procedure finalize when any typified value is greater than k
4. Typified values are influenced by outliers. So they exist a version to this method that consist to typifying values with the mean and variance obtained excluding this point.



P-variate outliers: Mahalanobis distance

in R:

package ***outliers, randtests***

runs.test

Hipótesis básicas

Statistical prerequisites: Normality

Why is important the data normality?

- It is the distribution of pure random processes.
- Is very adequate to be applied to the description of measurement errors
- Is the underlying statistical model for the majority of statistical analysis.
- Is the underlying hypothesis for the majority of the PAAMs.
- The model is easy to use and well-known.
- The model adequately models bias and dispersion, the two major concerns in spatial error analysis.

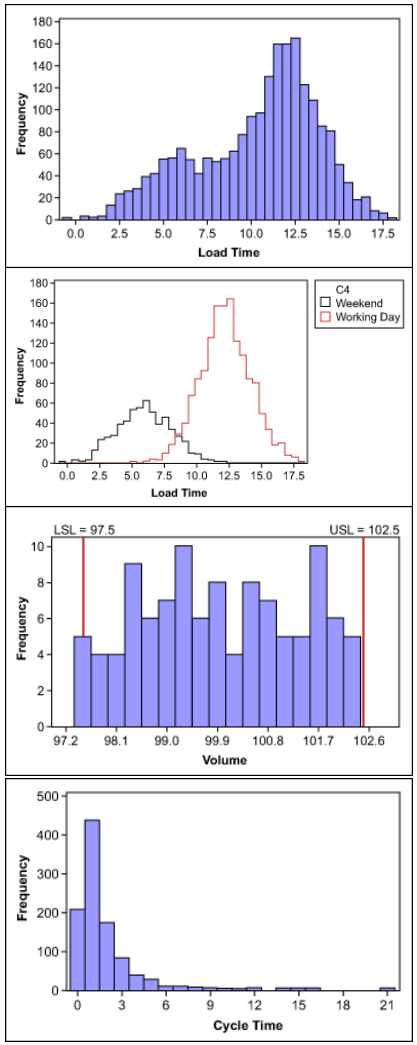
Hipótesis básicas

Statistical prerequisites: Normality

Which is the origin of non-normal distributed data?

There are six main reason:

- Presence of too many extreme values
- Overlap of two or more processes
- Insufficient data discrimination (round-off errors, poor resolution)
- Elimination of data from the sample
- Values close to cero or natural limit
- Data follows a different distribution (e.g. Weibull, log-normal, exponential, Poisson, Binomial, etc.)



Hipótesis básicas

Statistical prerequisites: Normality

How can affect my assessment non-normal distributed data?

- If the underlying hypothesis is normality, non-normal distributed data means that results are totally wrong.

How can be verified?

- They exist many statistical procedures to contrast this hypothesis.
- Some of them are: QQ-plot, Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors, Anderson–Darling, etc.

