Methods of statistical and numerical analysis (integrated course). Part I

Stefano Siboni

Main topics of the course:

(1) **General aspects of experimental measurements**

(2) **Random variables**
   (mean, variance, probability distributions, error propagation)

(3) **Functions of random variables**
   (indirect measurements, error propagation)

(4) **Sample estimates of mean and variance**
   (unbiased estimates of mean and variance confidence intervals)

(5) **Hypothesis testing**
   (tests of hypothesis testing, ANOVA, correlations)

(6) **Pairs of random variables**
   (linear correlation coefficient)

(7) **Data modeling**
   (notions of regression analysis, linear regression by chi-square method and least-squares method, PCA)
1. EXPERIMENTAL ERRORS

- **Measurement error:**
  - difference between the value experimentally measured of a quantity and the “true” value of the same quantity
  - never certain, it can only be estimated

- **Systematic errors:**
  - reproducible errors due to incorrect measurement techniques or to the use of facilities not correctly working or calibrated
  - if recognized, they can be removed or corrected

- **Random errors:**
  - describe the dispersion of results of repeated experimental measurements
  - not related to well-defined and recognized sources
  - not reproducible

- **Absolute error:**
  - estimate of the measurement error of a quantity (of which it has the same physical units)

- **Relative error:**
  - ratio between the absolute error and the estimated true value of a quantity
  - pure number, typically expressed as a percentage
  - it quantifies the precision of a measurement
**Accuracy of a measurement:**
determines how close the result of an experimental measurement can be considered to the “true” value of the measured quantity

\[
\text{accuracy} \iff \text{unimportance of systematic errors}
\]

**Precision of a measurement:**
expresses how exactly the result of a measurement is determined, no matter the “true” value of the measured quantity is. Synonym of **reproducibility**

\[
\text{precision} \iff \text{reproducibility} \\
\iff \text{unimportance of random errors} \iff \text{unimportance of the relative error}
\]

**Postulate of the statistical population:**
In the presence of random errors, the result of a measurement is regarded as the outcome of a random variable, which describes a **statistical population**

**Statistical population:**
infinite and hypothetical set of points (values) of which the experimental results are assumed to be a random sample
Precision vs. accuracy
2. RANDOM VARIABLES

2.1 General definitions and properties

- Random variable (RV) - elementary definition:
  a variable $x$ susceptible to assume at random some real values
  according to a suitable probability distribution

- Discrete RV:
  RV assuming only a finite number or a countable infinity of real values

- Continuous RV:
  RV taking an uncountable infinity of real values, typically any real value of an interval

- Probability distribution of a RV:
  a real nonnegative function $p(x)$ defined on the set of definition of the RV $x$:

  - in the discrete case (probability mass distribution) $p(x_i)$ provides the probability that $x = x_i$;
  
  - in the continuous case (probability density distribution) the integral over the interval $[x_1, x_2] \subseteq \mathbb{R}$

  \[
  \int_{x_1}^{x_2} p(x) \, dx
  \]

  defines the probability of the event $x \in [x_1, x_2]$

In both cases the probability distribution satisfies a condition of normalization
- **Cumulative probability distribution of a RV**
  
  in the continuous case is given by the integral

  \[ P(X) = \int_{-\infty}^{X} p(x) \, dx \]

  and represents the probability that \( x \leq X \), for a given real \( X \).
  
  The discrete case is analogous (replace sum/series to integral)

- **Mean of a RV (probability distribution):**

  \[ \mu = \mathbb{E}(x) = \int_{-\infty}^{+\infty} p(x) \, x \, dx \]

  specifies the mean value of the RV or probability distribution

- **Variance of a RV (probability distribution):**

  \[ \sigma^2 = \mathbb{E}[(x - \mathbb{E}(x))^2] = \int_{-\infty}^{+\infty} p(x) \, (x - \mu)^2 \, dx \]

  specifies how “concentrated” the probability distribution of a RV is around its mean

- **Skewness of a RV (probability distribution):**

  \[ \frac{\mathbb{E}[(x - \mathbb{E}(x))^3]}{\sigma^3} = \frac{1}{\sigma^3} \int_{-\infty}^{+\infty} p(x) \, (x - \mu)^3 \, dx \]

  measures the degree of symmetry of \( p(x) \) with respect to the mean
\[ \require{AMSfonts} \sqrt{\sigma^2} = \mathbb{E}[(x - \mathbb{E}(x))^2]^{1/2} = \left\{ \int_{-\infty}^{+\infty} p(x) (x - \mu)^2 \, dx \right\}^{1/2} \]

has a meaning analogous to that of the variance, but it is homogeneous to \( x \) and \( \mu \) — same physical units!

\boxed{\begin{align*}
\text{Tchebyshev theorem:} & \\
& \text{The probability that a continuous RV takes values in the interval centred on the mean \( \mu \) and of half-width \( k\sigma \), with \( k > 0 \), is not smaller than } 1 - \frac{1}{k^2}:
\end{align*}}

\[ p[\mu - k\sigma \leq x \leq \mu + k\sigma] = \int_{\mu-k\sigma}^{\mu+k\sigma} p(x) \, dx \geq 1 - \frac{1}{k^2} \]

\text{Proof}

\[ \sigma^2 = \int_{-\infty}^{+\infty} p(x)(x - \mu)^2 \, dx \geq \int_{-\infty}^{\mu-k\sigma} p(x)(x - \mu)^2 \, dx + \int_{\mu+k\sigma}^{+\infty} p(x)(x - \mu)^2 \, dx \geq \int_{-\infty}^{\mu-k\sigma} p(x)k^2\sigma^2 \, dx + \int_{\mu+k\sigma}^{+\infty} p(x)k^2\sigma^2 \, dx = k^2\sigma^2 \left[ 1 - \int_{\mu-k\sigma}^{\mu+k\sigma} p(x) \, dx \right] \]

\[ \implies \frac{1}{k^2} \geq 1 - \int_{\mu-k\sigma}^{\mu+k\sigma} p(x) \, dx \quad \square \]
2.2 Examples of discrete RVs

- **Bernoulli variable**
  This RV can assume only a finite number of real values $x_i$, $i = 1, \ldots, n$, with probabilities $w_i = p(x_i)$. The condition of normalization $\sum_{i=1}^{n} w_i = 1$ holds. Moreover:

  $$\mu = \sum_{i=1}^{n} w_i x_i \quad \sigma^2 = \sum_{i=1}^{n} w_i (x_i - \mu)^2$$

- **Binomial variable**
  The probability (mass) distribution is

  $$p(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \quad x = 0, 1, \ldots, n$$

  for given $n \in \mathbb{N}$ and $p \in (0, 1)$. Mean and variance are:

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

  This RV arises in a natural way by considering $n$ independent identically distributed Bernoulli RVs $(x_1, x_2, \ldots, x_n)$, with values $\{1, 0\}$ and probabilities $p$ and $1-p$ respectively: the sum $x = \sum_{i=1}^{n} x_i$ is a binomial variable.

- **Poisson variable**
  It has probability (mass) distribution

  $$p(x) = e^{-m} \frac{m^x}{x!} \quad x = 0, 1, 2, \ldots$$

  with $m$ a positive constant. Mean and variance take the same value:

  $$\mu = m \quad \sigma^2 = m$$
2.3 Examples of continuous RVs

☐ Uniform variable
All the values of the RV $x$ within the interval $[a, b]$ are equally likely, the others are impossible. The probability (density) distribution is:

$$p(x) = \frac{1}{b-a} \begin{cases} 1 & \text{if } x \in [a, b] \\ 0 & \text{otherwise} \end{cases} \quad x \in \mathbb{R}$$

Mean and variance follow immediately:

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

The graph of the uniform distribution is the characteristic function of the interval $[a, b]$
\section*{Normal (or Gaussian) variable}

The probability distribution writes:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad x \in \mathbb{R}$$

where $\mu \in \mathbb{R}$ and $\sigma > 0$ are given constants. The parameter $\mu$ coincides with the mean, while $\sigma^2$ is the variance.

The graph of the distribution is illustrated in the following figure

A normal RV (or distribution) of mean $\mu$ and standard deviation $\sigma$ is sometimes denoted with $N(\mu, \sigma)$

The RV $z = (x - \mu)/\sigma$ still follows a normal probability distribution, but with zero mean and unit variance:

$$p(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \quad z \in \mathbb{R}$$

Such a RV is known as \textbf{standard normal} — $N(0, 1)$ —
Chi-square variable with \( n \) degrees of freedom \((\chi^2)\)

It is conventionally denoted with \(\chi^2\) and its probability distribution takes the form:

\[
p_n(\chi^2) = \frac{1}{\Gamma(n/2) 2^{n/2}} e^{-\chi^2/2} (\chi^2)^{n/2-1} \quad \chi^2 \geq 0
\]

where \(\Gamma\) is the **Euler Gamma function**:

\[
\Gamma(a) = \int_0^{+\infty} e^{-x} x^{a-1} dx \quad a > 0
\]

in turn related to the **factorial function**:

\[
a! = \Gamma(a + 1) = \int_0^{+\infty} e^{-x} x^a dx \quad a > -1
\]

From the characteristic recurrence relation of the function \(\Gamma\)

\[
\Gamma(a + 1) = a \Gamma(a) \quad \forall a > 0
\]

and from the particular values

\[
\Gamma(1) = 1 \quad \Gamma(1/2) = \sqrt{\pi}
\]

one can easily deduce that

\[
\Gamma(n + 1) = n! \quad n = 0, 1, 2, \ldots
\]

\[
\Gamma(n + 1) = n(n-1)(n-2) \ldots (3/2)(1/2)\sqrt{\pi} \quad n = 1/2, 3/2, \ldots
\]

The sum of the squares of \(n\) independent standard normal variables \(z_1, z_2, \ldots, z_n\) is a \(\chi^2\) variable with \(n\) degrees of freedom

\[
\chi^2 = \sum_{i=1}^{n} z_i^2 \quad p(z_i) = \frac{1}{\sqrt{2\pi}} e^{-z_i^2/2} , \quad i = 1, \ldots, n
\]
Graph of the $\chi^2$ probability distribution for an increasing number $n$ of d.o.f.
\[ p_n(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n} \, \Gamma\left(\frac{n}{2}\right)} \frac{1}{\left(1 + \frac{t^2}{n}\right)^{\frac{n+1}{2}}} \quad t \in \mathbb{R} \]

The probability distribution has zero mean and is symmetric with respect to the mean.

For large \( n \) the probability distribution is approximated by a standard normal, to which it tends in the limit \( n \to +\infty \).

Student’s \( t \) with \( n \) degrees of freedom is the probability distribution of the ratio

\[ t = \sqrt{n} \frac{z}{\sqrt{\chi^2}} \]

where:

\( z \) denotes a standard normal RV;

\( \chi^2 \) is a chi-square RV with \( n \) degrees of freedom;

the RVs \( z \) and \( \chi^2 \) are stochastically independent

\[ p(z, \chi^2) = p(z) p_n(\chi^2) \]
Graph of the Student t probability distribution for an increasing number n of d.o.f.
\[ \begin{align*}
\square & \textbf{ Fisher variable with } (n_1, n_2) \text{ degrees of freedom } (F) \\
A \text{ Fisher RV with } (n_1, n_2) \text{ d.o.f. is denoted with } F \text{ and follows the probability distribution} \\
p_{(n_1, n_2)}(F) & = \frac{\Gamma\left(\frac{n_1 + n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right) \Gamma\left(\frac{n_2}{2}\right)} \left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} \frac{F^{\frac{n_1}{2} - 1}}{\left(1 + \frac{n_1}{n_2}F\right)^{\frac{n_1 + n_2}{2}}} \\
\text{with } F & \geq 0. \\
\text{The variable can be written in the form} \\
F & = \frac{n_2}{n_1} \frac{X_1^2}{X_2^2} \\
\text{where:} \\
X_1^2 & \text{ is a chi-square RV with } n_1 \text{ d.o.f.}; \\
X_2^2 & \text{ denotes a chi-square RV with } n_2 \text{ d.o.f.;} \\
\text{the variables } X_1^2 \text{ and } X_2^2 & \text{ are stochastically independent} \\
p(X_1^2, X_2^2) & = p_{n_1}(X_1^2) \ p_{n_2}(X_2^2) \\
\text{We get, in particular, that } \forall (n_1, n_2) \in \mathbb{N}^2 \text{ the reciprocal of a Fisher } F \text{ with } (n_1, n_2) \text{ d.o.f. is distributed as a Fisher } F \text{ with } (n_2, n_1) \text{ d.o.f.:} \\
\frac{1}{F_{(n_1, n_2)}} & = F_{(n_2, n_1)} \\
\end{align*} \]
Graph of the Fisher F probability distribution for a fixed number $n_1$ of d.o.f. and an increasing number $n_2$ of d.o.f.
2.4 Remark. Relevance of normal variables

Random errors affecting many kinds of experimental measurements can be assumed to follow a normal probability distribution with appropriate mean and variance.

Such a large diffusion of normal RVs can be partially justified by the following, well-known result:

☐ Central limit theorem (Lindeberg-Lévy-Turing)

Let \((x_n)_{n \in \mathbb{N}} = x_1, x_2, \ldots\) be a sequence of independent, identically distributed RVs with finite variance \(\sigma^2\) and mean \(\mu\).

For any fixed \(n \in \mathbb{N}\) let us consider the RV

\[
z_n = \frac{1}{\sigma \sqrt{n}} \sum_{i=1}^{n} (x_i - \mu) \quad x_i \in \mathbb{R}
\]

whose probability distribution will be denoted with \(p_n(z_n)\).

We have then

\[
p_n(z) \xrightarrow[n \to +\infty]{} N(0, 1)(z) \quad \forall z \in \mathbb{R}
\]

being \(N(0, 1)(z)\) the standard normal probability distribution

\[
N(0, 1)(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \quad z \in \mathbb{Z}
\]

☐ Note:

We say that the sequence of RVs \((z_n)_{n \in \mathbb{N}}\) converges in distribution to the standard normal \(z\).

The condition that variables are identically distributed can be weakened (\(\Longrightarrow\) more general results).
A sum of \( n \) continuous, independent and identically distributed RVs follows a distribution which is **approximately normal** for \( n \) large enough.

Tipically this happens as \( n \) is of the order of some tens, **but it may be enough much less** (it depends on the common probability distribution of the RVs).

**Example: RVs with uniform distribution in \([0, 1]\)**

By summing 1, 2, 3, 4 independent RVs of the same kind we obtain RVs with the probability distributions below:
2.5 Probability distributions in more dimensions

They are the so-called **joint probability distributions** of two or more RVs (discrete or continuous).

The \( n \)-dimensional (or joint) probability distribution

\[
p(x_1, x_2, \ldots, x_n)
\]

of the RVs \( x_1, x_2, \ldots, x_n \):

- **(●) in the discrete case** describes the probability that the \( n \) variables take **simultaneously** the values indicated in the arguments (composed event);

- **(●) in the continuous case** determines, by means of the \( n \)-dimensional integral

\[
\int_{\Omega} p(x_1, x_2, \ldots, x_n) \, dx_1 dx_2 \ldots dx_n,
\]

the probability that the vector of RVs \( (x_1, \ldots, x_n) \) takes any value in the domain of integration \( \Omega \subseteq \mathbb{R}^n \).

In both cases the distribution satisfies a condition of normalization, that for continuous variables takes the form

\[
\int_{\mathbb{R}^n} p(x_1, x_2, \ldots, x_n) \, dx_1 dx_2 \ldots dx_n = 1
\]

while a similar expression holds for discrete variables.
2.5.1 (Stochastically) independent RVs:
the $n$ RVs $x_1, x_2, \ldots, x_n$ are called independent if their joint
probability distribution can be written as a product of $n$ distributions, one per each variable

$$p(x_1, x_2, \ldots, x_n) = p_1(x_1) p_2(x_2) \ldots p_n(x_n)$$

The realization of a given value of one variable of the set does
not affect the joint probability distribution of the residual vari-
ables (conditional probability)

**Example:**

$$p(x_1, x_2) = \frac{1}{2\pi \sigma_1 \sigma_2} e^{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2} - \frac{(x_2-\mu_2)^2}{2\sigma_2^2}} =$$

$$= \frac{1}{\sqrt{2\pi} \sigma_1} e^{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}} \cdot \frac{1}{\sqrt{2\pi} \sigma_2} e^{-\frac{(x_2-\mu_2)^2}{2\sigma_2^2}}$$

describes a pair of independent Gaussian variables, $x_1$ e $x_2$

2.5.2 (Stochastically) dependent RVs:
are not independent RVs. Their joint probability distribution
cannot be factorized as in the case of independent variables.

**Example:**
the RVs $x, y$ with the joint probability distribution

$$p(x, y) = \frac{3}{2\pi} e^{-x^2 - xy - y^2}$$

are stochastically dependent
2.5.3 Mean, covariance, correlation of a set of RVs

For a given set \((x_1, \ldots, x_n)\) of RVs, we define the following quantities

(●) **Mean of** \(x_i\):

\[
\mu_i = \mathbb{E}(x_i) = \int_{\mathbb{R}^n} x_i p(x_1, \ldots, x_n) \, dx_1 \ldots dx_n \in \mathbb{R}
\]

(●) **Covariance of** \(x_i\) and \(x_j\), \(i \neq j\):

\[
\text{cov}(x_i, x_j) = \mathbb{E}[(x_i - \mu_i)(x_j - \mu_j)] = \\
= \int_{\mathbb{R}^n} (x_i - \mu_i)(x_j - \mu_j) p(x_1, \ldots, x_n) \, dx_1 \ldots dx_n \in \mathbb{R}
\]

(●) **Variance of** \(x_i\):

\[
\text{var}(x_i) = \mathbb{E}[(x_i - \mu_i)^2] = \\
= \int_{\mathbb{R}^n} (x_i - \mu_i)^2 p(x_1, \ldots, x_n) \, dx_1 \ldots dx_n > 0
\]

(●) **Covariance matrix of the set:**

is the \(n \times n\), real and symmetric matrix \(C\) of entries

\[
C_{ij} = \begin{cases} 
\text{cov}(x_i, x_j) & \text{if } i \neq j \\
\text{var}(x_i) & \text{if } i = j 
\end{cases} \quad i, j = 1, 2, \ldots, n
\]
(●) **Correlation of** $x_i$ **and** $x_j$, $i \neq j$:

$$\text{corr}(x_i, x_j) = \frac{\text{cov}(x_i, x_j)}{\sqrt{\text{var}(x_i)} \sqrt{\text{var}(x_j)}} \in [-1, +1]$$

The correlation of a variable with itself is defined but not meaningful, since $\text{corr}(x_i, x_i) = 1 \ \forall \ i = 1, \ldots, n$

(●) **Correlation matrix of the set:**

is the $n \times n$, real symmetric matrix $V$, given by

$$V_{ij} = \text{corr}(x_i, x_j) \quad i, j = 1, 2, \ldots, n$$

with the diagonal entries equal to 1

☐ **Uncorrelated RVs:**

RVs with null correlation

$x_i$ and $x_j$ uncorrelated $\iff \text{corr}(x_i, x_j) = 0$

$(x_1, \ldots, x_n)$ uncorrelated $\iff V$ diagonal

$\iff C$ diagonal

☐ **Theorem**

The stochastic independence is a **sufficient** but **not necessary condition** to the uncorrelation of two or more RVs.

For instance, if $x_1$ and $x_2$ are independent we get

$$\text{cov}(x_1, x_2) =$$

$$= \int_{\mathbb{R}^2} (x_1 - \mu_1)(x_2 - \mu_2)p_1(x_1) p_2(x_2) \, dx_1 dx_2 =$$

$$= \int_{\mathbb{R}} (x_1 - \mu_1) p_1(x_1) \, dx_1 \int_{\mathbb{R}} (x_2 - \mu_2) p_2(x_2) \, dx_2 = 0$$
2.5.4 Multivariate normal (or Gaussian) RVs

A vector $x^T = (x_1, \ldots, x_n)$ of RVs constitutes a set of multivariate normal RVs if the joint probability distribution is of the form

$$p(x_1, \ldots, x_n) = \frac{(\det A)^{1/2}}{(2\pi)^{n/2}} e^{-\frac{1}{2}(x-\mu)^T A(x-\mu)}$$

where $A$ is a $n \times n$ real symmetric positive definite matrix (structure matrix), $\det A > 0$ its determinant and $\mu \in \mathbb{R}^n$ an arbitrary column vector (the superscript $^T$ denotes the transpose of a matrix or column vector).

Mean of $x$:

$$\mathbb{E}(x) = \int_{\mathbb{R}^n} x \frac{(\det A)^{1/2}}{(2\pi)^{n/2}} e^{-\frac{1}{2}(x-\mu)^T A(x-\mu)} \, dx_1 \ldots dx_n = \mu$$

Covariance matrix of $x$:

coincides with the inverse of the structure matrix

$$C = A^{-1}$$

☐ Theorem

For multivariate normal RVs, stochastic independence is equivalent to uncorrelation

$$(x_1, \ldots, x_n) \text{ uncorrelated} \iff C \text{ diagonal} \iff$$

$$\iff A \text{ diagonal} \iff (x_1, \ldots, x_n) \text{ independent}$$
**Example:** Let us consider the vector of the mean values and the structure matrix

\[ \mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \in \mathbb{R}^2 \quad A = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \]

the latter being real, symmetric and positive definite (both its eigenvalues are positive)

\[
\det(A - \lambda \mathbb{I}) = \det \begin{pmatrix} 2 - \lambda & 1 \\ 1 & 1 - \lambda \end{pmatrix} = \lambda^2 - 3\lambda + 1 = 0
\]

\[ \implies \lambda = \lambda_{\pm} = \frac{3 \pm \sqrt{5}}{2} > 0 \]

The probability distribution of the bivariate RV \((x, y)\) takes the form

\[
p(x, y) = \frac{\sqrt{1}}{(2\pi)^{2/2}} e^{-\frac{1}{2}(x y) \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}} = \frac{e^{-\frac{1}{2}(2x^2 + 2xy + y^2)}}{2\pi}
\]

and has the graph shown in the following figure.
**Warning:** If two RVs $x$ and $y$ are normal, the pair $(x, y)$ **may not be** a bivariate normal RV!

**1st counterexample:**
If $x$ is standard normal and $y = -x$, then $y$ is standard normal

$$p_y(y) = p_x(x) \left. \frac{dx}{dy} \right|_{x=-y} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(-y)^2} \left| -1 \right| = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}$$

but $(x, y)$ is not a bivariate normal, due to the joint distribution

$$p(x, y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \delta(x + y),$$

where $\delta(x + y)$ denotes a Dirac delta

**2nd counterexample:**
For a standard normal $x$, the RV

$$y = \begin{cases} -x & \text{if } |x| \leq 1 \\ x & \text{if } |x| > 1 \end{cases}$$

is a standard normal as in the previous counterexample, but $(x, y)$ is not a bivariate normal, because the joint probability distribution has not the correct form

$$p(x, y) = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \delta(y + x) & \text{if } |x| \leq 1 \\ \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \delta(y - x) & \text{if } |x| > 1 \end{cases}$$
3. INDIRECT MEASUREMENTS
FUNCTIONS OF RANDOM VARIABLES

3.1 Generalities

(•) Direct measurement:
obtained by a comparison with a specimen or the use of a calibrated instrument

(•) Indirect measurement:
performed by a calculation, from a relationship among quantities which are in turn measured directly or indirectly

In the presence of random errors, the result of an indirect measurement must be regarded as a RV which is a function of other RVs

\[ x = f(x_1, x_2, \ldots, x_n) \]

The joint probability distribution of the RVs \((x_1, \ldots, x_n)\) being known, we want to determine that of \(x\)

\[
\text{joint distribution of } (x_1, x_2, \ldots, x_n) \quad \Rightarrow \quad \text{distribution of } x = f(x_1, x_2, \ldots, x_n)
\]
The probability distribution of the function

\[ x = f(x_1, x_2, \ldots, x_n) \]

of the RVs \( x_1, x_2, \ldots, x_n \) can be calculated in an explicit way only in very particular cases, if the joint probability distribution of \( p(x_1, x_2, \ldots, x_n) \) is given.

It often happens that the joint probability distribution of \( x_1, x_2, \ldots, x_n \) is not known, but that (estimates of) means, variances and covariances are available.

It may even happen that we only have an estimate of the true values \( \bar{x}_1, \ldots, \bar{x}_n \) and of the errors \( \Delta x_1, \ldots, \Delta x_n \)

\[ \Downarrow \]

It may be impossible to reckon explicitly the probability distribution of \( x \).

In that case, we must confine ourselves to more limited goals:

1. calculate mean and variance of the random function \( x \);
2. estimate only the true value and the error of \( x \);
3. obtain a numerical approximation of the probability distribution of \( x \) by a Monte-Carlo method.
3.2 Linear combination of RVs

For given RVs \((x_1, x_2, \ldots, x_n) = x\), let us consider the RV defined by the linear combination

\[ z = \sum_{i=1}^{n} a_i x_i = a^T x \]

where \(a_i\) are real constants and \(a^T = (a_1, \ldots, a_n)\). We have therefore that:

(\bullet) the mean of \(z\) is the linear combination, with the same coefficients \(a_i\), of the means of the RVs

\[ \mathbb{E}(z) = \mathbb{E} \left[ \sum_{i=1}^{n} a_i x_i \right] = \sum_{i=1}^{n} \mathbb{E}(a_i x_i) = \sum_{i=1}^{n} a_i \mathbb{E}(x_i) \]

(\bullet) the variance of \(z\) is given by the expression

\[ \mathbb{E}[(z - \mathbb{E}(z))^2] = \mathbb{E} \left[ \sum_{i,j=1}^{n} a_i [x_i - \mathbb{E}(x_i)] a_j [x_j - \mathbb{E}(x_j)] \right] = \]

\[ = \sum_{i,j=1}^{n} a_i a_j \mathbb{E}[[x_i - \mathbb{E}(x_i)][x_j - \mathbb{E}(x_j)]] = \]

\[ = \sum_{i,j=1}^{n} a_i a_j C_{ij} = a^T C a \]

where \(C\) denotes the covariance matrix of the variables \(x\). For uncorrelated variables (and better if independent)

\[ \mathbb{E}[(z - \mathbb{E}(z))^2] = \sum_{i=1}^{n} a_i^2 \text{var}(x_i) \]
3.3 Linear combination of multivariate normal RVs

Let \((x_1, x_2, \ldots, x_n) = x\) be a set of multivariate normal RVs with a joint probability distribution specified by the structure matrix \(A\) and by the mean value \(\mu \in \mathbb{R}^n\).

Then:

\((\bullet)\) given an arbitrary real nonsingular \(n \times n\) matrix \(R\), the set of RVs defined by \((y_1, \ldots, y_n) = y = Rx\) still constitutes a set of multivariate normal RVs, with joint probability distribution

\[
p(y) = \frac{[\det[(R^{-1})^T A R^{-1}]]^{1/2}}{(2\pi)^{n/2}} e^{-\frac{1}{2}(y-R\mu)^T (R^{-1})^T A R^{-1} (y-R\mu)}
\]

and therefore mean \(R\mu\) and structure matrix \((R^{-1})^T A R^{-1}\);

\((\bullet)\) as a particular case of the previous result, if the normal variables \(x\) are stochastically independent and standard and if moreover the matrix \(R\) is orthogonal \((R^T = R^{-1})\), the RVs \(y = Rx\) are in turn normal, independent and standard;

\((\bullet)\) for an arbitrary nonzero vector \(a^T = (a_1, \ldots, a_n) \in \mathbb{R}^n\), the linear combination

\[
z = a^T x = \sum_{i=1}^{n} a_i x_i
\]

is a normal RV with mean and variance

\[
\mathbb{E}(z) = a^T \mu \quad \mathbb{E}[(z - \mathbb{E}(z))^2] = a^T Ca = a^T A^{-1} a;
\]
(●) if $B$ is a $m \times n$ matrix with $m \leq n$ and of maximum rank (i.e., the $m$ rows of $B$ are linearly independent), then the $m$ RVs $y = (y_1, \ldots, y_m)$ defined by the linear transformation

$$y = Bx$$

are not only normal, but form also a set of $m$ multivariate normal RVs with a mean value vector

$$\mathbb{E}(Bx) = B \mathbb{E}(x) = B\mu,$$

a (positive definite) covariance matrix

$$\mathbb{E}[(Bx - B\mu)(Bx - B\mu)^T] = \mathbb{E}[B(x - \mu)(x - \mu)^TB^T] = B\mathbb{E}[(x - \mu)(x - \mu)^TB] = BA^{-1}B^T$$

and a (positive definite) structure matrix

$$(BA^{-1}B^T)^{-1};$$

(●) any subset of $(x_1, \ldots, x_n)$ is still a set of multivariate normal RVs. The result follows from $(iv)$ by choosing an appropriate matrix $B$. As an illustration, the subset $(x_1, x_3)$ would be defined by the linear transformation $y = Bx$ with

$$B = \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \end{pmatrix}$$

which is a $2 \times n$ matrix of rank 2.
3.4 Quadratic forms of standard variables

3.4.1 Craig theorem
(independent quadratic forms of independent standard variables)
Let $x$ be a vector of $n$ independent standard variables. The real symmetric matrices $A$ and $B$, let us consider the quadratic forms

$$Q_1 = x^T Ax$$
$$Q_2 = x^T Bx$$

The RVs $Q_1$ and $Q_2$ are then stochastically independent if and only if $AB = 0$.

3.4.2 Characterization theorem of chi-square RVs
(quadratic forms of independent standard variables which follow a chi-square distribution)
Let $x^T Ax$ be a positive semidefinite quadratic form of the independent standard variables $(x_1, x_2, \ldots, x_n) = x$.
The RV $x^T Ax$ is then a $\chi^2$ variable if and only if the matrix $A$ is idempotent

$$A^2 = A$$

In such a case the number of d.o.f. of $x^T Ax$ coincides with the rank $p$ of the matrix $A$. 
**Example:** let $x_1$ and $x_2$ be two independent standard normal RVs. Let us consider the RVs

\[ Q_1 = 2x_1x_2 = (x_1 \ x_2) A \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \]

\[ Q_2 = x_1^2 - 4x_1x_2 + x_2^2 = (x_1 \ x_2) B \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \]

with

\[ A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 1 & -2 \\ -2 & 1 \end{pmatrix} \]

We have that:

\[ AB = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & -2 \\ -2 & 1 \end{pmatrix} = \begin{pmatrix} -2 & 1 \\ 1 & -2 \end{pmatrix} \neq 0 \]

and therefore the RVs $Q_1$ and $Q_2$ are **stochastically dependent** by Craig’s theorem.

On the other hand, there holds

\[ A^2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \neq A \]

and analogously

\[ B^2 = \begin{pmatrix} 1 & -2 \\ -2 & 1 \end{pmatrix} \begin{pmatrix} 1 & -2 \\ -2 & 1 \end{pmatrix} = \begin{pmatrix} 5 & -4 \\ -4 & 5 \end{pmatrix} \neq B \]

so that **neither** $Q_1$ **nor** $Q_2$ **are** $\chi^2$ RVs.
3.4.3 Fisher-Cochran theorem

Let \((x_1, x_2, \ldots, x_n) = x\) be a set of independent standard RVs. Let \(Q_1, Q_2, \ldots, Q_k\) be positive semidefinite quadratic forms of the variables \((x_1, x_2, \ldots, x_n)\)

\[
Q_r = x^T A_r x = \sum_{i,h=1}^{n} (A_r)_{ih} x_i x_h \quad r = 1, 2, \ldots, k
\]

with matrices \(A_1, A_2, \ldots, A_k\), of rank \(n_1, n_2, \ldots, n_k\) respectively. Moreover, let

\[
\sum_{i=1}^{n} x_i^2 = \sum_{j=1}^{k} Q_j
\]

\(Q_1, Q_2, \ldots, Q_k\) are then mutually independent \(\chi^2\) RVs if and only if

\[
\sum_{j=1}^{k} n_j = n
\]

and in that case \(Q_r\) has \(n_r\) d.o.f., \(\forall r = 1, \ldots, k\)
### 3.4.4 Two-variable cases

(i) Let
\[ Q_1 = x^T A_1 x \quad Q_2 = x^T A_2 x \]
be two positive semidefinite quadratic forms of the independent standard variables \((x_1, x_2, \ldots, x_n) = x\). Let the RVs \(Q_1\) and \(Q_2\) satisfy the condition
\[ Q_1 + Q_2 = \sum_{i=1}^{n} x_i^2 \]
If \(Q_1\) is a \(\chi^2\) RV with \(p < n\) d.o.f., \(Q_2\) is then a \(\chi^2\) RV with \(n-p < n\) d.o.f. and the two RVs are stochastically independent.

**Proof**
It follows from Craig’s theorem and the characterization theorem of \(\chi^2\) variables, by noting that \(A_1 + A_2 = I\).

(ii) Let \(Q, Q_1\) and \(Q_2\) be three positive semidefinite quadratic forms of the independent standard variables \((x_1, x_2, \ldots, x_n) = x\). Let the RVs \(Q, Q_1\) and \(Q_2\) obey the condition
\[ Q_1 + Q_2 = Q \]
If \(Q\) and \(Q_1\) are \(\chi^2\) RVs, with \(n\) and \(p < n\) d.o.f. respectively, then \(Q_2\) is a \(\chi^2\) RV with \(n-p < n\) d.o.f. and stochastically independent on \(Q_1\).

**Proof**
This theorem can be easily reduced to the previous one (as we will discuss in the following).
3.5 Error propagation in indirect measurements

3.5.1 Gauss law

If:

- variances and covariances of the RVs $x_1, \ldots, x_n$ are known and sufficiently small,
- the means $\mu_1, \ldots, \mu_n$ of the same RVs are given,
- the function $f$ is sufficiently smooth (e.g. $C^2$),

then the mean and the variance of $x = f(x_1, x_2, \ldots, x_n)$ can be estimated by using the Taylor approximation:

$$x = f(\mu_1, \ldots, \mu_n) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\mu_1, \ldots, \mu_n)(x_i - \mu_i)$$

and hold therefore

$$\mathbb{E}(x) = f(\mu_1, \ldots, \mu_n)$$

$$\mathbb{E}[(x - \mathbb{E}(x))^2] = \sum_{i,j=1}^{n} \frac{\partial f}{\partial x_i}(\mu_1, \ldots, \mu_n) \frac{\partial f}{\partial x_j}(\mu_1, \ldots, \mu_n) C_{ij}$$

being $C$ the covariance matrix

For uncorrelated (or independent) RVs the following relationship holds

$$\mathbb{E}[(x - \mathbb{E}(x))^2] = \sum_{i=1}^{n} \left[ \frac{\partial f}{\partial x_i}(\mu_1, \ldots, \mu_n) \right]^2 \text{var}(x_i)$$

which expresses the so-called Gauss law of propagation of random errors
**Example: An application of Gauss law**

Let \( x \) and \( y \) be two independent RVs such that:

- \( x \) has mean \( \mu_x = 0.5 \) and variance \( \sigma^2_x = 0.1 \)
- \( y \) has mean \( \mu_y = 0.2 \) and variance \( \sigma^2_y = 0.5 \)

Let us consider the RV

\[
z = f(x, y) = x \sin y + x^2 y
\]

The first partial derivatives of the function \( f(x, y) \) are

\[
\frac{\partial f}{\partial x}(x, y) = \sin y + 2xy
\]
\[
\frac{\partial f}{\partial y}(x, y) = x \cos y + x^2
\]

and in the mean values \( (x, y) = (\mu_x, \mu_y) = (0.5, 0.2) \) they hold

\[
\frac{\partial f}{\partial x}(0.5, 0.2) = \sin 0.2 + 2 \cdot 0.5 \cdot 0.2 = 0.398669331
\]
\[
\frac{\partial f}{\partial y}(0.5, 0.2) = 0.5 \cdot \cos 0.2 + 0.5^2 = 0.740033289
\]

Thus, the RV \( z = f(x, y) \) has mean

\[
\mu_z = f(0.5, 0.2) = 0.5 \cdot \sin 0.2 + 0.5^2 \cdot 0.2 = 0.1493346654
\]

and variance

\[
\sigma^2_z = 0.398669331^2 \cdot 0.1 + 0.740033289^2 \cdot 0.5 = 0.2897183579
\]
3.5.2 Logarithmic differential

If no statistical information about the quantities \( x_1, x_2, \ldots, x_n \) is available, but only the estimated true values \( \bar{x}_1, \ldots, \bar{x}_n \) and errors \( \Delta x_1, \ldots, \Delta x_n \) are known

we can calculate the true value of \( x = f(x_1, x_2, \ldots, x_n) \) as

\[
\bar{x} = f(\bar{x}_1, \ldots, \bar{x}_n)
\]

and the error \( \Delta x \) by the differential estimate

\[
\Delta x = \sum_{i=1}^{n} \left| \frac{\partial f}{\partial x_i}(\bar{x}_1, \ldots, \bar{x}_n) \right| \Delta x_i
\]

Such a relationship is often deduced by reckoning the differential of the natural logarithm of \( f \)

\[
\frac{\Delta x}{\bar{x}} = \sum_{i=1}^{n} \left| \frac{\partial \ln f}{\partial x_i}(\bar{x}_1, \ldots, \bar{x}_n) \right| \Delta x_i \tag{1}
\]

particularly useful when \( f \) is a product of factors. This approach is then known as logarithmic differential method

In the design of an experiment the logarithmic differential method is often used by following the so-called principle of equivalent effects.

This means that the experiment is designed, and the errors \( \Delta x_i \) eventually reduced, in such a way that all the terms in the sum (1) have the same order of magnitude.
Example: An application of the logarithmic differential

The Reynolds number of a fluid dynamic system is defined by the relationship

\[ R = \frac{\rho v \ell}{\eta} \]

where:

- \( \eta = (1.05 \pm 0.02) \cdot 10^{-3} \text{ kg m}^{-1}\text{s}^{-1} \) (coefficient of dynamic viscosity of the liquid)
- \( \rho = (0.999 \pm 0.001) \cdot 10^{3} \text{ kg m}^{-3} \) (density of the liquid)
- \( \ell = (0.950 \pm 0.005) \text{ m} \) (a characteristic length)
- \( v = (2.5 \pm 0.1) \text{ m s}^{-1} \) (a characteristic speed)

The presumable true value of \( R \) is calculated by the assumed true values of \( \rho, v, \ell, \eta \):

\[ R = \frac{0.999 \cdot 10^{3} \cdot 2.5 \cdot 0.950}{1.05 \cdot 10^{-3}} = 2.259642857 \cdot 10^{6} \]

and the relative error estimated by the logarithmic differential:

\[ \frac{\Delta R}{R} = \frac{\Delta \rho}{\rho} + \frac{\Delta v}{v} + \frac{\Delta \ell}{\ell} + \frac{\Delta \eta}{\eta} = \]

\[ = \frac{0.001}{0.999} + \frac{0.1}{2.5} + \frac{0.005}{0.950} + \frac{0.02}{1.05} = 0.06531177795 \]

The relative error on Reynolds number is then 6.5% and corresponds to an absolute error

\[ \Delta R = R \cdot \frac{\Delta R}{R} = 2.259642857 \cdot 10^{6} \cdot 0.06531177795 = \]

\[ = 0.1475812925 \cdot 10^{6} \]
3.5.3 Error propagation in solving a set of linear algebraic equations

Let us consider the set of linear algebraic equations

$$Ax = b$$

where $A$ is a $n \times n$ nonsingular matrix and $b$ a column vector of $\mathbb{R}^n$.

Suppose that the entries of $A$ and $b$ are affected by some errors, expressed by means of suitable matrices $\Delta A$ and $\Delta b$.