Hipótesis básicas

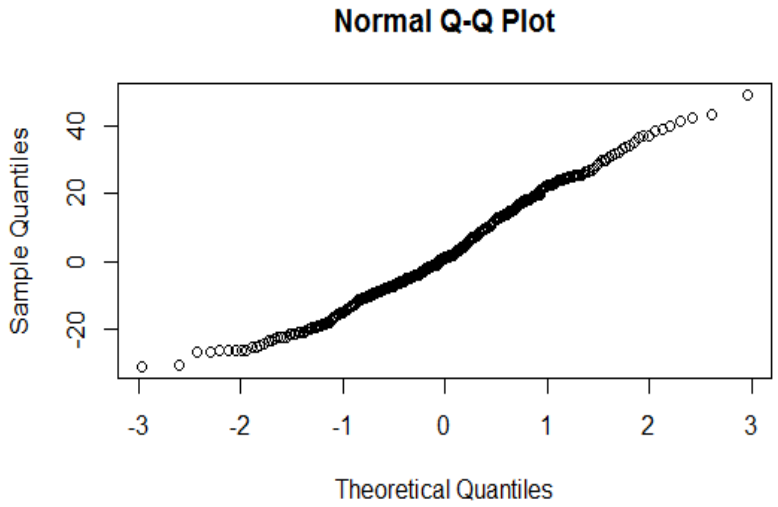
Statistical prerequisites: Normality

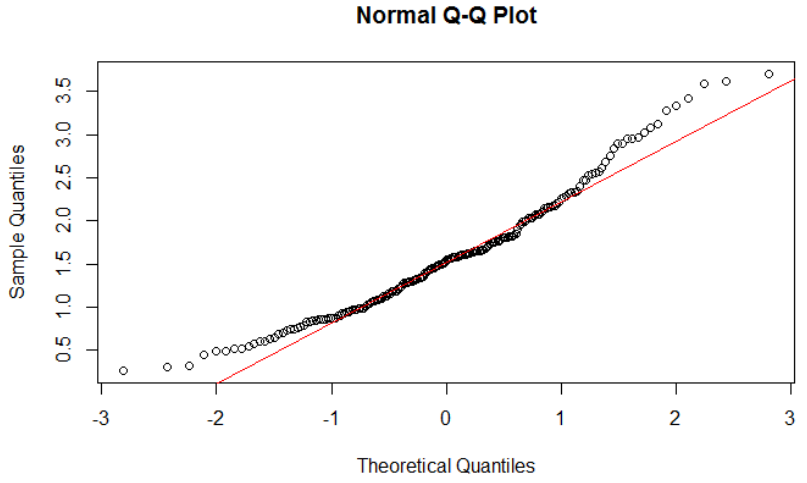
QQ-plot: is a plot (in normal probabilistic paper). If data are normally distributed, then they appear as an straight line in the diagonal.

In R

qqnorm(y, …)

qqline(y, col=#) shows the reference diagonal line





Hipótesis básicas

Statistical prerequisites: Normality

Inferential methods: Are referred to hypothesis test where the null hypothesis is the assumption of normality. They exist many of them. We recall:

Kolmogorov-Smirnov, In R ***fBasic***:

ksnormTest(x, title = NULL, description = NULL
ks.test(x, y="pnorm", alternative = c("two.sided", "less", "greater"), exact = NULL)

Shapiro-Wilks:

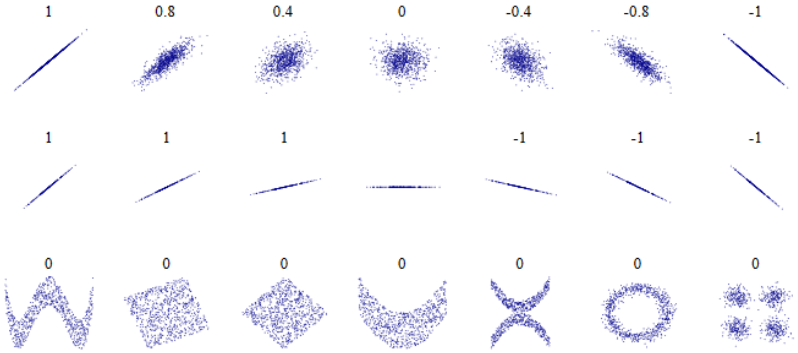
shapiroTest(x, title=NULL, description=NULL)

Hipótesis básicas

Statistical prerequisites: Correlation

Why is important the absence of correlation?

- No correlation means independence between variables.
- No correlation is a logical assumption for many data-capture situations.
- Usually the statistical analysis are based on non correlated data.
- Independence is an assumption of a lot of statistical analysis.
- Independence is an assumption of a lot of PAAMs.
- The presence of correlation means more complex statistical models and calculations.



Hipótesis básicas

Statistical prerequisites: **Correlation**

Which is the origin of correlated data?

- Correlation means dependence or association is any statistical relationship, whether causal or not.
- Correlation does not necessarily mean causation, but some degree of causation can exist.
- Correlation can be originated by structures, processes, operators, weather, devices, etc.

How can affect my assessment the presence of correlated data?

- If the underlying hypothesis is independence, the presence of data correlation means that results are totally wrong.

How can be measured?

- They exist many statistical coefficients to obtain a value for correlations.
- Some of them are: Pearson's Coefficient, Spearman's Coefficient, Kendall coefficient, Quotient correlation, etc.

Hipótesis básicas

Statistical prerequisites: Correlation

How can be verified?