As a consequence, the solution $x \in \mathbb{R}^n$ of the equations will be in turn affected by a certain error $\Delta x$.

In the (usual) hypothesis that the errors $\Delta A$ on $A$ are small, the following estimate of the error $\Delta x$ on $x$ holds

$$\frac{|\Delta x|}{|x|} \leq \text{cond}(A) \left( \frac{|\Delta b|}{|b|} + \frac{\|\Delta A\|}{\|A\|} \right) \frac{1}{1 - \text{cond}(A) \frac{\|\Delta A\|}{\|A\|}}$$

where:

$|\cdot|$ denotes any **vector norm** of $\mathbb{R}^n$;

$\|\cdot\|$ stands for the **subordinate matrix norm** of the previous vector norm $|\cdot|$;

$\text{cond}(A) = \|A\| \|A^{-1}\|$ is the so-called **condition number** of the matrix $A$.
Some examples of vector norms commonly used

For an arbitrary vector \( x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n \) we can define the vector norms listed below:

- the **1-norm** or \( \ell^1 \)-norm, given by
  \[
  |x|_1 = \sum_{i=1}^{n} |x_i|
  \]

- the **2-norm** or \( \ell^2 \)-norm, which is the usual Euclidean norm
  \[
  |x|_2 = \left[ \sum_{i=1}^{n} x_i^2 \right]^{1/2}
  \]

- the **infinity norm** or \( \ell^\infty \)-norm, expressed as
  \[
  |x|_\infty = \max_{i=1,\ldots,n} |x_i|
  \]

Subordinate matrix norms

The matrix norm \( \| \cdot \| \) subordinated to a vector norm \( | \cdot | \) is formally defined as

\[
\|A\| = \inf \{ c \in \mathbb{R} ; |Ax| \leq c|x| \forall x \in \mathbb{R}^n \}
\]

where \( \inf \) denotes the infimum of a real subset.
The definition is highly nontrivial, but for the previous vector norms the calculation of the subordinate matrix norms is quite simple. Indeed, for any square matrix $A$:

- the matrix norm subordinated to the vector norm $\| \cdot \|_1$ turns out to be

  \[ \| A \|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^{n} |A_{ij}| \]

  and therefore is obtained by adding the absolute values of the entries of each column of $A$ and taking the largest result;

- the matrix norm subordinated to the vector norm $\| \cdot \|_2$ is written as

  \[ \| A \|_2 = \sqrt{\lambda} \]

  where $\lambda$ denotes the largest eigenvalue of the real, symmetric, positive semidefinite matrix $A^T A$;

- the matrix norm subordinated to the vector norm $\| \cdot \|_\infty$ is

  \[ \| A \|_\infty = \max_{i=1,\ldots,n} \sum_{j=1}^{n} |A_{ij}| \]

  and can be obtained by adding the absolute values of the entries in each row and choosing the largest of the results.
Example: Error propagation in the solution of a linear system of two algebraic equations in two variables

Let us consider the system of equations

\[
\begin{align*}
    a_{11} x + a_{12} y &= b_1 \\
    a_{21} x + a_{22} y &= b_2
\end{align*}
\]

whose coefficients are affected by an error:

\[
\begin{align*}
    a_{11} &= 5.0 \pm 0.1 & a_{12} &= 6.2 \pm 0.1 \\
    a_{21} &= 3.5 \pm 0.1 & a_{22} &= 10.5 \pm 0.2 \\
    b_1 &= 12.0 \pm 0.2 & b_2 &= 15.2 \pm 0.1
\end{align*}
\]

The matrices to be taken into account are the following ones:

\[
A = \begin{pmatrix}
5.0 & 6.2 \\
3.5 & 10.5
\end{pmatrix} \quad \Delta A = \begin{pmatrix}
0.1 & 0.1 \\
0.1 & 0.2
\end{pmatrix}
\]

\[
b = \begin{pmatrix}
12.0 \\
15.2
\end{pmatrix} \quad \Delta b = \begin{pmatrix}
0.2 \\
0.1
\end{pmatrix}
\]

We must determine the inverse matrix of \( A \)

\[
A^{-1} = \frac{1}{5.0 \cdot 10.5 - 3.5 \cdot 6.2} \begin{pmatrix}
10.5 & -6.2 \\
-3.5 & 5.0
\end{pmatrix} = 
\]

\[
= \begin{pmatrix}
0.34090909 & -0.20129870 \\
-0.11363636 & 0.16233766
\end{pmatrix}
\]

by means of the determinant

\[
det A = 5.0 \cdot 10.5 - 3.5 \cdot 6.2 = 30.80
\]
The estimated solution of the linear system is derived by using the mean values of the coefficients:

\[
x = \begin{pmatrix} x \\ y \end{pmatrix} = A^{-1} b =
\begin{pmatrix}
0.34090909 & -0.20129870 \\
-0.11363636 & 0.16233766
\end{pmatrix}
\begin{pmatrix} 12.0 \\ 15.2 \end{pmatrix} =
\begin{pmatrix}
1.03116884 \\
1.10389611
\end{pmatrix}
\]

Let us estimate the error propagation by using the \( \ell^\infty \) norms. We get:

- \( |b|_\infty = \max\{12.0, 15.2\} = 15.2 \)
- \( |\Delta b|_\infty = \max\{0.2, 0.1\} = 0.2 \)
- \( \|A\|_\infty = \max\{11.2, 14.0\} = 14.0 \) since

\[
\begin{pmatrix}
|5.0| & |6.2| \\
|3.5| & |10.5|
\end{pmatrix} \rightarrow \text{sum} 11.2
\]

\[
\begin{pmatrix}
|0.1| & |0.1| \\
|0.1| & |0.2|
\end{pmatrix} \rightarrow \text{sum} 0.2
\]

- \( \|\Delta A\|_\infty = \max\{0.2, 0.3\} = 0.3 \) because

\[
\begin{pmatrix}
|+0.34090909| & |-0.20129870| \\
|-0.11363636| & |+0.16233766|
\end{pmatrix} \rightarrow \text{sum} 0.54220779 \]

\[
\begin{pmatrix}
| +0.34090909 | & | -0.20129870 | \\
| -0.11363636 | & | +0.16233766 |
\end{pmatrix} \rightarrow \text{sum} 0.27597402
\]
As a consequence, the condition number turns out to be

$$\text{cond}(A) = \|A\|_{\infty} \cdot \|A^{-1}\|_{\infty} = \frac{14.0 \cdot 0.54220779}{0.31354462} = 7.59090906$$

The norm of the solution finally holds

$$|x|_{\infty} = \max\{1.03116884, 1.10389611\} = 1.10389611$$

The error propagation on the solution is then estimated by

$$\frac{\|\Delta x\|_{\infty}}{|x|_{\infty}} = \max\{\|\Delta x\|, \|\Delta y\|\} \leq \text{cond}(A)(\frac{\|\Delta b\|_{\infty}}{|b|_{\infty}} + \frac{\|\Delta A\|_{\infty}}{\|A\|_{\infty}}) \leq\frac{1}{1 - \text{cond}(A)\|\Delta A\|_{\infty}}$$

$$= 7.59090906 \cdot \left(\frac{0.2}{15.2} + \frac{0.3}{14.0}\right) \cdot \frac{1}{1 - 7.59090906 \cdot \frac{0.3}{14.0}} = 0.31354462$$

The relative error on the solution is about 31% !

The absolute error is simply given by the upper bound:

$$\max\{\|\Delta x\|, \|\Delta y\|\} \leq 0.31354462 \cdot 1.10389611 = 0.34612068$$
3.5.4 Probability distribution of a function of RVs estimated by the Monte-Carlo methods

The probability distribution of the function \( x = f(x_1, \ldots, x_n) \) can be approximated by generating at random the values of the RVs \( x_1, \ldots, x_n \) according to their joint probability distribution, which is assumed to be known.

Collecting and plotting the results provide the approximate probability distribution of the RV \( x \).

The simplest case is that of stochastically independent RVs, with given probability distributions.

An algorithm is needed for the generation of random numbers.

In practice it is enough to have an algorithm for the generation of random numbers with uniform distribution in the interval \([0, 1]\).

Indeed, if \( p(x_i) \) is the probability distribution of the RV \( x_i \in \mathbb{R} \) and \( P(x_i) \) the cumulative one, it can be shown that the RV

\[
Z = P(X) = \int_{-\infty}^{X} p(x_i) \, dx_i \in (0, 1)
\]

follows a uniform distribution in \([0, 1]\).
\section*{Theorem}

Let \( x \) be a RV in \( \mathbb{R} \) with probability distribution \( p(x) \) and let \( g : \mathbb{R} \rightarrow \mathbb{R} \) be a monotonic increasing \( C^1 \) function. The RV

\[ y = g(x) \]

has then the range \((g(-\infty), g(+\infty))\), with probability distribution

\[ \hat{p}(y) = p[g^{-1}(y)] \frac{1}{\frac{dg}{dx}(x) \bigg|_{x=g^{-1}(y)}} \]

where \( g(\pm\infty) = \lim_{x \to \pm\infty} g(x) \) and \( g^{-1} \) denotes the inverse function of \( g \).

\section*{Proof}

We simply have to introduce the change of variables \( x = g^{-1}(y) \) in the normalization integral and apply the theorem of derivation of the inverse functions

\[ 1 = \int_{-\infty}^{+\infty} p(x) \, dx = \int_{g(-\infty)}^{g(+\infty)} p[g^{-1}(y)] \frac{d}{dy} [g^{-1}(y)] \, dy = \]

\[ = \int_{g(-\infty)}^{g(+\infty)} p[g^{-1}(y)] \frac{1}{\frac{dg}{dx}(x) \bigg|_{x=g^{-1}(y)}} \, dy \quad \square \]

\section*{Remark}

If \( y = g(x) = P(x) \) we get

\[ \hat{p}(y) = p[g^{-1}(y)] \frac{1}{p(x) \bigg|_{x=g^{-1}(y)}} = 1 \]

for \( y \in (P(-\infty), P(+\infty)) = (0, 1) \)
It is then possible to proceed in the following way:

(1) Plotting of the cumulative distribution $P(x)$, for $x \in (-\infty, +\infty)$ (or in a sufficiently large interval)

(2) Generation of the RV $z = P(x)$, uniformly distributed in $[0, 1]$, by a **generator of uniform random numbers in $[0, 1]$**

(3) Calculation of the related values of

$$x = P^{-1}(z)$$

by means of the inverse function of $P$

The calculation is usually performed by interpolation of the tabulated values of $P$

The RV $x$ will be distributed according to $p(x)$!

**But how can we generate the random numbers?**
Generators of uniform random numbers in $[0, 1]$

They are always deterministic algorithms, which require a starting input (seed).

The generated values are not really random, but only pseudo-random.

The algorithms ensure that the sequences of values obtained satisfy some criteria of stochasticity.

For instance that, by generating a sufficiently large number of values, the related frequency histogram is essentially constant for a quite small bin width and for any choice of the initial seed.

The algorithms for random numbers can be more or less sophisticated.

A relatively simple class of generators is represented by the multiplication-modulus algorithms.

A standard example is the following:

$$y_{n+1} = 16807y_n \mod 2147483647 \quad \forall n = 0, 1, 2, \ldots$$

If the initial seed $y_0$ is a large odd integer, then the sequence

$$x_n = y_n/2147483647$$

simulates a set of outcomes of a uniform RV in $[0, 1)$.
4. SAMPLE THEORY
SAMPLE ESTIMATES OF $\mu$ AND $\sigma^2$

The results of a repeated measurement affected by random errors (the sample) are assumed to be elements (so-called individuals) belonging to a statistical population

The latter is described by an appropriate probability distribution

The problem is to deduce, from the sample available, the features of such probability distribution

Fundamental problem of the statistical inference

In particular:

Sample estimate of mean and variance of the distribution

In general:

Hypothesis testing on the distribution and the parameters therein
4.1 Sample estimate of the mean $\mu$

If $x_1, x_2, \ldots, x_n$ are the results of the same experimental measurement repeated $n$ times, they can be regarded as the realizations of $n$ independent identically distributed RVs, according to an unknown distribution.

For simplicity’s sake it is convenient to denote with the same symbols $x_1, x_2, \ldots, x_n$ these identically distributed RVs.

The mean $\mu$ of the unknown probability distribution is estimated, by using the available sample, in terms of the arithmetic mean

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

The arithmetic mean must be regarded as a random variable, function of the RVs $x_1, x_2, \ldots, x_n$. 
The estimate is **unbiased**. Indeed:

(●) the mean of $\overline{x}$ is $\mu$, for all $n$

$$E(\overline{x}) = E\left(\frac{1}{n} \sum_{i=1}^{n} x_i\right) = \frac{1}{n} \sum_{i=1}^{n} E(x_i) = \frac{1}{n} \sum_{i=1}^{n} \mu = \mu$$

(●) if $\sigma^2$ is the variance of each RV $x_i$, the RV $\overline{x}$ has variance $\sigma^2/n$ (**de Moivre’s law**)

$$E[(\overline{x} - \mu)^2] = \frac{1}{n^2} \sum_{i,j=1}^{n} E[(x_i - \mu)(x_j - \mu)] = \frac{1}{n^2} \sum_{i,j=1}^{n} \delta_{ij} \sigma^2 = \frac{\sigma^2}{n}$$

(●) more generally, **Khintchine theorem** holds, also known as the **Weak law of large numbers**

*If $(x_1, x_2, \ldots, x_n)$ is a sample of the statistical population whose probability distribution has mean $\mu$, then the arithmetic mean $\overline{x} = \overline{x}_n$ converges in probability to $\mu$ as $n \to \infty$*

$$\forall \varepsilon > 0 \quad \lim_{n \to \infty} p[\mu - \varepsilon \leq \overline{x}_n \leq \mu + \varepsilon] = 1$$

**Proof** For finite $\sigma^2$, the result immediately follows from Tchebychev’s theorem

This means that for larger $n$ the arithmetic mean $\overline{x}$ tends to provide a better estimate of the mean $\mu$ of the probability distribution
As an illustration, the following figure shows that when the number \( n \) of data on which the sample mean is calculated increases, the probability distribution of the sample mean tends to concentrate around the mean value \( \mu \).

As for the shape of the distribution, we can notice what follows:

- if the sample data \( x_1, \ldots, x_n \) follow a normal distribution, the sample mean \( \bar{x} \) is also a normal RV as a linear combination of normal RVs;

- even if the data distribution is not normal, however, the Central Limit Theorem ensures that for \( n \) large enough the distribution of \( \bar{x} \) is normal to a good approximation.
4.2 Sample estimate of the variance $\sigma^2$

Under the same hypotheses, the **variance of the probability distribution** can be estimated by means of the relationship

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

which must be considered a RV, a function of $x_1, x_2, \ldots, x_n$

Also in this case the estimate is unbiased, since:

- (●) the mean $\mathbb{E}(s^2)$ of $s^2$ coincides with $\sigma^2$

$$\frac{1}{n-1} \mathbb{E} \left[ \sum_{i=1}^{n} (x_i - \bar{x})^2 \right] = \frac{1}{n-1} \mathbb{E} \left[ \sum_{i=1}^{n} (x_i - \mu)^2 - n(\bar{x} - \mu)^2 \right]$$

$$= \frac{1}{n-1} \left[ \sum_{i=1}^{n} \mathbb{E}[(x_i - \mu)^2] - n\mathbb{E}[(\bar{x} - \mu)^2] \right] = \sigma^2$$

- (●) if the fourth central moment $\mathbb{E}[(x_i - \mu)^4]$ of the RVs $x_i$ is finite, the variance of $s^2$ becomes

$$\mathbb{E}[(s^2 - \sigma^2)^2] = \frac{1}{n} \mathbb{E}[(x_i - \mu)^4] + \frac{3-n}{n(n-1)} \sigma^4$$

and tends to zero as $n \to +\infty$. As a consequence, $s^2 = s_n^2$ converges in probability to $\sigma^2$

$$\forall \varepsilon > 0 \quad \lim_{n \to \infty} p[\sigma^2 - \varepsilon \leq s_n^2 \leq \sigma^2 + \varepsilon] = 1$$

**Remark** The estimate with $1/n$ instead of $1/(n-1)$ is biased (Bessel’s correction)
Therefore, also for the sample estimate of the variance the probability distribution tends to shrink and concentrate around the variance $\sigma^2$

As for the form of the distribution, we have that:

- if the sample data $x_1, \ldots, x_n$ follow a normal distribution with variance $\sigma^2$, the RV
  $$(n - 1) \frac{s^2}{\sigma^2}$$
  follows a $\chi^2$ distribution with $n - 1$ degrees of freedom (this topic will be discussed later);

- if the data distribution is not normal, for $n$ large enough ($n > 30$) the distribution of $s^2$ can be assumed almost normal, with mean $\sigma^2$ and variance $E[(s^2 - \sigma^2)^2]$. Notice that the result does not trivially follow from the Central Limit Theorem, since the RVs $(x_i - \bar{x})^2$ in $s^2$ are not independent.
4.3 Normal RVs. Confidence intervals for $\mu$ and $\sigma^2$

Whenever the statistical population is normal, we can determine the so-called confidence intervals of the mean $\mu$ and of the variance $\sigma^2$

We simply have to notice that:

(●) the RV, function of $x_1, x_2, \ldots, x_n$,

$$z = \sqrt{n} \frac{\bar{x} - \mu}{\sigma}$$

is normal and standard

**Proof** $\bar{x}$ is normal, with mean $\mu$ and variance $\sigma^2/n$  

(●) the RV

$$X^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

obeys a $X^2$ distribution with $n - 1$ d.o.f.

**Proof**

$X^2$ is a positive semidefinite quadratic form of the independent standard RVs $(x_i - \mu)/\sigma$

$$X^2 = \sum_{i=1}^{n} \left[ \frac{x_i - \mu}{\sigma} \right]^2 - n \left[ \frac{\bar{x} - \mu}{\sigma} \right]^2 = \sum_{ij=1}^{n} \left[ \delta_{ij} - \frac{1}{n} \right] \frac{x_i - \mu}{\sigma} \frac{x_j - \mu}{\sigma}$$

with an idempotent representative matrix

$$\sum_{j=1}^{n} \left[ \delta_{ij} - \frac{1}{n} \right] \left[ \delta_{jk} - \frac{1}{n} \right] = \sum_{j=1}^{n} \left[ \delta_{ij} \delta_{jk} - \frac{1}{n} \delta_{ij} - \frac{1}{n} \delta_{jk} + \frac{1}{n^2} \right] = \delta_{ik} - \frac{1}{n}$$

of rank $n - 1$ (the eigenvalues are 1 and 0, with multiplicities $n - 1$ and 1, respectively)  

Stefano Siboni
(●) the RVs $z$ and $X^2$ are stochastically independent

**Proof**

It is enough to prove that the quadratic forms $z^2$ and $X^2$

$$z^2 = \sum_{ij=1}^{n} \frac{1}{n} \frac{x_i - \mu}{\sigma} x_j - \mu = \sum_{ij=1}^{n} A_{ij} \frac{x_i - \mu}{\sigma} x_j - \mu$$

$$X^2 = \sum_{ij=1}^{n} \left[ \delta_{ij} - \frac{1}{n} \right] \frac{x_i - \mu}{\sigma} x_j - \mu = \sum_{ij=1}^{n} B_{ij} \frac{x_i - \mu}{\sigma} x_j - \mu$$

are independent RVs, by using Craig’s theorem. Indeed

$$(AB)_{ik} = \sum_{j=1}^{n} A_{ij} B_{jk} = \sum_{j=1}^{n} \frac{1}{n} \left[ \delta_{jk} - \frac{1}{n} \right] = \frac{1}{n} \left[ 1 - \frac{1}{n} \right] = 0$$

∀ik = 1, ..., n. Therefore $AB = 0$  

(●) the RV

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

is a Student’s $t$ with $n - 1$ d.o.f.

**Proof**

There holds

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \sqrt{n}\frac{\bar{x} - \mu}{s} = \sqrt{n - 1}\frac{\bar{x} - \mu}{\sigma}\frac{1}{\sqrt{(n - 1)s^2}}$$

so that

$$t = \sqrt{n - 1}\frac{z}{\sqrt{X^2}}$$

with $z$ and $X^2$ stochastically independent, the former standard normal and the latter $X^2$ with $n - 1$ d.o.f.  

Stefano Siboni
4.3.1 Confidence interval for the mean $\mu$

The cumulative distribution of the Student’s $t$ with $n$ d.o.f. is given by:

$$P(\tau) = p[t \leq \tau] = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n} \Gamma\left(\frac{n}{2}\right)} \int_{-\infty}^{\tau} \frac{1}{t^{n/2}} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} dt \quad \tau \in \mathbb{R}$$

This describes the probability that the RV $t$ takes a value $\leq \tau$.

For any $\alpha \in (0, 1)$ we pose $t_{[\alpha]}(n) = P^{-1}(\alpha)$, so that

$$P(t_{[\alpha]}(n)) = \alpha$$

and the symmetry of the Student’s distribution implies

$$t_{[\alpha]}(n) = -t_{[1-\alpha]}(n)$$

For an arbitrary $\alpha$, the inequality

$$t_{[\frac{\alpha}{2}]}(n-1) \leq \frac{\bar{x} - \mu}{s/\sqrt{n}} \leq t_{[1-\frac{\alpha}{2}]}(n-1)$$

is then satisfied with probability $1 - \alpha$ and defines the (two-sided) confidence interval of the mean $\mu$, in the form

$$\bar{x} - t_{[1-\frac{\alpha}{2}]}(n-1) \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} - t_{[\frac{\alpha}{2}]}(n-1) \frac{s}{\sqrt{n}}$$

i.e.

$$\bar{x} - t_{[1-\frac{\alpha}{2}]}(n-1) \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{[1-\frac{\alpha}{2}]}(n-1) \frac{s}{\sqrt{n}}$$

$1 - \alpha$ is known as the confidence level of the interval.
Calculation of $t_{[\alpha]}(n)$ by Maple 11
It can be performed by the following Maple commands:

\[
\text{with(stats):} \\
\text{statevalf[icdf, students[n]](\alpha)};
\]

As an example, for $n = 12$ and $\alpha = 0.7$ we obtain

\[
\text{statevalf[icdf, studentst[12]](0.7)}; \\
0.5386176682
\]

so that

\[
t_{[12]}(0.7) = 0.5386176682
\]

Calculation of $t_{[\alpha]}(n)$ by Excel
The Excel function TINV provides the required value, but the calculation is a bit more involved. Indeed, for any $\varepsilon \in (0, 1)$ the worksheet function

\[
\text{TINV}(\varepsilon; n)
\]

returns directly the critical value $t_{[1-\frac{\varepsilon}{2}]}(n)$. Therefore, it is easily seen that

\[
\text{TINV}(2 - 2\alpha; n)
\]
gives $t_{[n]}(\alpha)$ for any $\alpha \in (0.5, 1)$.

In the same example as before the value of $t_{[12]}(0.7)$ is returned by computing

\[
\text{TINV}(0.6; 12)
\]
4.3.2 Confidence interval for the variance $\sigma^2$

The cumulative distribution of the $\chi^2$ RV with $n$ d.o.f.

$$P(\tau) = p[\chi^2 \leq \tau] = \int_0^\tau \frac{1}{\Gamma(n/2)} \frac{1}{2^{n/2}} e^{-\chi^2/2} (\chi^2)^{n/2-1} d\chi^2$$

cGary the probability that the RV takes a value $\leq \tau$

For any $\alpha \in (0, 1)$ we pose $\chi^2_{[\alpha]}(n) = P^{-1}(\alpha)$, in such a way that

$$P(\chi^2_{[\alpha]}(n)) = \alpha$$

The inequality

$$\chi^2_{[\alpha]}(n-1) \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{[1-\alpha]}(n-1)$$

is then verified with probability $1 - \alpha$ and defines the \textit{two-sided} confidence interval of the variance as

$$\frac{1}{\chi^2_{[1-\alpha]}(n-1)} (n-1)s^2 \leq \sigma^2 \leq \frac{1}{\chi^2_{[\alpha]}(n-1)} (n-1)s^2$$

Alternatively, we can introduce a \textit{one-sided confidence interval}, by means of the inequalities

$$\frac{(n-1)s^2}{\sigma^2} \geq \chi^2_{[\alpha]}(n-1) \iff \sigma^2 \leq (n-1)s^2 \frac{1}{\chi^2_{[\alpha]}(n-1)}$$

In both cases, the \textbf{confidence level} of the interval is given by $1 - \alpha$
Calculation of $\chi^2[\alpha](n)$ by Maple 11

It can be performed by the following Maple commands:

\[
\text{with(stats):} \\
\text{statevalf[icdf, chisquare[n]]}(\alpha);
\]

As an example, for $n = 13$ and $\alpha = 0.3$ we get

\[
\text{statevalf[icdf, chisquare[13]]}(0.3); \\
9.925682415
\]

and therefore

\[
\chi^2_{13}(0.3) = 9.925682415
\]

Calculation of $\chi^2[\alpha](n)$ by Excel

The relevant Excel function is CHIINV. For any $\alpha \in (0, 1)$ the worksheet function

\[
\text{CHIINV(\alpha; n)}
\]

returns the value of $\chi^2_{1-\alpha}(n)$. Consequently, $\chi^2[\alpha](n)$ can be calculated as

\[
\text{CHIINV}(1 - \alpha; n)
\]

In the same example as before the value of $\chi^2_{13}(0.3)$ is returned by

\[
\text{CHIINV}(0.7; 13)
\]
Example: CI for $\mu$ and $\sigma^2$ (normal sample)
Let us consider the following table of data, concerning some repeated measurements of a quantity (in arbitrary units). The sample is assumed to be normal.

<table>
<thead>
<tr>
<th>$i$</th>
<th>$x_i$</th>
<th>$i$</th>
<th>$x_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.0851</td>
<td>13</td>
<td>1.0768</td>
</tr>
<tr>
<td>2</td>
<td>834</td>
<td>14</td>
<td>842</td>
</tr>
<tr>
<td>3</td>
<td>782</td>
<td>15</td>
<td>811</td>
</tr>
<tr>
<td>4</td>
<td>818</td>
<td>16</td>
<td>829</td>
</tr>
<tr>
<td>5</td>
<td>810</td>
<td>17</td>
<td>803</td>
</tr>
<tr>
<td>6</td>
<td>837</td>
<td>18</td>
<td>811</td>
</tr>
<tr>
<td>7</td>
<td>857</td>
<td>19</td>
<td>789</td>
</tr>
<tr>
<td>8</td>
<td>768</td>
<td>20</td>
<td>831</td>
</tr>
<tr>
<td>9</td>
<td>842</td>
<td>21</td>
<td>829</td>
</tr>
<tr>
<td>10</td>
<td>786</td>
<td>22</td>
<td>825</td>
</tr>
<tr>
<td>11</td>
<td>812</td>
<td>23</td>
<td>796</td>
</tr>
<tr>
<td>12</td>
<td>784</td>
<td>24</td>
<td>841</td>
</tr>
</tbody>
</table>

We want to determine the confidence interval (CI) of the mean and that of the variance, both with confidence level $1 - \alpha = 0.95$

**Solution**

Number of data: $n = 24$

Sample mean: $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = 1.08148$

Sample variance: $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 = 6.6745 \cdot 10^{-6}$

Estimated standard deviation: $s = \sqrt{s^2} = 0.002584$
The CI of the mean is
\[
\bar{x} - t_{\left[1 - \frac{\alpha}{2}\right]}(n-1) \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\left[1 - \frac{\alpha}{2}\right]}(n-1) \frac{s}{\sqrt{n}}
\]
and for \(\alpha = 0.05\), \(n = 24\) has therefore the limits
\[
\bar{x} - t_{[0.975](23)} \frac{s}{\sqrt{24}} = 1.08148 - 2.069 \cdot \frac{0.002584}{\sqrt{24}} = 1.08039
\]
\[
\bar{x} + t_{[0.975](23)} \frac{s}{\sqrt{24}} = 1.08148 + 2.069 \cdot \frac{0.002584}{\sqrt{24}} = 1.08257
\]
so that the CI writes
\[
1.08039 \leq \mu \leq 1.08257
\]
or, equivalently,
\[
[1.08148 \pm 0.00109]
\]

The CI of the variance takes the form
\[
\frac{1}{\chi^2_{\left[1 - \frac{\alpha}{2}\right]}(n-1)} (n-1)s^2 \leq \sigma^2 \leq \frac{1}{\chi^2_{\left[\frac{\alpha}{2}\right]}(n-1)} (n-1)s^2
\]
with \(\alpha = 0.05\) and \(n = 24\). Thus
\[
\frac{1}{\chi^2_{[0.975](23)}} 23s^2 = \frac{1}{38.076} 23 \cdot 6.6745 \cdot 10^{-6} = 4.03177 \cdot 10^{-6}
\]
\[
\frac{1}{\chi^2_{[0.025](23)}} 23s^2 = \frac{1}{11.689} 23 \cdot 6.6745 \cdot 10^{-6} = 13.1332 \cdot 10^{-6}
\]
and the CI becomes
\[
4.03177 \cdot 10^{-6} \leq \sigma^2 \leq 13.1332 \cdot 10^{-6}
\]
4.4 Large samples of an arbitrary population.
Approximate confidence intervals of $\mu$ and $\sigma^2$

In this case the definition of the confidence intervals is possible because for large samples the distribution of $\bar{x}$ and $s^2$ is approximately normal for any population.

4.4.1 Confidence interval for the mean $\mu$

For any population of finite mean $\mu$ and variance $\sigma^2$, the Central Limit Theorem ensures that when $n$ is large enough the RV

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma} = \sqrt{n} \frac{\bar{x} - \mu}{\sigma} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

follows a normal standard distribution (and the larger $n$ the better the approximation): the RVs $x_1, \ldots, x_n$ are indeed independent and identically distributed.

The standard deviation $\sigma$ is unknown, but $s^2$ constitutes a RV of mean $\sigma^2$ and its standard deviation tends to zero as $1/\sqrt{n}$: for large samples it becomes very likely that $s^2 \simeq \sigma^2$. We may then assume the RV

$$z_n = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

approximately standard normal for large samples
(large sample: $n > 30$)

Such an approximation is the key to determine the CI of the mean $\mu$. 

Stefano Siboni

62
For a fixed $z_{1-\frac{\alpha}{2}} > 0$, the probability that the RV $z_n$ takes any value within the interval (symmetric with respect to $z_n = 0$)

$$-z_{1-\frac{\alpha}{2}} \leq z_n \leq z_{1-\frac{\alpha}{2}}$$

is approximately given by the integral between $-z_{1-\frac{\alpha}{2}}$ and $z_{1-\frac{\alpha}{2}}$ of the standard normal distribution

$$\int_{-z_{1-\frac{\alpha}{2}}}^{z_{1-\frac{\alpha}{2}}} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$

and is denoted with $1 - \alpha$. As a matter of fact, we do the contrary: we assign the confidence level $1 - \alpha$ and determine the corresponding value of $z_{1-\frac{\alpha}{2}}$. We simply have to notice that, due to the symmetry of the distribution, the tails $\{z < -z_{1-\frac{\alpha}{2}}\}$ and $\{z > z_{1-\frac{\alpha}{2}}\}$ must have the same probability $\alpha/2$

the critical value $z_{1-\frac{\alpha}{2}} > 0$ being thus defined by the equation

$$\frac{1}{\sqrt{2\pi}} \int_{0}^{z_{1-\frac{\alpha}{2}}} e^{-\xi^2/2} d\xi = \frac{1 - \alpha}{2}$$

Stefano Siboni
The integral between 0 and $z_{[1-\alpha/2]}$ of the standard normal distribution (or vice versa, the value of $z_{[1-\alpha/2]}$ for a given value $(1-\alpha)/2$ of the integral) can be calculated by using a table of the kind partially reproduced here

<table>
<thead>
<tr>
<th>$X$</th>
<th>0.00</th>
<th>0.01</th>
<th>0.02</th>
<th>0.03</th>
<th>0.04</th>
<th>0.05</th>
<th>0.06</th>
<th>0.07</th>
<th>0.08</th>
<th>0.09</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.00000</td>
<td>0.00399</td>
<td>0.00798</td>
<td>0.01197</td>
<td>0.01595</td>
<td>0.01994</td>
<td>0.02392</td>
<td>0.02790</td>
<td>0.03188</td>
<td>0.03586</td>
</tr>
<tr>
<td>0.1</td>
<td>0.03983</td>
<td>0.04380</td>
<td>0.04776</td>
<td>0.05172</td>
<td>0.05567</td>
<td>0.05962</td>
<td>0.06356</td>
<td>0.06749</td>
<td>0.07142</td>
<td>0.07535</td>
</tr>
<tr>
<td>0.2</td>
<td>0.07926</td>
<td>0.08317</td>
<td>0.08706</td>
<td>0.09095</td>
<td>0.09483</td>
<td>0.09871</td>
<td>0.10257</td>
<td>0.10642</td>
<td>0.11026</td>
<td>0.11409</td>
</tr>
<tr>
<td>0.3</td>
<td>0.11791</td>
<td>0.12172</td>
<td>0.12552</td>
<td>0.12930</td>
<td>0.13307</td>
<td>0.13683</td>
<td>0.14058</td>
<td>0.14431</td>
<td>0.14803</td>
<td>0.15173</td>
</tr>
<tr>
<td>0.4</td>
<td>0.15542</td>
<td>0.15910</td>
<td>0.16276</td>
<td>0.16640</td>
<td>0.17003</td>
<td>0.17364</td>
<td>0.17724</td>
<td>0.18082</td>
<td>0.18439</td>
<td>0.18793</td>
</tr>
<tr>
<td>0.5</td>
<td>0.19146</td>
<td>0.19497</td>
<td>0.19847</td>
<td>0.20194</td>
<td>0.20540</td>
<td>0.20884</td>
<td>0.21226</td>
<td>0.21566</td>
<td>0.21904</td>
<td>0.22240</td>
</tr>
<tr>
<td>0.6</td>
<td>0.22575</td>
<td>0.22907</td>
<td>0.23237</td>
<td>0.23565</td>
<td>0.23891</td>
<td>0.24215</td>
<td>0.24537</td>
<td>0.24857</td>
<td>0.25175</td>
<td>0.25490</td>
</tr>
<tr>
<td>0.7</td>
<td>0.25804</td>
<td>0.26115</td>
<td>0.26424</td>
<td>0.26730</td>
<td>0.27035</td>
<td>0.27337</td>
<td>0.27637</td>
<td>0.27935</td>
<td>0.28230</td>
<td>0.28524</td>
</tr>
<tr>
<td>0.8</td>
<td>0.28814</td>
<td>0.29103</td>
<td>0.29389</td>
<td>0.29673</td>
<td>0.29955</td>
<td>0.30234</td>
<td>0.30515</td>
<td>0.30785</td>
<td>0.31057</td>
<td>0.31327</td>
</tr>
<tr>
<td>0.9</td>
<td>0.31594</td>
<td>0.31859</td>
<td>0.32121</td>
<td>0.32381</td>
<td>0.32639</td>
<td>0.32894</td>
<td>0.33147</td>
<td>0.33398</td>
<td>0.33646</td>
<td>0.33891</td>
</tr>
<tr>
<td>1.0</td>
<td>0.34134</td>
<td>0.34375</td>
<td>0.34614</td>
<td>0.34849</td>
<td>0.35083</td>
<td>0.35314</td>
<td>0.35543</td>
<td>0.35769</td>
<td>0.35993</td>
<td>0.36214</td>
</tr>
<tr>
<td>1.1</td>
<td>0.36433</td>
<td>0.36650</td>
<td>0.36864</td>
<td>0.37076</td>
<td>0.37286</td>
<td>0.37493</td>
<td>0.37698</td>
<td>0.37900</td>
<td>0.38100</td>
<td>0.38298</td>
</tr>
<tr>
<td>1.2</td>
<td>0.38493</td>
<td>0.38686</td>
<td>0.38877</td>
<td>0.39065</td>
<td>0.39251</td>
<td>0.39435</td>
<td>0.39617</td>
<td>0.39796</td>
<td>0.39973</td>
<td>0.40147</td>
</tr>
</tbody>
</table>

(which provides the values of the integral for $z_{[1-\alpha/2]}$ between 0.00 and 1.29, sampled at intervals of 0.01)

Alternatively, we may express the integral in the form

$$
\frac{1}{\sqrt{2\pi}} \int_0^{z_{[1-\alpha/2]}} e^{-\xi^2/2} d\xi = \frac{1}{2} \text{erf} \left( \frac{z_{[1-\alpha/2]}}{\sqrt{2}} \right)
$$

in terms of the error function:

$$
\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt
$$

whose values are tabulated in an analogous way
To obtain the CI of the mean $\mu$ we simply have to remember that the double inequality

$$-z_{[1-\frac{\alpha}{2}]} \leq \frac{\bar{x} - \mu}{s/\sqrt{n}} \leq z_{[1-\frac{\alpha}{2}]}$$

is satisfied with probability $1 - \alpha$, so that

$$\bar{x} - z_{[1-\frac{\alpha}{2}]} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{[1-\frac{\alpha}{2}]} \frac{s}{\sqrt{n}}$$

holds with the same probability $1 - \alpha$. What we have obtained is the CI of the mean with a **confidence level** $1 - \alpha$.

**Calculation of $z_{[1-\frac{\alpha}{2}]$ by Maple 11**

By definition $z_{[1-\frac{\alpha}{2}]}$ is the inverse of the standard normal cumulative distribution at $1 - \frac{\alpha}{2}$, which can be calculated by the Maple command lines:

```
with(stats):
statevalf[icdf, normald[0, 1]](1 - \frac{\alpha}{2});
```

As an example, for $\alpha = 0.4$ we have $1 - \frac{\alpha}{2} = 0.8$ and the value of $z_{[1-\frac{\alpha}{2}]}$ is computed by means of the command

```
statevalf[icdf, normald[0, 1]](0.8);
```

0.8416212336

so that

$$z_{0.8} = 0.8416212336$$
**Calculation of** $z_{[1-\frac{\alpha}{2}]}$ **by Excel**

The Excel function to be used in this case is NORMINV. For any $\alpha \in (0, 1)$ the worksheet command

$$\text{NORMINV}(\alpha; 0; 1)$$

returns the value at $\alpha$ of the inverse of the standard normal cumulative distribution. Therefore, the value of $z_{[1-\frac{\alpha}{2}]}$ is obtained by

$$\text{NORMINV}(1 - \frac{\alpha}{2}; 0; 1)$$

In the previous example, the command would take the form

$$\text{NORMINV}(0, 8; 0; 1)$$

### 4.4.2 Confidence interval for the variance $\sigma^2$

For large samples ($n > 30$) the sample variance $s^2$ can be regarded as a normal RV of mean $\sigma^2$ and variance

$$\mathbb{E}[(s^2 - \sigma^2)^2] = \frac{1}{n} \mathbb{E}[(x_i - \mu)^4] + \frac{3 - n}{n(n - 1)} \sigma^4.$$  

The variance is estimated, as usual, by the sample variance

$$\sigma^2 \simeq s^2$$

while the fourth central moment can be calculated by using the corresponding sample estimate

$$\mathbb{E}[(x_i - \mu)^4] \simeq m_4 = \frac{1}{n - 1} \sum_{i=1}^{n} (x_i - \bar{x})^4$$

or, for known special distributions, possible relationships which relate the fourth moment to the variance.
We can then claim that for large samples the RV

\[ z_n = \frac{s^2 - \sigma^2}{\sqrt{\frac{1}{n} m_4 + \frac{3 - n}{n(n-1)} s^4}} \]

is approximately standard normal

We can then determine the approximate CI for the variance \( \sigma^2 \) in a way analogous to that discussed for the mean \( \mu \).