- They exist many statistical procedures to contrast this hypothesis.
- Each correlation coefficient has its own contrast
- In general, we ask for absence of correlation (coefficient equals to 0). *This implies independence only if data are normally distributed!!!!*
- Pearson's Coefficient is the best in presence of normality. Trhe remaining coefficients are useful in other cases (outliers, ordinal measurements, etc).
- In R, we can use

cor(x, y = NULL, use = "everything", method = c("pearson", "kendall", "spearman"))

cor.test(x, y, alternative = c("two.sided", "less", "greater"), method = c("pearson", "kendall", "spearman"), exact = NULL, conf.level = 0.95, continuity = FALSE, ...)

Hipótesis básicas

Statistical prerequisites: Homoscedasticity

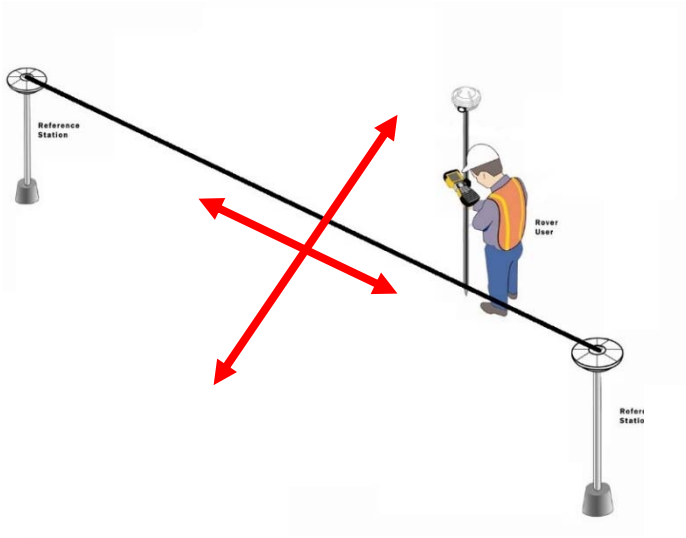
Why is important the homoscedasticity?

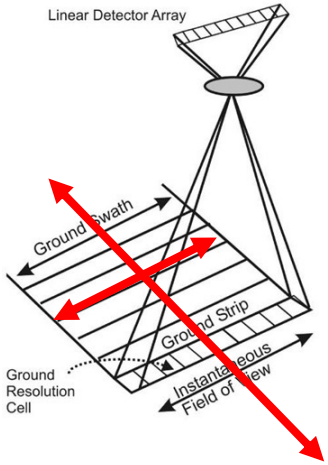
- Homoscedasticity means an uniform behavior in error components (X, Y…)
- Uniform behavior is a logical assumption of isotropic situations.
- Homoscedasticity is an assumption of a lot of statistical analysis.
- Homoscedasticity is an assumption of a lot of PAAMs.
- The absence of homoscedasticity means more complex statistical models and calculations.

Hipótesis básicas

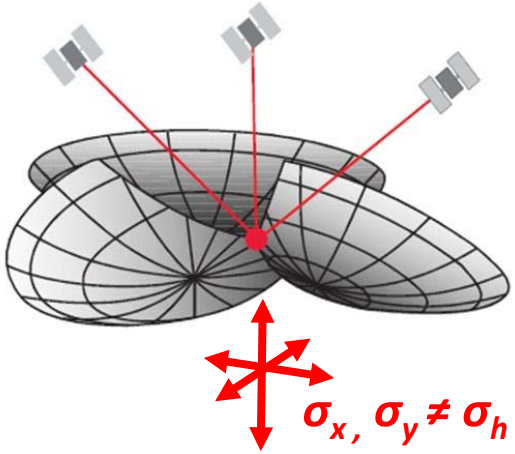
Statistical prerequisites: Homoscedasticity

Which is the origin of heteroscedasticity?





$\sigma_x \neq \sigma_y$



$\sigma_x, \sigma_y \neq \sigma_h$

Hipótesis básicas

Statistical prerequisites: Homoscedasticity

How can affect my assessment the presence heteroscedasticity?

- If the underlying hypothesis is homoscedasticity, heteroscedasticity means that results are totally wrong.
- Regression analysis, analysis of variance, statistical tests of significance, and may other statistical techniques can be affected (bias, ordinary least squares are not the Best Linear Unbiased Estimators, etc.).

How can be verified?

- They exist many statistical procedures to contrast this hypothesis.
- In presence of normality and in two dimensions we can use the F-test (in R, ***var.test(x, ...)***

varianceTest

$$F = \frac{S_x^2}{S_y^2} \sim F_{n-1,n-1}$$

- In other situations, we can use Bartlett's test or Levene's test, among others