The CI for the variance turns out to be:

\[
s^2 - z_{[1-\frac{\alpha}{2}]} \sqrt{\frac{1}{n} m_4 + \frac{3 - n}{n(n-1)} s^4} \leq \sigma^2 \leq s^2 + z_{[1-\frac{\alpha}{2}]} \sqrt{\frac{1}{n} m_4 + \frac{3 - n}{n(n-1)} s^4}
\]

and that of the standard deviation \( \sigma \) is therefore:

\[
\sqrt{s^2 - z_{[1-\frac{\alpha}{2}]} \sqrt{\frac{1}{n} m_4 + \frac{3 - n}{n(n-1)} s^4}} \leq \sigma \leq \sqrt{s^2 + z_{[1-\frac{\alpha}{2}]} \sqrt{\frac{1}{n} m_4 + \frac{3 - n}{n(n-1)} s^4}}
\]

both with confidence level \( 1 - \alpha \).
**Example: CI for the mean (non-normal sample)**

The measurements of the diameters of a random sample of 200 balls for ball bearing, produced by a machine in a week, show a mean of 0.824 cm and a standard deviation of 0.042 cm. Determine, for the mean diameter of the balls:

(a) the 95%-confidence interval;
(b) the 99%-confidence interval.

**Solution**

As \( n = 200 > 30 \), we can apply the theory of large samples and it is not necessary to assume that the population is normal.

(a) The confidence level is \( 1 - \alpha = 0.95 \), so that \( \alpha = 0.05 \) and

\[
\frac{\alpha}{2} = \frac{0.05}{2} = 0.025 \implies 1 - \frac{\alpha}{2} = 1 - 0.025 = 0.975
\]

On the table of the standard normal distribution we read then

\[
z_{[1 - \frac{\alpha}{2}]} = z_{[0.975]} = 1.96
\]

and the confidence interval becomes

\[
\bar{x} - z_{[0.975]} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{[0.975]} \frac{s}{\sqrt{n}}
\]

i.e.

\[
0.824 - 1.96 \cdot \frac{0.042}{\sqrt{200}} \leq \mu \leq 0.824 + 1.96 \cdot \frac{0.042}{\sqrt{200}}
\]

and finally

\[
0.81812 \text{ cm} \leq \mu \leq 0.8298 \text{ cm}.
\]
The same CI can be expressed in the equivalent form:

\[ \mu = 0.824 \pm 0.0058 \text{ cm}. \]

(b) In the present case the confidence level holds \(1 - \alpha = 0.99\), in such a way that \(\alpha/2 = 0.005\) and

\[ 1 - \frac{\alpha}{2} = 1 - 0.005 = 0.995 \]

The table of the standard normal distribution provides then

\[ z_{[1-\frac{\alpha}{2}]} = z_{[0.995]} = 2.58 \]

and the CI of the mean becomes

\[
\bar{x} - z_{[0.995]} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{[0.995]} \frac{s}{\sqrt{n}}
\]

that is

\[
0.824 - 2.58 \cdot \frac{0.042}{\sqrt{200}} \leq \mu \leq 0.824 + 2.58 \cdot \frac{0.042}{\sqrt{200}}
\]

and finally

\[
0.8163 \text{ cm} \leq \mu \leq 0.8317 \text{ cm}.
\]

The alternative form below puts into evidence the absolute error:

\[ \mu = 0.824 \pm 0.0077 \text{ cm}. \]
5. HYPOTHESIS TESTING

The study of the characteristics of a probability distribution is performed by the formulation of appropriate hypotheses (kind of distribution, value of the parameters, etc.)

The correctness of the hypothesis is not certain, but it must be tested.

The hypothesis whose correctness is to be tested is known as null hypothesis, usually denoted with $H_0$.

The alternative hypothesis is denoted with $H_1$.

The acceptance or the rejection of the null hypothesis may give rise to:

- a type I error, when the hypothesis is rejected although actually correct;
- a type II error, if the null hypothesis is accepted although incorrect.

The test must lead to the acceptance or rejection of the null hypothesis, quantifying the related probability error (significance level of the test).

Many tests are based on the rejection of the null hypothesis, by specifying the probability of a type I error.
Tests based on the rejection of the null hypothesis

The basis of the test is a suitable RV (test variable), a function of the individuals of the small sample. The choice of the test variable depends on the kind of null hypothesis $H_0$ to be tested.

The set of the possible values of the test variable is divided into two regions: a rejection region and an acceptance region. As a rule, these regions are (possibly unions of) real intervals.

If the value of the test variable for the given sample falls in the rejection region, the null hypothesis is rejected.

The test variable must be chosen in such a way that it follows a known probability distribution whenever the null hypothesis is satisfied.

It is then possible to calculate the probability that, although the null hypothesis is correct, the test variable takes a value within the rejection region, leading thus to a type I error.

Tests based on the rejection of the null hypothesis allow us to quantify the probability of a type I error, i.e. of rejecting a correct hypothesis as false.
An illustrative example

(●) Process to be analysed: repeated toss of a die (independent tosses)

$H_0$: the outcome 1 has a probability $\pi = 1/6$ (fair die)

$H_1$: the outcome 1 has a probability $\pi = 1/4$ (fixed die)

(●) Probability of $x$ outcomes of the side 1 out of $n$ tosses

$$p_{n,\pi}(x) = \frac{n!}{x!(n-x)!}\pi^x(1-\pi)^{n-x} \quad (x = 0, 1, \ldots, n)$$

with mean $n\pi$ and variance $n\pi(1-\pi)$

(●) Frequency of the side 1 out of $n$ tosses

$$f = \frac{x}{n} \quad (f = 0, \frac{1}{n}, \frac{2}{n}, \ldots, 1)$$

(●) Probability of a frequency $f$ observed out of $n$ tosses

$$p_{n,\pi}(f) = \frac{n!}{(nf)!(n(1-f))}\pi^{nf}(1-\pi)^{n(1-f)}$$

with mean $\frac{1}{n}n\pi = \pi$ and variance $\frac{1}{n^2}n\pi(1-\pi) = \frac{\pi(1-\pi)}{n}$
(•) Probability distribution of \( f \):

\[
\begin{align*}
H_0 \text{ true} & \implies \pi = 1/6 \\
H_1 \text{ true} & \implies \pi = 1/4
\end{align*}
\]

\[\downarrow\]

\( p_{n,1/6}(f) \): probability distribution of \( f \) if \( H_0 \) holds true

\( p_{n,1/4}(f) \): probability distribution of \( f \) if \( H_1 \) is true

Trend of the two distributions for \( n = 20 \)
(●) Testing strategy:

\( f \) is the **test variable**

if \( f \) is small enough **we accept** \( H_0 \) (and reject \( H_1 \))

if \( f \) is sufficiently large **we reject** \( H_0 \) (and accept \( H_1 \))

We define a critical value \( f_{cr} \) of \( f \) such that

\[
\begin{align*}
  f &\leq f_{cr} \implies \text{acceptance of } H_0 \\
  f &> f_{cr} \implies \text{rejection of } H_0
\end{align*}
\]

\( \{f \in [0, 1] : f \leq f_{cr}\} \) denotes the **region of acceptance**

\( \{f \in [0, 1] : f > f_{cr}\} \) constitutes the **region of rejection**
Interpretation of the shadowed areas

\( \alpha \): probability of rejecting \( H_0 \), although it holds true
   — probability of a **type I error**

\( \beta \): probability of accepting \( H_0 \), although actually false
   — probability of a **type II error**

\( \alpha \): **“significance level”** of the test

\( 1 - \beta \): **“power”** of the test

**Remark** It is convenient that \( \alpha \) and \( \beta \) are both small

But it is impossible to reduce both \( \alpha \) and \( \beta \) by modifying the only value of \( f_{cr} \). Indeed:

- if \( f_{cr} \) increases \( \implies \) \( \alpha \) decreases, but \( \beta \) grows;
- if \( f_{cr} \) decreases \( \implies \) \( \alpha \) grows and \( \beta \) decreases;
In order to reduce both $\alpha$ and $\beta$ it is necessary to consider a larger sample (i.e., to increase $n$)

When $n$ increases, the variance $\pi(1-\pi)/n$ of $f$ decreases and the distribution $p_{n,\pi}(f)$ tends to concentrate around the mean value $\pi$

In this example the occurrence of the alternative hypothesis $H_1 : \pi = 1/4$ characterizes completely the probability distribution of $f$

A hypothesis which, when true, specifies completely the probability distribution of the test variable is known as simple; in the opposite case it is named composed hypothesis

The calculation of $\alpha$ and $\beta$ is then possible only if both the hypotheses $H_0$ and $H_1$ are simple
Summary of the main hypothesis tests

\[
\begin{align*}
\chi^2 & \text{-test for adapting a distribution to a sample} \\
\text{Kolmogorov-Smirnov test for adapting a distribution to a sample}
\end{align*}
\]

PARAMETRIC TESTS

\[
\begin{align*}
t \text{-test on the mean of a normal population} \\
\chi^2 & \text{-test on the variance of a normal population} \\
\end{align*}
\]

\[
\begin{align*}
z \text{-test to compare the means of 2 independent normal populations} & \quad \text{(of known variances)} \\
F & \text{-test to compare the variances of 2 independent normal populations} \\
\text{Unpaired } t \text{-test to compare the means of 2 independent normal populations} & \quad \text{(of unknown variances)} \\
\text{- test for the case of equal variances} \\
\text{- test for the case of unequal variances} \\
\text{Paired } t \text{-test to compare the means of 2 normal populations}
\end{align*}
\]

\[
\begin{align*}
F \text{-tests for ANOVA}
\end{align*}
\]

\[
\begin{align*}
\text{Chauvenet criterion}
\end{align*}
\]

NON-PARAMETRIC TESTS

\[
\begin{align*}
\text{Sign test for the median} \\
\text{Sign test to check if 2 paired samples belong to the same population}
\end{align*}
\]
5.1 $\chi^2$-test for adapting a distribution to a sample

It is used to establish whether a statistical population follows a given probability distribution $p(x)$.

The individuals of the sample are plotted as a histogram on a suitable number of bins of appropriate width (“binned distribution”). Let $f_i$ be the number of individuals in the $i$-th bin (“empirical frequency”).

The average number $n_i$ of individuals expected in the $i$-th bin is calculated by means of the distribution $p(x)$ (“theoretical frequency”).

If the assumed probability distribution $p(x)$ is correct, the RV

$$\chi^2 = \sum_i \frac{(f_i - n_i)^2}{n_i}$$

wherein the sum is over all the bins of the histogram, follows approximately a $\chi^2$ distribution with $k - 1 - c$ d.o.f., being:

- $k$ the number of bins;
- $c$ the number of parameters of the distribution which are not a priori assigned, but calculated by using the same sample (the so-called “constraints”)

The approximation is satisfactory whenever the observed frequencies $f_i$ are not too small ($\geq 3$ about).
It seems reasonable to accept the null hypothesis

\[ H_0 : p(x) \text{ is the probability distribution of the population} \]

whenever \( \chi^2 \) is not too large, and to reject it in the opposite case

Therefore the rejection region is defined by

\[ R = \{ \chi^2 \geq \chi^2_{[1-\alpha](k-1-c)} \} \]

If the hypothesis \( H_0 \) is true, the probability that \( \chi^2 \) takes anyway a value in the rejection interval \( R \) holds

\[ 1 - (1 - \alpha) = \alpha \]

which represents therefore the probability of a type I error, and also the significance level of the test
Example: $\chi^2$-test for a uniform distribution
Repeated measurements of a certain quantity $x$ led to the following data table

<table>
<thead>
<tr>
<th>$x_i$</th>
<th>$x_i$</th>
<th>$x_i$</th>
<th>$x_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.101</td>
<td>0.020</td>
<td>0.053</td>
<td>0.120</td>
</tr>
<tr>
<td>0.125</td>
<td>0.075</td>
<td>0.033</td>
<td>0.130</td>
</tr>
<tr>
<td>0.210</td>
<td>0.180</td>
<td>0.151</td>
<td>0.145</td>
</tr>
<tr>
<td>0.245</td>
<td>0.172</td>
<td>0.199</td>
<td>0.252</td>
</tr>
<tr>
<td>0.370</td>
<td>0.310</td>
<td>0.290</td>
<td>0.349</td>
</tr>
<tr>
<td>0.268</td>
<td>0.333</td>
<td>0.265</td>
<td>0.287</td>
</tr>
<tr>
<td>0.377</td>
<td>0.410</td>
<td>0.467</td>
<td>0.398</td>
</tr>
<tr>
<td>0.455</td>
<td>0.390</td>
<td>0.444</td>
<td>0.497</td>
</tr>
<tr>
<td>0.620</td>
<td>0.505</td>
<td>0.602</td>
<td>0.544</td>
</tr>
<tr>
<td>0.526</td>
<td>0.598</td>
<td>0.577</td>
<td>0.556</td>
</tr>
<tr>
<td>0.610</td>
<td>0.568</td>
<td>0.630</td>
<td>0.749</td>
</tr>
<tr>
<td>0.702</td>
<td>0.657</td>
<td>0.698</td>
<td>0.731</td>
</tr>
<tr>
<td>0.642</td>
<td>0.681</td>
<td>0.869</td>
<td>0.761</td>
</tr>
<tr>
<td>0.802</td>
<td>0.822</td>
<td>0.778</td>
<td>0.998</td>
</tr>
<tr>
<td>0.881</td>
<td>0.950</td>
<td>0.922</td>
<td>0.975</td>
</tr>
<tr>
<td>0.899</td>
<td>0.968</td>
<td>0.945</td>
<td>0.966</td>
</tr>
</tbody>
</table>

The whole number of data is 64. We want to check, with a 5% significance level, whether the population from which the sample comes may follow a uniform distribution in $[0, 1]$.

Solution
The 64 data are arranged in increasing order and plotted in a histogram with 8 bins of equal amplitude in the interval $[0, 1]$, closed on the right and open on the left (apart from the first, which is closed at both the endpoints).
We obtain then the following histogram of frequency

which seems satisfactory for the application of a $\chi^2$-test, since the empirical frequencies in each bin are never smaller than $3 \div 5$.

If the uniform distribution in $[0, 1]$ holds true, the expected frequency is the same for each bin:

$$\text{number of sample data} \times \text{probability of the bin} =$$

$$= 64 \times \frac{1}{8} = 8$$

The $\chi^2$ of the sample for the suggested distribution holds then

$$\chi^2 = \frac{(7-8)^2}{8} + \frac{(8-8)^2}{8} + \frac{(9-8)^2}{8} + \frac{(8-8)^2}{8} +$$

$$+ \frac{(10-8)^2}{8} + \frac{(8-8)^2}{8} + \frac{(5-8)^2}{8} + \frac{(9-8)^2}{8} = 2$$
The number of d.o.f. is simply the number of bins minus 1

\[ k - 1 - c = 8 - 1 - 0 = 7 \]

as no parameter of the distribution must be estimated by using the same data of the sample \( (c = 0) \).

With a significance level \( \alpha = 0.05 \), the hypothesis that the uniform be correct is rejected if

\[ \chi^2 \geq \chi^2_{[1-\alpha](k-1-c)} = \chi^2_{[1-0.05](7)} = \chi^2_{[0.95](7)} \]

In the present case, the table of the cumulative \( \chi^2 \) distribution gives

\[ \chi^2_{[0.95](7)} = 14.067 \]

so that the sample \( \chi^2 \) is smaller than the critical one

\[ \chi^2 = 2 < 14.067 = \chi^2_{[0.95](7)} \]

Therefore we conclude that, with a 5% significance level, we cannot exclude that the population follows a uniform distribution in \([0, 1]\). The null hypothesis must be accepted, or however not rejected.
Example: \( \chi^2 \)-test for a normal distribution
An anthropologist is interested in the heights of the natives of a certain island. He suspects that the heights of male adults should be normally distributed and measures the heights of a sample of 200 men. By using these data, he computes the sample mean \( \bar{x} \) and standard deviation \( s \), and applies the results as an estimate for the mean \( \mu \) and the standard deviation \( \sigma \) of the expected normal distribution. Then he chooses eight equal intervals where he groups the results of his records (empirical frequencies). He deduces the table below:

<table>
<thead>
<tr>
<th>i</th>
<th>height interval</th>
<th>empirical frequency</th>
<th>expected frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( x &lt; \bar{x} - 1.5s )</td>
<td>14</td>
<td>13.362</td>
</tr>
<tr>
<td>2</td>
<td>( \bar{x} - 1.5s \leq x &lt; \bar{x} - s )</td>
<td>29</td>
<td>18.370</td>
</tr>
<tr>
<td>3</td>
<td>( \bar{x} - s \leq x &lt; \bar{x} - 0.5s )</td>
<td>30</td>
<td>29.976</td>
</tr>
<tr>
<td>4</td>
<td>( \bar{x} - 0.5s \leq x &lt; \bar{x} )</td>
<td>27</td>
<td>38.292</td>
</tr>
<tr>
<td>5</td>
<td>( \bar{x} \leq x &lt; \bar{x} + 0.5s )</td>
<td>28</td>
<td>38.292</td>
</tr>
<tr>
<td>6</td>
<td>( \bar{x} + 0.5s \leq x &lt; \bar{x} + s )</td>
<td>31</td>
<td>29.976</td>
</tr>
<tr>
<td>7</td>
<td>( \bar{x} + s \leq x &lt; \bar{x} + 1.5s )</td>
<td>28</td>
<td>18.370</td>
</tr>
<tr>
<td>8</td>
<td>( \bar{x} + 1.5s \leq x )</td>
<td>13</td>
<td>13.362</td>
</tr>
</tbody>
</table>

Check the hypothesis that the distribution is actually normal with a significance level (a) of 5% and (b) of 1%.

Solution
The empirical frequencies \( f_i \) are rather large (\( f_i \gg 5 \)), so that the \( \chi^2 \)-test is applicable. The expected frequencies \( n_i \) are calculated multiplying by the whole number of individuals — 200 — the integrals of the standard normal distribution over each interval. For intervals \( i = 5, 6, 7, 8 \) the table of integrals be-
between 0 and \( z \) of the standard normal provides indeed:

\[
P(\bar{x} \leq x < \bar{x} + 0.5s) = \int_{0}^{0.5} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}dz = 0.19146
\]

\[
P(\bar{x} + 0.5s \leq x < \bar{x} + s) = \int_{0.5}^{1} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}dz = \int_{0}^{0.5} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}dz = 0.19146
\]

\[
P(\bar{x} + s \leq x < \bar{x} + 1.5s) = \int_{1}^{1.5} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}dz = \int_{0}^{1.5} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}dz - \int_{0}^{0.5} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}dz = 0.34134 - 0.19146 = 0.14674
\]

\[
P(\bar{x} + 1.5s \leq x) = \int_{1.5}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}dz = \int_{0}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}dz - \int_{0}^{1.5} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}dz = 0.5 - 0.34319 = 0.15681
\]
while the probabilities of the intervals \( i = 1, 2, 3, 4 \) are symmetrically equal to the previous ones (due to the symmetry of the standard normal distribution with respect to the origin).

The \( \chi^2 \) of the sample is then given by:

\[
\chi^2 = \sum_{i=1}^{8} \frac{(f_i - n_i)^2}{n_i} = 17.370884
\]

If the proposed probability distribution is correct, the test variable obeys approximately a \( \chi^2 \) distribution with

\[8 - 1 - 2 = 5\]

d.o.f. since \( k = 8 \) are the bins introduced and \( c = 2 \) the parameters of the distribution (\( \mu \) and \( \sigma \)) which are estimated by using the same data of the sample.

The table of the cumulative distributions of \( \chi^2 \) provides the critical values:

\[
\chi^2_{[1-\alpha](5)} = \chi^2_{[0.95](5)} = 11.070 \quad \text{for} \quad \alpha = 0.05
\]

\[
\chi^2_{[1-\alpha](5)} = \chi^2_{[0.99](5)} = 15.086 \quad \text{for} \quad \alpha = 0.01
\]

In both cases the \( \chi^2 \) calculated on the sample is larger: we conclude that, with both levels of significance of 5 and 1%, the null hypothesis must be rejected. The data of the sample suggest that the height distribution is not normal for the population of the natives.
5.2 Kolmogorov-Smirnov test

It is an alternative method to test if a set of independent data follows a given probability distribution.

For a continuous random variable \( x \in \mathbb{R} \) with distribution \( p(x) \), we consider \( n \) outcomes \( x_1, x_2, \ldots, x_n \).

Equivalently, \( x_1, x_2, \ldots, x_n \) are assumed to be continuous i.i.d. random variables of distribution \( p(x) \).

The empirical cumulative frequency distribution at \( n \) data of the variable \( x \) is defined as

\[
P_{e,n}(x) = \frac{\#\{x_i, \ i = 1, \ldots, n, \ x_i \leq x\}}{n}
\]

and is compared with the cumulative distribution of \( x \)

\[
P(x) = \int_{-\infty}^{x} p(\xi) \, d\xi
\]
Test variable

We consider the maximum residual between the theoretical and empirical cumulative distribution

\[ D_n = \sup_{x \in \mathbb{R}} |P(x) - P_{e,n}(x)| \]

\(D_n\) is a function of the random variables \(x_1, x_2, \ldots, x_n\) and constitutes in turn a random variable — known as Kolmogorov-Smirnov variable

Kolmogorov-Smirnov test is based on the following limit theorem

**Theorem**

If the random variables \(x_1, x_2, \ldots\) are i.i.d. with continuous cumulative distribution \(P(x)\), for any given \(\lambda > 0\) there holds

\[
\text{Probability}\left\{ D_n > \frac{\lambda}{\sqrt{n}} \right\} \xrightarrow{n \to +\infty} Q(\lambda)
\]

where \(Q(\lambda)\) denotes the decreasing function defined in \(\mathbb{R}^+\) by

\[
Q(\lambda) = 2 \sum_{j=1}^{\infty} (-1)^{j-1} e^{-2j^2 \lambda^2} \quad \forall \lambda > 0
\]

with

\[
\lim_{\lambda \to 0^+} Q(\lambda) = 1 \quad \lim_{\lambda \to +\infty} Q(\lambda) = 0 \quad \square
\]
The theorem ensures that for $n$ large enough there holds

$$\text{Probability}\left\{ D_n > \frac{\lambda}{\sqrt{n}} \right\} \simeq Q(\lambda)$$

whatever the distribution $P(x)$ is.

For a fixed value $\alpha \in (0, 1)$ of the previous probability, we have

$$\alpha = Q(\lambda) \iff \lambda = Q^{-1}(\alpha)$$

in such a way that

$$\text{Probability}\left\{ D_n > \frac{Q^{-1}(\alpha)}{\sqrt{n}} \right\} \simeq \alpha$$

**Hypothesis to be tested**

$H_0$: the cumulative distribution of data $x_1, x_2, \ldots, x_n$ is $P(x)$

$H_1$: the cumulative distribution is not $P(x)$

**Rejection region**

A large value of the variable $D_n$ must be considered as an indication of a bad overlap between the empirical cumulative distribution $P_{e,n}(x)$ and the theoretical distribution $P(x)$ conjectured.

Therefore the null hypothesis will be rejected, with significance level $\alpha$, if

$$D_n > \frac{Q^{-1}(\alpha)}{\sqrt{n}}$$

In particular

$$Q^{-1}(0.10) = 1.22384 \quad Q^{-1}(0.05) = 1.35809$$

$$Q^{-1}(0.01) = 1.62762 \quad Q^{-1}(0.001) = 1.94947$$
Example: Kolmogorov-Smirnov test for a uniform distribution

An algorithm for the generation of random numbers has given the following sequence of values in the interval $[0, 1]$:

<table>
<thead>
<tr>
<th>$i$</th>
<th>$x_i$</th>
<th>$i$</th>
<th>$x_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.750</td>
<td>17</td>
<td>0.473</td>
</tr>
<tr>
<td>2</td>
<td>0.102</td>
<td>18</td>
<td>0.779</td>
</tr>
<tr>
<td>3</td>
<td>0.310</td>
<td>19</td>
<td>0.266</td>
</tr>
<tr>
<td>4</td>
<td>0.581</td>
<td>20</td>
<td>0.508</td>
</tr>
<tr>
<td>5</td>
<td>0.199</td>
<td>21</td>
<td>0.979</td>
</tr>
<tr>
<td>6</td>
<td>0.859</td>
<td>22</td>
<td>0.802</td>
</tr>
<tr>
<td>7</td>
<td>0.602</td>
<td>23</td>
<td>0.176</td>
</tr>
<tr>
<td>8</td>
<td>0.020</td>
<td>24</td>
<td>0.645</td>
</tr>
<tr>
<td>9</td>
<td>0.711</td>
<td>25</td>
<td>0.473</td>
</tr>
<tr>
<td>10</td>
<td>0.422</td>
<td>26</td>
<td>0.927</td>
</tr>
<tr>
<td>11</td>
<td>0.958</td>
<td>27</td>
<td>0.368</td>
</tr>
<tr>
<td>12</td>
<td>0.063</td>
<td>28</td>
<td>0.693</td>
</tr>
<tr>
<td>13</td>
<td>0.517</td>
<td>29</td>
<td>0.401</td>
</tr>
<tr>
<td>14</td>
<td>0.895</td>
<td>30</td>
<td>0.210</td>
</tr>
<tr>
<td>15</td>
<td>0.125</td>
<td>31</td>
<td>0.465</td>
</tr>
<tr>
<td>16</td>
<td>0.631</td>
<td>32</td>
<td>0.827</td>
</tr>
</tbody>
</table>

By using Kolmogorov-Smirnov test, check with a significance level of 5% whether the obtained sequence of data is really consistent with a uniform distribution in the interval $[0, 1]$. 
Solution
To apply the test we need to compute the empirical cumulative distribution and compare it to the cumulative distribution of a uniform RV in \([0, 1]\).

The cumulative distribution of a uniform RV in \([0, 1]\) is determined in an elementary way

\[
P(x) = \begin{cases} 
0 & x < 0 \\
x & 0 \leq x < 1 \\
1 & x \geq 1
\end{cases}
\]

To compute the empirical cumulative distribution we must rearrange the sample data into increasing order:

<table>
<thead>
<tr>
<th>(x_i)</th>
<th>(x_i)</th>
<th>(x_i)</th>
<th>(x_i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.020</td>
<td>0.297</td>
<td>0.517</td>
<td>0.779</td>
</tr>
<tr>
<td>0.063</td>
<td>0.310</td>
<td>0.581</td>
<td>0.802</td>
</tr>
<tr>
<td>0.102</td>
<td>0.368</td>
<td>0.602</td>
<td>0.827</td>
</tr>
<tr>
<td>0.125</td>
<td>0.401</td>
<td>0.631</td>
<td>0.859</td>
</tr>
<tr>
<td>0.176</td>
<td>0.422</td>
<td>0.645</td>
<td>0.895</td>
</tr>
<tr>
<td>0.199</td>
<td>0.465</td>
<td>0.693</td>
<td>0.927</td>
</tr>
<tr>
<td>0.210</td>
<td>0.473</td>
<td>0.711</td>
<td>0.958</td>
</tr>
<tr>
<td>0.266</td>
<td>0.508</td>
<td>0.750</td>
<td>0.979</td>
</tr>
</tbody>
</table>

The empirical cumulative frequency distribution of the sample is then defined by

\[
P_{e,n}(x) = \# \{ x_i, \ i = 1, \ldots, n, \ x_i \leq x \} / n
\]

in this case with \(n = 32\). Each “step” of the empirical cumulative distribution has thus an height of \(1/32 = 0.03125\).
The graphs of the two cumulative distributions are compared in the following figure for the interval $[0.000, 0.490]$ and in figure below for the residual interval $[0.490, 1.000]$.
The test variable is that of Kolmogorov-Smirnov

\[ D_n = D_{32} = \sup_{x \in \mathbb{R}} |P(x) - P_{e,32}(x)| \]

For this sample it takes the value

\[ D_{32} = 0.048 \]

which can be determined either graphically by analysing the graph of the function \(|P(x) - P_{e,32}(x)|\) in the interval \([0, 1]\):

or numerically (for instance, by the Maple command “maximize” applied to the same function in the same interval).

As shown in the figure, the calculated maximum corresponds to the value of the function at \(x = 0.827 + 0\) — the right limit of the (discontinuous) function in \(x = 0.827\).
The null hypothesis must be rejected, with a significance level $\alpha$, if

$$ D_n > \frac{Q^{-1}(\alpha)}{\sqrt{n}} $$

In this case $n = 32$ and $\alpha = 0.05$, so that

$$ Q^{-1}(0.05) = 1.35809 $$

and the rejection condition becomes

$$ D_{32} > \frac{Q^{-1}(0.05)}{\sqrt{32}} = \frac{1.35809}{5.656854249} = 0.2400786621 $$

Since the calculated value of the KS variable is smaller than the critical one

$$ D_{32} = 0.048 < 0.2400786621 $$

we conclude that the null hypothesis cannot be rejected on the basis of the sample data:

we accept the hypothesis that the sample data are drawn from a uniform population in the interval $[0,1]$

In other words, the generator of random numbers in $[0,1]$ can be considered satisfactory according to the KS test, at the given significance level.
5.2.1 Remark. Calculation of the KS variable

The calculation of the variable $D_n$ does not really require an analysis of the graph of $|P(x) - P_{e,n}(x)|$ in $x \in \mathbb{R}$.

Firstly, we must arrange the empirical values $x_1, \ldots, x_n$ in an increasing order:

$$x(1) \leq x(2) \leq \ldots \leq x(n-1) \leq x(n)$$

on having introduced $\forall j = 1, \ldots, n$ the notation

$$x(j) = j\text{-th smallest value among } x_1, x_2, \ldots, x_n.$$ 

Since $P_{e,n}(x)$ is constant in each interval of $x \in \mathbb{R}$ individuated by the empirical points:

$$(-\infty, x(1)) \quad [x(1), x(2)) \quad \ldots \quad [x(n-1), x(n)) \quad [x(n), +\infty)$$

while the cumulative distribution $P(x)$ is non-decreasing, the largest differences between the empirical and theoretical distribution may only occur at the points $x_1, \ldots, x_n$.

Therefore, it is sufficient to compute the maximum of a finite set of values:

$$\max \left\{ \left| P(x(j)) - \frac{j - 1}{n} \right|, \left| P(x(j)) - \frac{j}{n} \right|, j = 1, \ldots, n \right\}$$
5.2.2 Remark: for any given $\lambda > 0$ and $n \in \mathbb{N}$,

\[
\text{Probability}\left\{ \sup_{x \in \mathbb{R}} \left| P(x) - P_{e,n}(x) \right| > \frac{\lambda}{\sqrt{n}} \right\}
\]

in independent on the cumulative distribution $P(x)$

By the definition of $P_{e,n}(x)$, such a probability holds indeed:

\[
\text{Probability}\left\{ \sup_{x \in \mathbb{R}} \left| P(x) - \frac{\# \{ i : x_i \leq x \}}{n} \right| > \frac{\lambda}{\sqrt{n}} \right\}
\]

but since $P(x)$ is non-decreasing we have:

\[
x_i \leq x \iff P(x_i) \leq P(x)
\]

so that the previous probability becomes

\[
\text{Probability}\left\{ \sup_{x \in \mathbb{R}} \left| P(x) - \frac{\# \{ i : P(x_i) \leq P(x) \}}{n} \right| > \frac{\lambda}{\sqrt{n}} \right\}.
\]

Moreover, the definition of cumulative distribution ensures that

\[
P(x) \in [0, 1] \quad \forall x \in \mathbb{R}
\]

and that the RVs

\[
u_i = P(x_i) , \quad i = 1, \ldots, n,
\]

are independent and uniformly distributed in $[0, 1]$. As a conclusion, by posing $u = P(x)$, the required probability can be expressed in the equivalent form

\[
\text{Probability}\left\{ \sup_{u \in [0,1]} \left| u - \frac{\# \{ i : u_i \leq u \}}{n} \right| > \frac{\lambda}{\sqrt{n}} \right\}.
\]

which only depends on $n$ and $\lambda$, as claimed.
5.3 *t*-test on the mean of a normal population

It is used to check if the mean $\mu$ of a normal population is respectively equal, smaller or larger than a given value $\mu_0$.

For a normal population the test variable is

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

which for $\mu = \mu_0$ follows a Student’s distribution with $n - 1$ d.o.f.

We can distinguish three cases

(i) $H_0$: $\mu = \mu_0$ versus $H_1$: $\mu \neq \mu_0$

The test is two-sided. The rejection region consists in the union of two intervals placed symmetrically with respect to the origin $t = 0$

$$\{t \leq -t_{[1 - \frac{\alpha}{2}]}(n - 1)\} \cup \{t \geq t_{[1 - \frac{\alpha}{2}]}(n - 1)\}$$
(ii) $H_0: \mu = \mu_0$ versus $H_1: \mu > \mu_0$

In this case the test is one-sided. The rejection region is an interval of sufficiently large values of $t$. More precisely

$$\{ t \geq t_{[1-\alpha]}(n-1) \}$$

(iii) $H_0: \mu = \mu_0$ versus $H_1: \mu < \mu_0$

Just another one-sided test, but with a rejection region consisting in an interval of sufficiently small values of $t$

$$t \leq t_{[\alpha]}(n-1) = -t_{[1-\alpha]}(n-1)$$
Example: CI and hypothesis testing on the mean

Let us consider the data of the following table:

<table>
<thead>
<tr>
<th>i</th>
<th>$x_i$</th>
<th>i</th>
<th>$x_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>449</td>
<td>16</td>
<td>398</td>
</tr>
<tr>
<td>2</td>
<td>391</td>
<td>17</td>
<td>472</td>
</tr>
<tr>
<td>3</td>
<td>432</td>
<td>18</td>
<td>449</td>
</tr>
<tr>
<td>4</td>
<td>459</td>
<td>19</td>
<td>435</td>
</tr>
<tr>
<td>5</td>
<td>389</td>
<td>20</td>
<td>386</td>
</tr>
<tr>
<td>6</td>
<td>435</td>
<td>21</td>
<td>388</td>
</tr>
<tr>
<td>7</td>
<td>430</td>
<td>22</td>
<td>414</td>
</tr>
<tr>
<td>8</td>
<td>416</td>
<td>23</td>
<td>376</td>
</tr>
<tr>
<td>9</td>
<td>420</td>
<td>24</td>
<td>463</td>
</tr>
<tr>
<td>10</td>
<td>381</td>
<td>25</td>
<td>344</td>
</tr>
<tr>
<td>11</td>
<td>417</td>
<td>26</td>
<td>353</td>
</tr>
<tr>
<td>12</td>
<td>407</td>
<td>27</td>
<td>400</td>
</tr>
<tr>
<td>13</td>
<td>447</td>
<td>28</td>
<td>438</td>
</tr>
<tr>
<td>14</td>
<td>391</td>
<td>29</td>
<td>437</td>
</tr>
<tr>
<td>15</td>
<td>480</td>
<td>30</td>
<td>373</td>
</tr>
</tbody>
</table>

By assuming a normal population, we want to test the null hypothesis that the mean $\mu$ is $\mu_0 = 425$ against the alternative hypothesis that $\mu \neq 425$, with significance level $\alpha = 0.02$.

Solution

Number of data: $n = 30$

Sample mean: $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = 415.66$
Sample variance: \( s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = 1196.44 \)

Estimated standard deviation: \( s = \sqrt{s^2} = 34.59 \)

The test variable is
\[
 t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}
\]

and the two-sided rejection region writes
\[
 \{ t < -t_{[1-\alpha/2]}(n-1) \} \cup \{ t > t_{[1-\alpha/2]}(n-1) \}
\]

with \( \alpha = 0.02, \ n = 30. \)

We calculate the value of the test variable
\[
 t = \frac{415.66 - 425}{34.59/\sqrt{30}} = -1.47896
\]

and the critical point
\[
 t_{[1-\alpha/2]}(n-1) = t_{[0.99]}(29) = 2.462
\]

so that the rejection region becomes
\[
 \{ t < -2.462 \} \cup \{ t > 2.462 \}
\]

and does not contain the value of \( t. \)

The hypothesis \( \mu = 425 \) cannot be rejected on the basis of the given sample at the significance level of 0.02.
**Remark - Confidence interval for \( \mu \)**

The CI of the mean has the lower and upper limits:

\[
\bar{x} - t_{[0.99](29)} \frac{s}{\sqrt{30}} = 415.66 - 2.462 \cdot \frac{34.59}{\sqrt{30}} = 400.11
\]
\[
\bar{x} + t_{[0.99](29)} \frac{s}{\sqrt{30}} = 415.66 + 2.462 \cdot \frac{34.59}{\sqrt{30}} = 431.21
\]

and reduces then to

\[400.11 \leq \mu \leq 431.21\]

The suggested mean \( \mu_0 = 425 \) actually belongs to the confidence interval of \( \mu \).

**Generally speaking:**

The test variable takes a value in the acceptance region of \( H_0 : \mu = \mu_0 \) if and only if the tested value \( \mu_0 \) of the mean belongs to the confidence interval of \( \mu \) (the significance level of the test, \( \alpha \), and the confidence level of the CI, \( 1 - \alpha \), are complementary to 1).
### 5.4 $\chi^2$-test on the variance of a normal population

It is used to verify if the variance $\sigma^2$ of a **normal population** is respectively equal, smaller or larger than a given value $\sigma_0^2$.

The test variable takes the form

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

which, for a normal population and $\sigma^2 = \sigma_0^2$ follows a $\chi^2$ distribution with $n-1$ d.o.f.

Again, we must distinguish three cases

(i) $H_0$: $\sigma^2 = \sigma_0^2$ versus $H_1$: $\sigma^2 \neq \sigma_0^2$

Two-sided test with rejection region of $H_0$

$$\{ \chi^2 \leq \chi^2_{[\frac{\alpha}{2}]}(n-1) \} \cup \{ \chi^2 \geq \chi^2_{[1-\frac{\alpha}{2}]}(n-1) \}$$

so that the null hypothesis is rejected when $\chi^2$ is too small or too large.
(ii) $H_0: \sigma^2 = \sigma_0^2$ versus $H_1: \sigma^2 > \sigma_0^2$

The test is one-sided and the rejection region corresponds to large values of the test variable

$$\{ \chi^2 \geq \chi^2_{[1-\alpha](n-1)} \}$$

(iii) $H_0: \sigma^2 = \sigma_0^2$ versus $H_1: \sigma^2 < \sigma_0^2$

In this case also the test is one-sided; rejection occurs for very small values of the test variable

$$\{ \chi^2 \leq \chi^2_{[\alpha](n-1)} \}$$
Example: CI and $\chi^2$-test on the variance

Repeated measurements of the electromotive force at the ends of a photovoltaic cell, exposed to a radiation of constant intensity and spectrum, have given the following table of data (in V), which can be assumed to belong to a normal population:

<table>
<thead>
<tr>
<th>$i$</th>
<th>$V_i$</th>
<th>$i$</th>
<th>$V_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.426</td>
<td>16</td>
<td>1.497</td>
</tr>
<tr>
<td>2</td>
<td>1.353</td>
<td>17</td>
<td>1.484</td>
</tr>
<tr>
<td>3</td>
<td>1.468</td>
<td>18</td>
<td>1.466</td>
</tr>
<tr>
<td>4</td>
<td>1.379</td>
<td>19</td>
<td>1.383</td>
</tr>
<tr>
<td>5</td>
<td>1.481</td>
<td>20</td>
<td>1.411</td>
</tr>
<tr>
<td>6</td>
<td>1.413</td>
<td>21</td>
<td>1.428</td>
</tr>
<tr>
<td>7</td>
<td>1.485</td>
<td>22</td>
<td>1.358</td>
</tr>
<tr>
<td>8</td>
<td>1.476</td>
<td>23</td>
<td>1.483</td>
</tr>
<tr>
<td>9</td>
<td>1.498</td>
<td>24</td>
<td>1.473</td>
</tr>
<tr>
<td>10</td>
<td>1.475</td>
<td>25</td>
<td>1.482</td>
</tr>
<tr>
<td>11</td>
<td>1.467</td>
<td>26</td>
<td>1.435</td>
</tr>
<tr>
<td>12</td>
<td>1.478</td>
<td>27</td>
<td>1.424</td>
</tr>
<tr>
<td>13</td>
<td>1.401</td>
<td>28</td>
<td>1.521</td>
</tr>
<tr>
<td>14</td>
<td>1.492</td>
<td>29</td>
<td>1.399</td>
</tr>
<tr>
<td>15</td>
<td>1.376</td>
<td>30</td>
<td>1.489</td>
</tr>
</tbody>
</table>

(a) Calculate the CI of the variance, at a confidence level of 98%. (b) Moreover, check with a significance level of 2% the hypothesis that $\sigma^2 = \sigma_0^2 = 32.0 \cdot 10^{-4}$.

Solution

Firstly we compute the sample estimates of the mean, variance and standard deviation.
Number of data: \( n = 30 \)

Sample mean: \( \bar{V} = \frac{1}{n} \sum_{i=1}^{n} V_i = 1.4467 \)

Sample variance:
\[
s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (V_i - \bar{V})^2 = 0.00221318
\]

Sample standard deviation: \( s = \sqrt{s^2} = 0.04704448 \)

We can now proceed to determine the CI of the variance and verify the hypothesis that \( \sigma^2 = \sigma_0^2 = 32.0 \cdot 10^{-4} \).

**a)** The CI of the variance takes the form
\[
\frac{1}{\chi^2_{[1-\alpha/2]}(n-1)}(n-1)s^2 \leq \sigma^2 \leq \frac{1}{\chi^2_{[\alpha/2]}(n-1)}(n-1)s^2
\]
with \( \alpha = 0.02 \) and \( n = 30 \). Therefore
\[
\frac{29 \cdot s^2}{\chi^2_{[0.99]}(29)} = \frac{29 \cdot 0.00221318}{49.588} = 12.94311 \cdot 10^{-4}
\]
whereas
\[
\frac{29 s^2}{\chi^2_{[0.01]}(29)} = \frac{29 \cdot 0.00221318}{14.256} = 45.02125 \cdot 10^{-4}
\]
and the CI becomes
\[
12.94311 \cdot 10^{-4} \leq \sigma^2 \leq 45.02125 \cdot 10^{-4}.
\]
(b) The RV suitable for the test is
\[ \chi^2 = \frac{(n - 1)s^2}{\sigma_0^2} \]
which, if the null hypothesis \( H_0 : \sigma^2 = \sigma_0^2 \) holds true, follows a \( \chi^2 \) distribution with \( n - 1 \) d.o.f.

The required significance level is \( \alpha = 2\% = 0.02 \).

By assuming the alternative hypothesis \( H_1 : \sigma^2 \neq \sigma_0^2 \), the critical region takes the two-sided form
\[ \{ \chi^2 \leq \chi^2_{[\frac{\alpha}{2}]}(n-1) \} \cup \{ \chi^2 \geq \chi^2_{[1-\frac{\alpha}{2}]}(n-1) \} \]
with \( \alpha = 0.02 \) and \( n = 30 \). As before, the table of \( \chi^2 \) provides
\[ \chi^2_{[\frac{\alpha}{2}]}(n-1) = \chi^2_{[0.01]}(29) = 14.256 \]
\[ \chi^2_{[1-\frac{\alpha}{2}]}(n-1) = \chi^2_{[0.99]}(29) = 49.588 \]
so that the critical region of \( H_0 \) becomes
\[ \{ \chi^2 \leq 14.256 \} \cup \{ \chi^2 \geq 49.588 \} \]

In the present case the test statistic holds
\[ \chi^2 = \frac{(n - 1)s^2}{\sigma_0^2} = \frac{29 \cdot 0.002213183}{32 \cdot 10^{-4}} = 20.057 \]
which does not belong to the critical region: \( H_0 \) cannot be rejected.

Notice that accepting the hypothesis \( \sigma^2 = \sigma_0^2 \) at a significance level of \( \alpha \) is equivalent to check that \( \sigma_0^2 \) belongs to the CI of \( \sigma^2 \) with a confidence level of \( 1 - \alpha \).
5.5 z-test to compare the means of two independent normal populations (of known variances)

The sample estimates of the means

\[ \bar{y} = \frac{1}{p} \sum_{i=1}^{p} y_i \quad \bar{z} = \frac{1}{q} \sum_{j=1}^{q} z_j \]

are normal RVs with means \( \mu_1 \) and \( \mu_2 \), while their variances write

\[ \frac{\sigma_1^2}{p} \quad \text{and} \quad \frac{\sigma_2^2}{q} \]

As a consequence, when \( \mu_1 = \mu_2 \) the random variable

\[ z = \frac{\bar{y} - \bar{z}}{\sqrt{\frac{\sigma_1^2}{p} + \frac{\sigma_2^2}{q}}} \]

is a standard normal \( N(0,1) \). With a significance level \( \alpha \in (0.5, 1) \), the two-sided rejection region for the null hypothesis \( H_0: \mu_1 = \mu_2 \) versus the alternative \( H_1: \mu_1 \neq \mu_2 \) can then be expressed as

\[ \{ z \leq -z_{[1-\frac{\alpha}{2}]} \} \cup \{ z_{[1-\frac{\alpha}{2}]} \leq z \} \]

where the critical value \( z_{[1-\frac{\alpha}{2}]} > 0 \) is defined by the equation

\[ \frac{1}{\sqrt{2\pi}} \int_{0}^{z_{[1-\frac{\alpha}{2}]}} e^{-\xi^2/2} d\xi = \frac{1 - \alpha}{2} \]

i.e. by \( 1 - \alpha = \text{erf}(z_{[1-\frac{\alpha}{2}]/\sqrt{2}}) \)
5.6 \textit{F-test to compare the variances of two independent normal populations}

It is used to test if two independent normal populations of variances $\sigma_1^2$ and $\sigma_2^2$, respectively, have the same variance

Therefore, we have to test the null hypothesis $H_0: \sigma_1^2 = \sigma_2^2$ versus the alternative hypothesis $H_1: \sigma_1^2 \neq \sigma_2^2$

We assume that two small samples are available, of $p$ and $q$ individuals, respectively

\[(y_1, y_2, \ldots, y_p) \quad (z_1, z_2, \ldots, z_q)\]

obtained from the two independent populations

The estimated means of the two populations are given by

\[
\bar{y} = \frac{1}{p} \sum_{i=1}^{p} y_i \quad \bar{z} = \frac{1}{q} \sum_{j=1}^{q} z_j
\]

while the estimates of the variances can be written as

\[
s_y^2 = \frac{1}{p-1} \sum_{i=1}^{p} (y_i - \bar{y})^2 \quad s_z^2 = \frac{1}{q-1} \sum_{j=1}^{q} (z_j - \bar{z})^2
\]
Since the samples are obtained from independent normal populations, the random variables
\[
\frac{(p - 1)s^2_y}{\sigma_1^2} = \frac{1}{\sigma_1^2} \sum_{i=1}^{p} (y_i - \overline{y})^2 \quad \frac{(q - 1)s^2_z}{\sigma_2^2} = \frac{1}{\sigma_2^2} \sum_{j=1}^{q} (z_j - \overline{z})^2
\]
are independent $X^2$ variables with $p - 1$ and $q - 1$ d.o.f., respectively.

We deduce that the ratio
\[
F = \frac{q - 1}{p - 1} \frac{(p - 1)s^2_y}{\sigma_1^2} \left( \frac{(q - 1)s^2_z}{\sigma_2^2} \right)^{-1} = \frac{\sigma_2^2 s^2_y}{\sigma_1^2 s^2_z}
\]
is by definition a Fisher $F$ with $p - 1$, $q - 1$ d.o.f.

If the null hypothesis $\sigma_1^2 = \sigma_2^2$ is correct, the sample variable reduces to
\[
F = \frac{s^2_y}{s^2_z}
\]
The critical region can be then characterized by means of the $F$ cumulative distribution with $(n_1, n_2)$ d.o.f.

\[
P(\tau) = \int_0^\tau \frac{\Gamma\left(\frac{n_1 + n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right) \Gamma\left(\frac{n_2}{2}\right)} \left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} \frac{F^{n_1-1}}{\left(1 + \frac{n_1}{n_2} F\right)^{\frac{n_1 + n_2}{2}}} dF
\]
which expresses the probability that $F$ takes a value smaller than a fixed $\tau \geq 0$. For any $\alpha \in (0, 1)$ we pose $F_{[\alpha]}(n_1, n_2) = P^{-1}(\alpha)$, so that
\[
P(F_{[\alpha]}(n_1, n_2)) = \alpha
\]
Very small or very large values of $F$ must be regarded as an indication of a ratio $\sigma_1^2/\sigma_2^2$ significantly different from 1, so that the rejection region of $H_0$ versus $H_1$ is two-sided

$$\{ F \leq F_{[\frac{\alpha}{2}]}(p-1,q-1) \} \cup \{ F \geq F_{[1-\frac{\alpha}{2}]}(p-1,q-1) \}$$

If the alternative hypothesis is $H_1$: $\sigma_1^2 > \sigma_2^2$, the rejection region is defined in a one-sided way

$$\{ F \geq F_{[1-\alpha]}(p-1,q-1) \}$$

since large values of $s_y^2/s_z^2$ are expected when $H_1$ holds true.
Finally, in the case that the alternative hypothesis is $H_1: \sigma_1^2 < \sigma_2^2$, the rejection region is still one-sided

$$\left\{ F \leq F_{[\alpha]}(p-1,q-1) \right\}$$

and small values of $s_y^2/s_z^2$ will suggest the correctness of $H_1$

![Diagram showing the rejection region](image)

**Calculation of $F_{[\alpha]}(n_1,n_2)$ by Maple 11**

By definition $F_{[\alpha]}(n_1,n_2)$ is simply the inverse of the Fisher cumulative distribution with $(n_1,n_2)$ at $\alpha$, which can be calculated by the Maple command lines:

```maple
with(stats):
statevalf[icdf, fratio[n1, n2]](alpha);
```

As an example, for $\alpha = 0.85$ and $(n_1,n_2) = (10,13)$ we have

```
statevalf[icdf, fratio[10, 13]](0.85);
```

1.841918109
so that
\[ F_{[0.85]}(10,13) = 1.841918109 \]

**Calculation of** $F_{[\alpha]}(n_1,n_2)$ **by Excel**

The Excel worksheet function to be applied in this case is FINV. For any $\alpha \in (0, 1)$ the command

\[ \text{FINV}(\alpha; n_1; n_2) \]

computes the value of the Fisher RV with $(n_1, n_2)$ d.o.f. which the RV may *exceed* with probability $\alpha$. Thus, it corresponds to the inverse of the Fisher cumulative distribution calculated at $1 - \alpha$. Therefore, the value of $F_{[\alpha]}(n_1,n_2)$ can be obtained by computing

\[ \text{FINV}(1 - \alpha; n_1; n_2) \]

In the previous example, the command would become

\[ \text{FINV}(0,15; 10; 13) \]
Example: F-test on the variances of two independent normal populations
In order to enhance a chemical reaction we must choose between two different catalysts, say, $A$ and $B$. To check if the variance of the amount of product is the same or not, for the two catalysts, we carry out $p = 10$ experiments with $A$ and $q = 12$ experiments with $B$ and obtain the following sample estimates:

$$s_{A}^2 = 0.1405 \quad s_{B}^2 = 0.2820$$

At a significance level of 5%, can we reject the hypothesis that $\sigma_A^2 = \sigma_B^2$? The two populations are assumed independent and normal.

Solution
We can use the test variable

$$F = \frac{s_{A}^2}{s_{B}^2}$$

which follows a Fisher distribution with $(p - 1, q - 1) = (9, 11)$ d.o.f. if the null hypothesis $H_0 : \sigma_A^2 = \sigma_B^2$ is satisfied.

We have

$$F = \frac{0.1405}{0.2820} = 0.49822695$$

The rejection region is

$$\{F \leq F_{[0.05]}(9,11)\} \cup \{F \geq F_{[1-0.05]}(9,11)\}$$

i.e. (since $p = 10$, $q = 12$ and $\alpha = 0.05$)

$$\{F \leq F_{[0.025]}(9,11)\} \cup \{F \geq F_{[0.975]}(9,11)\}$$
where

\[
F_{0.025}(9,11) = 0.2556188 \\
F_{0.975}(9,11) = 3.587898
\]

The $F$-values are not available on the table, but can be easily calculated by numerical tools. For instance, by the following Maple commands

\[
\text{statevalf}[\text{icdf}, \text{fratio}[9, 11]](0.025)
\]

\[
\text{statevalf}[\text{icdf}, \text{fratio}[9, 11]](0.975)
\]

In the present case

\[
0.49822695 \notin \{ F \leq 0.2556188 \} \cup \{ F \geq 3.587898 \}
\]

thus we must conclude that our samples cannot exclude that the variances of the two populations are actually the same. $H_0$ **cannot be rejected!**
5.7 Unpaired $t$-test to compare the means of two independent normal populations (unknown variances)

For two given normal populations of mean $\mu_1$ and $\mu_2$, we want to test the null hypothesis $H_0$: $\mu_1 = \mu_2$ versus the alternative hypothesis $H_1$: $\mu_1 \neq \mu_2$

We distinguish the case where the populations have the same variance $\sigma_1^2 = \sigma_2^2$ and that of unequal variances $\sigma_1^2 \neq \sigma_2^2$

The correctness of the hypothesis $\sigma_1^2 = \sigma_2^2$ or $\sigma_1^2 \neq \sigma_2^2$ may be checked by the suitable $F$-test previously described (variances are unknown!)

5.7.1 Case of equal variances: $\sigma_1^2 = \sigma_2^2 = \sigma^2$

If $\mu_1 = \mu_2$ and $s^2$ denotes the so-called pooled variance

$$s^2 = \frac{(p - 1)s_y^2 + (q - 1)s_z^2}{p + q - 2},$$

the random variable

$$t = \frac{\bar{y} - \bar{z}}{s \sqrt{\frac{1}{p} + \frac{1}{q}}}$$

follows a Student’s distribution with $p + q - 2$ d.o.f.

$t$ will then be used as a test variable, with a rejection region given by the union of two intervals, corresponding to large absolute values of $t$

$$\{t \leq -t_{\alpha/2}(p+q-2)\} \cup \{t \geq t_{\alpha/2}(p+q-2)\}$$
5.7.2 Case of unequal variances $\sigma_1^2 \neq \sigma_2^2$

If $\mu_1 = \mu_2$ the random variable

$$t = \frac{\overline{y} - \overline{z}}{\sqrt{\frac{1}{p} s_y^2 + \frac{1}{q} s_z^2}}$$

follows approximately a Student’s distribution with a generally non-integer number of d.o.f.

$$n = \frac{\left(\frac{s_y^2}{p} + \frac{s_z^2}{q}\right)^2}{\frac{1}{p - 1} \left(\frac{s_y^2}{p}\right)^2 + \frac{1}{q - 1} \left(\frac{s_z^2}{q}\right)^2}$$

Student’s distribution is defined anyway and the rejection region can be written as

$$\{ t \leq -t_{[1-\frac{\alpha}{2}]}(n) \} \cup \{ t \geq t_{[1-\frac{\alpha}{2}]}(n) \}$$

No exact test is available! (Behrens-Fisher’ problem)
5.7.3 Remark for the case $\sigma_1^2 = \sigma_2^2$

Proof that the test variable actually follows a Student’s distribution with $p + q - 2$ d.o.f.

Notice that for $\mu_1 = \mu_2$:

(i) the random variable

$$\xi = \frac{\bar{y} - \bar{z}}{\sigma \sqrt{\frac{1}{p} + \frac{1}{q}}} = \frac{\bar{y} - \bar{z}}{\sigma} \sqrt{\frac{pq}{p + q}}$$

follows a standard normal distribution

(ii) the random variable

$$\chi^2 = \frac{1}{\sigma^2} \left[ \sum_{i=1}^{p} (y_i - \bar{y})^2 + \sum_{j=1}^{q} (z_j - \bar{z})^2 \right]$$

is a $\chi^2$ variable with $p - 1 + q - 1 = p + q - 2$ d.o.f.

(iii) the random variables $\xi$ and $\chi^2$ are stochastically independent, as it can be deduced from Craig’s theorem

As a consequence, the test variable

$$t = \sqrt{p + q - 2} \frac{\xi}{\chi^2} =$$

$$= \frac{\bar{y} - \bar{z}}{\sigma \sqrt{\frac{1}{p} + \frac{1}{q}}} \frac{\sqrt{p + q - 2}}{\sigma^{-1} \sqrt{(p - 1)s^2_y + (q - 1)s^2_z}} = \frac{\bar{y} - \bar{z}}{s \sqrt{\frac{1}{p} + \frac{1}{q}}}$$

follows a Student’s distribution with $p + q - 2$ d.o.f.

Stefano Siboni
A justification of the result can be obtained by observing that if we pose

\[(x_1, \ldots, x_p, x_{p+1}, \ldots, x_{p+q}) = (y_1, \ldots, y_p, z_1, \ldots, z_q)\]

and

\[\bar{x} = \frac{1}{p + q} \sum_{k=1}^{p+q} x_k = \frac{1}{p + q} \left[ \sum_{i=1}^{p} y_i + \sum_{j=1}^{q} z_j \right]\]

there holds the identify below

\[\sum_{k=1}^{p+q} (x_k - \bar{x})^2 = \sum_{i=1}^{p} (y_i - \bar{y})^2 + \sum_{j=1}^{q} (z_j - \bar{z})^2 + \frac{pq}{p + q} (\bar{y} - \bar{z})^2\]

where:

\[\frac{1}{\sigma^2} \sum_{k=1}^{p+q} (x_k - \bar{x})^2\] is a \(\chi^2\) variable with \(p + q - 1\) d.o.f.

\[\frac{1}{\sigma^2} \left[ \sum_{i=1}^{p} (y_i - \bar{y})^2 + \sum_{j=1}^{q} (z_j - \bar{z})^2 \right]\] is a \(\chi^2\) with \(p + q - 2\) d.o.f.

\[\frac{1}{\sigma^2} \frac{pq}{p + q} (\bar{y} - \bar{z})^2\] is a \(\chi^2\) with 1 d.o.f.

The previous properties imply the stochastic independence of the random variables

\[\frac{1}{\sigma^2} \left[ \sum_{i=1}^{p} (y_i - \bar{y})^2 + \sum_{j=1}^{q} (z_j - \bar{z})^2 \right]\] and \[\frac{1}{\sigma^2} \frac{pq}{p + q} (\bar{y} - \bar{z})^2\]

and therefore of

\[\frac{1}{\sigma^2} \left[ \sum_{i=1}^{p} (y_i - \bar{y})^2 + \sum_{j=1}^{q} (z_j - \bar{z})^2 \right]\] and \[\frac{1}{\sigma} \sqrt{\frac{pq}{p + q} (\bar{y} - \bar{z})}\]
Example: Unpaired \( t \)-test for the comparison of two means (same variance)

Two processes are characterized by the values listed in the table below:

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>14</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td>7</td>
<td>12</td>
</tr>
<tr>
<td>9</td>
<td>13</td>
</tr>
<tr>
<td>10</td>
<td>8</td>
</tr>
</tbody>
</table>

We want to test, with a 5%-significance level, the hypothesis that the mean values \( \mu_1 \) and \( \mu_2 \) of the processes are the same, by assuming normal populations and equal variances \( (\sigma_1^2 = \sigma_2^2) \).

**Solution**

The two samples are composed by \( p = 6 \) and \( q = 7 \) elements, respectively. The mean and the variance of the first sample can be written as:

\[
\bar{x} = \frac{1}{p} \sum_{i=1}^{p} x_i = 8.17 \quad s^2_x = \frac{1}{p-1} \sum_{i=1}^{p} (x_i - \bar{x})^2 = 5.37
\]

while those of the second sample hold:

\[
\bar{y} = \frac{1}{q} \sum_{j=1}^{q} y_j = 11.29 \quad s^2_y = \frac{1}{q-1} \sum_{j=1}^{q} (y_j - \bar{y})^2 = 5.24
\]
The pooled variance of the two samples is then calculated in the form:

\[ s^2 = \frac{(p - 1)s_x^2 + (q - 1)s_y^2}{p + q - 2} = \frac{5 \cdot 5.37 + 6 \cdot 5.24}{11} = 5.299 \]

As a test variable for testing the hypothesis \( H_0 : \mu_1 = \mu_2 \), we use the quotient:

\[ t = \frac{\bar{x} - \bar{y}}{s \sqrt{\frac{1}{p} + \frac{1}{q}}} = \frac{8.17 - 11.29}{\sqrt{5.299 \left( \frac{1}{6} + \frac{1}{7} \right)}} = -2.436 \]

If the null hypothesis holds true, such a random variable is a Student’s t with \( p + q - 2 \) d.o.f. so that the two-sided acceptance region must be taken as:

\[ |t| \leq t_{[1-\frac{\alpha}{2}](p+q-2)} \]

In the present case we have \( \alpha = 0.05 \), \( p = 6 \), \( q = 7 \) and the acceptance region becomes

\[ |t| \leq t_{[0.975](11)} = 2.201 \]

Since \( |t| = 2.436 > 2.201 = t_{[0.975](11)} \), our samples suggest that the null hypothesis \( \mu_1 = \mu_2 \) must be rejected: we conclude that the difference between the means of the two processes appears significant.
5.8 Paired \(t\)-test to compare the means of two normal populations (unknown variances)

It should be used when the samples obtained from the two normal populations, of means \(\mu_1\), \(\mu_2\) and arbitrary variances, cannot be considered independent since data are \textit{organized by pairs of values belonging to each population} \((p = q = n)\)

\[(y_1, z_1), (y_2, z_2), \ldots, (y_n, z_n)\]

The null hypothesis to be tested is always \(H_0 : \mu_1 = \mu_2\) versus the alternative hypothesis \(H_1 : \mu_1 \neq \mu_2\)

The differences between data

\[d_i = y_i - z_i, \quad i = 1, \ldots, n\]

are i.i.d. normal variables with mean \(\mu_1 - \mu_2\) and variance \(\sigma_1^2 + \sigma_2^2\)

If \(H_0\) is true the \(d_i\) constitute a sample of a normal variable \(d\) of zero mean and variance \(\sigma_d^2 = \sigma_1^2 + \sigma_2^2\), so that

\[
t = \frac{\bar{d}}{s_d/\sqrt{n}} = \frac{\bar{y} - \bar{z}}{1/\sqrt{n} \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - z_i - \bar{y} + \bar{z})^2}}
\]

is a Student’s \(t\) with \(n - 1\) d.o.f.

The rejection region, with significance level \(\alpha\), can be written as

\[
\{ t \leq -t_{1-\frac{\alpha}{2}}(n-1) \} \cup \{ t \geq t_{1-\frac{\alpha}{2}}(n-1) \}
\]

The test is known as \textit{paired} \(t\)-\textit{test}
Example: Paired $t$-test to compare the means of two normal populations

A certain set of specimens is subjected to a particular physico-chemical treatment. A physical quantity is measured for each specimen prior to and after the treatment. The results are summarized in the following table

<table>
<thead>
<tr>
<th>before the treatment</th>
<th>after the treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.5</td>
<td>10.4</td>
</tr>
<tr>
<td>4.2</td>
<td>9.5</td>
</tr>
<tr>
<td>10.1</td>
<td>11.6</td>
</tr>
<tr>
<td>7.3</td>
<td>8.5</td>
</tr>
<tr>
<td>9.6</td>
<td>11.0</td>
</tr>
<tr>
<td>7.9</td>
<td>10.9</td>
</tr>
<tr>
<td>10.1</td>
<td>9.3</td>
</tr>
<tr>
<td>8.7</td>
<td>10.3</td>
</tr>
</tbody>
</table>

Check, with a significance level of 5%, the hypothesis that the means $\mu_1$ and $\mu_2$ of the measured quantity are the same before and after the treatment, by assuming normal populations.

Solution

In this case it seems natural to apply a paired $t$-test for the comparison of the means, as the quantity is measured before and after the treatment on each specimen. Therefore, we will couple up the values relative to the same specimen:

$$(y_i, z_i) \quad i = 1, \ldots, n$$

denoting with $y_i$ and $z_i$ the values measured before and after the treatment, respectively, on the whole number $n = 8$ of specimens.
We check the hypothesis $H_0 : \mu_1 = \mu_2$, that the treatment has no effect on the mean (and thus on the true value) of the measured quantity, versus the alternative hypothesis $H_1 : \mu_1 \neq \mu_2$. The test variable is

$$t = \sqrt{n} \frac{\bar{y} - \bar{z}}{\sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - z_i - \bar{y} + \bar{z})^2}}$$

which for $H_0$ true follows a Student distribution with $n - 1 = 7$ d.o.f.

The rejection region, with a significance level $\alpha$, is of the form

$$\{ t \leq -t_{[1-\frac{\alpha}{2}] (n-1)} \} \cup \{ t \geq t_{[1-\frac{\alpha}{2}] (n-1)} \}$$

and for $n = 8$, $\alpha = 0.05$ becomes

$$\{ t \leq -2.365 \} \cup \{ t \geq 2.365 \}$$

since

$$t_{[1-\frac{\alpha}{2}] (n-1)} = t_{[0.975]}(7) = 2.365$$

In the present case, we have:

$$\bar{y} = \frac{1}{8} \sum_{i=1}^{8} y_i = 8.425$$

$$\bar{z} = \frac{1}{8} \sum_{i=1}^{8} z_i = 10.1875$$
while

\[ \frac{1}{n-1} \sum_{i=1}^{n} (y_i - z_i - \overline{y} + \overline{z})^2 = \frac{1}{7} \cdot 21.89875 = 3.12839286 \]

and therefore

\[ \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - z_i - \overline{y} + \overline{z})^2} = \sqrt{3.12839286} = 1.76872634 \]

so that the test variable takes the value

\[ t = \sqrt{8} \cdot \frac{8.425 - 10.1875}{1.76872634} = -2.8184704 \]

The calculated value belongs to the lower tail of the critical region

\[ -2.8184704 < -2.365 \]

and therefore we have to exclude, with a significance level of 5%, that the mean value of the quantity prior to and after the treatment be the same.

**The null hypothesis is rejected!**

**Remark**

All the \( t \)-tests for the comparison of the means of two normal populations are implemented in Excel by the function TTEST:

- unpaired with equal variances
- unpaired with different variances
- paired

in both the two-sided and the one-sided cases.
5.9 F-test for the comparison of means

Typical problem
A material is subjected to various kinds of treatment treatments $1, 2, \ldots, r$ and on each treated material some measurements of a given property $x$ are performed measured values of $x$ on the material subjected $x_{ij}$ $i = 1, 2, \ldots, r$
to the treatment $i$ $j = 1, 2, \ldots, c$

(we assume for simplicity that the number of measurements is always the same for all the treatments)

<table>
<thead>
<tr>
<th>treatment</th>
<th>measured values of $x$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 2 \ldots \ldots c-1  c</td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>$i$</td>
<td></td>
</tr>
<tr>
<td>$r$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$i$</th>
<th>\ldots</th>
<th>$x_{ij}$</th>
<th>\ldots</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Question
Are the mean values of the quantity $x$ significantly different for the various treatments?

Or are they essentially equal, so that the various treatments do not modify the property $x$ in a significant way?
The model

Set of random variables with two indices

\[ x_{ij} \quad i = 1, \ldots, r \quad j = 1, \ldots, c \]

\( i \) specifies the population (treatment)

\( j \) identifies an element of the \( i \)-th population

(value measured on the sample subjected to the treatment \( i \))

For all \( i = 1, \ldots, r \) and \( j = 1, \ldots, c \) the random variable \( x_{ij} \) is normal of mean \( \mu_i \) and variance \( \sigma^2 \)

Remark All the populations are characterized by the same variance \( \sigma^2 \), but they could have different means \( \mu_i \)

Hypothesis to be tested

\[ H_0 : \mu_i = \mu \quad \forall \ i = 1, \ldots, r \] (equal means)

\[ H_1 : \mu_h \neq \mu_k \text{ for at least one } h \text{ and one } k \text{ in } 1, \ldots, r \] (at least two unequal means)

Method

1-factor analysis of variance
(ANalysis Of VAriance - ANOVA)
ANOVA

For a set of data ranked in \( r \) samples of \( c \) data each, the 1-factor analysis of variance consists in the interpretation (or “explanation”) of data spread by distinguishing

- a contribution due to the choice of the sample (row)
- an intrinsic contribution of the single samples

The samples are classified according to an appropriate experimental condition, or “factor”

In the present example, the only factor which classify the various samples is the kind of treatment

The ANOVA we apply is a 1-factor ANOVA

It may happen that the samples are classified by two or more experimental conditions varied in a controlled way

For instance, the treatment could consist in the action of a chemical reagent at various temperatures

\[
\text{sample} \iff \text{chemical reagent} + \text{temperature}
\]

In that case each sample will be identified by a pair of indices and we will use a 2-factor ANOVA

Further, the sample could be specified by the chemical reagent employed, by the temperature and by the duration of the treatment

\[
\text{sample} \iff \text{chemical reagent} + \text{temperature} + \text{time}
\]

that is by a triplet of indices. We will have to apply a 3-factor ANOVA
General mathematical formulation of ANOVA

1-factor ANOVA
when the factor is related to the first index

\[ x_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad \sum_i \alpha_i = 0 \quad \varepsilon_{ij} \text{ i.i.d. } N(0, \sigma) \]

or to the second index

\[ x_{ij} = \mu + \beta_j + \varepsilon_{ij} \quad \sum_j \beta_j = 0 \quad \varepsilon_{ij} \text{ i.i.d. } N(0, \sigma) \]

2-factor ANOVA without interaction

\[ x_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \quad \sum_i \alpha_i = 0 \quad \sum_j \beta_j = 0 \quad \varepsilon_{ij} \text{ i.i.d. } N(0, \sigma) \]

2-factor ANOVA with interaction

\[ x_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ij} \quad \sum_i \alpha_i = 0 \quad \sum_j \beta_j = 0 \]

\[ \sum_i (\alpha\beta)_{ij} = 0 \quad \forall j \quad \sum_j (\alpha\beta)_{ij} = 0 \quad \forall i \quad \varepsilon_{ij} \text{ i.i.d. } N(0, \sigma) \]

3-factor ANOVA without interaction

\[ x_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \varepsilon_{ijk} \quad \varepsilon_{ijk} \text{ i.i.d. } N(0, \sigma) \]

\[ \sum_i \alpha_i = 0 \quad \sum_j \beta_j = 0 \quad \sum_k \gamma_k = 0 \]

eetc. etc.
5.9.1 1-factor ANOVA

The 1-factor ANOVA makes use of some fundamental definitions

Mean of data

$$\bar{x} = \frac{1}{rc} \sum_{i=1}^{r} \sum_{j=1}^{c} x_{ij} = \frac{1}{rc} \sum_{i,j} x_{ij}$$

Mean of a sample

the mean of the data in a row, i.e. within a sample

$$\bar{x}_i. = \frac{1}{c} \sum_{j=1}^{c} x_{ij}$$

Residual

difference between the value of a variable and the mean of data

$$x_{ij} - \bar{x}$$

Residual “explained” by the factor corresponding to index $i$

the part of the residual which can be attributed to the different means of the various samples, as estimated by the relative sample mean

$$\bar{x}_i. - \bar{x}$$

“Unexplained” residual

the remaining component of the residual, which describes the spread of the data within the sample (i.e. within the rows)

$$x_{ij} - \bar{x} - (\bar{x}_i. - \bar{x}) = x_{ij} - \bar{x}_i.$$
**Total variation**
the (residual) sum of squares

\[ SS_t = \sum_{i,j} (x_{ij} - \bar{x})^2 \]

**Variation explained by the factor related to index i**
the sum of squares of residuals explained by the factor

\[ SS_1 = \sum_{i,j} (\bar{x}_i - \bar{x})^2 = c \sum_{i=1}^{r} (\bar{x}_i - \bar{x})^2 \]

It expresses the data variation due to the fact that the samples (rows) have unequal means

**Unexplained variation**
the sum of squares of unexplained residuals

\[ SS_u = \sum_{i,j} (x_{ij} - \bar{x}_i)^2 \]

It describes the data spread within each sample (row), due to purely random phenomena and not to belonging to one sample or another

It is independent on having the samples the same mean or not
1-factor ANOVA

Model:

\[ x_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad i = 1, \ldots, r \quad j = 1, \ldots, c \]

with:

\[ \sum_{i} \alpha_i = 0 \]

and

\[ \varepsilon_{ij} \text{ i.i.d. random variables } N(0, \sigma) \forall i, j \]

We make the strong assumption that \( \sigma \) is the same for all the variables

The algebraic identity

\[
\sum_{i,j} (x_{ij} - \overline{x})^2 = \text{total variation } SS_t
\]

\[ = \sum_{i,j} (\overline{x}_i. - \overline{x})^2 \quad \text{explained variation } SS_1 \]

\[ + \sum_{i,j} (x_{ij} - \overline{x}_i.)^2 \quad \text{unexplained variation } SS_u \]

holds \textbf{independently on the correctness of the model}
If the model is correct there also holds

$$\frac{1}{\sigma^2} \sum_{i,j} (x_{ij} - \bar{x} - \alpha_i)^2 =$$

$$= \left\langle \frac{1}{\sigma} (\varepsilon_{ij}) \left| (I - P_1 P_2) \frac{1}{\sigma} (\varepsilon_{ij}) \right\rangle = \chi^2_{rc-1}$$

$$\frac{1}{\sigma^2} \sum_{i,j} (\bar{x}_i. - \bar{x} - \alpha_i)^2 =$$

$$= \left\langle \frac{1}{\sigma} (\varepsilon_{ij}) \left| P_2 (I - P_1) \frac{1}{\sigma} (\varepsilon_{ij}) \right\rangle = \chi^2_{r-1}$$

$$\frac{1}{\sigma^2} \sum_{i,j} (x_{ij} - \bar{x}_i.)^2 =$$

$$= \left\langle \frac{1}{\sigma} (\varepsilon_{ij}) \left| (I - P_2) \frac{1}{\sigma} (\varepsilon_{ij}) \right\rangle = \chi^2_{r(c-1)}$$

with $\chi^2_{r-1}$ and $\chi^2_{r(c-1)}$ stochastically independent

List of symbols:

- $(v_{ij}) = (v_{11}, v_{12}, \ldots, v_{rc}) \in \mathbb{R}^{rc}$ (arbitrary vector of $\mathbb{R}^{rc}$)
- $P_1(v_{ij}) = (\overline{v}.j)$ (operator mean over the 1st index)
- $P_2(v_{ij}) = (\overline{v}.i)$ (operator mean over the 2nd index)
- $\alpha(v_{ij}) = (\alpha v_{ij}) \quad \forall \alpha \in \mathbb{R}, \forall (v_{ij}) \in \mathbb{R}^{rc}$
- $\langle (v_{ij}) | (u_{ij}) \rangle = \sum_{ij} v_{ij} u_{ij} \quad \forall (v_{ij}), (u_{ij}) \in \mathbb{R}^{rc}$ (usual inner product in $\mathbb{R}^{rc}$)
**Fundamental theorem**

For any set of data \( x_{ij} \) the total variation is the sum of the variation explained by the factor related to \( i \) and of the unexplained one

\[
SS_t = SS_1 + SS_u
\]

If moreover the variables \( x_{ij} \) are independent and identically distributed, with normal distribution of mean \( \mu \) and variance \( \sigma^2 \), we have that

\[
SS_t/\sigma^2 \quad \text{is a } \chi^2 \text{ variable with } rc - 1 \text{ d.o.f.}
\]

\[
SS_1/\sigma^2 \quad \text{is a } \chi^2 \text{ variable with } r - 1 \text{ d.o.f.}
\]

\[
SS_u/\sigma^2 \quad \text{is a } \chi^2 \text{ variable with } rc - r \text{ d.o.f.}
\]

\[
SS_1/\sigma^2 \text{ and } SS_u/\sigma^2 \text{ are stochastically independent}
\]

**Remarks**

- we have a balance on the number of d.o.f.

\[
rc - 1 = (r - 1) + (rc - r)
\]

- the (residual) sums of squares divided by the respective number of d.o.f. are defined as **variances**

\[
SS_t/(rc - 1) \text{ total variance}
\]

\[
SS_1/(r - 1) \text{ variance explained by the factor considered}
\]

\[
SS_u/(rc - r) \text{ unexplained variance}
\]

- the ratio between the explained variance and the unexplained one

\[
F = \frac{SS_1 \cdot r(c - 1)}{SS_u \cdot r - 1} = \frac{r(c - 1)}{r - 1} \cdot \frac{SS_1}{SS_u}
\]

follows a Fisher distribution with \((r - 1, rc - r)\) d.o.f.
Coming back to the model, if the null hypothesis \( H_0 \) is satisfied

\[
H_0 : \mu_i = \mu \quad \forall i = 1, \ldots, r
\]

- the ratio \( F \) follows a Fisher distribution with \((r - 1, rc - r)\) d.o.f.
- the explained variance \( SS_1 / (r - 1) \) is very close to zero
- the unexplained variance — \( SS_u / r(c - 1) \) — can be assumed of the order of \( \sigma^2 \)

In contrast, if \( H_0 \) is false (so that at least one of the means \( \mu_i \) turns out to be different from the others) we expect that

\[
\frac{1}{r - 1} SS_1 \gtrsim \sigma^2
\]

whereas the unexplained variance is however of the order of \( \sigma^2 \)

\[ \Downarrow \]

\( F \) is taken as the test variable

and the rejection region of \( H_0 \) is of the form

\[
F \geq F_{\text{critical}} = F_{[1-\alpha](r-1,r(c-1))}
\]

with significance level \( \alpha \)

(\( H_0 \) is rejected whenever \( F \) turns out to be large enough)
**Remark**

In the conditions assumed for the model — same variance $\sigma^2$ and means $\mu_i$ for the various samples — it is often important to determine the confidence intervals of the mean differences.

We can apply the following result

**Scheffé theorem**

For any $r$-uple of real coefficients $(C_1, C_2, \ldots, C_r)$ with zero mean

$$\sum_{i=1}^{r} C_i = 0$$

and for any given $\alpha \in (0, 1)$, the inequality

$$\left\| \sum_{i=1}^{r} C_i (\bar{x}_i - \mu_i) \right\|^2 \leq F_{[\alpha]}(r-1, rc-r) \frac{SS_u}{r(c-1)} \frac{r - 1}{c} \sum_{k=1}^{r} C_k^2$$

is verified with probability $\alpha$.

We deduce the confidence interval, with confidence level $\alpha$, of the linear combination $\sum_{i=1}^{r} C_i \mu_i$

$$\sum_{i=1}^{r} C_i \bar{x}_i. \pm \sqrt{F_{[\alpha]}(r-1, rc-r)} \sqrt{\frac{SS_u}{r(c-1)}} \sqrt{\frac{r - 1}{c} \sum_{k=1}^{r} C_k^2}$$

In particular, for the difference of any two means $\mu_i$ and $\mu_h$ the confidence interval is

$$\bar{x}_i. - \bar{x}_h. \pm \sqrt{F_{[\alpha]}(r-1, rc-r)} \sqrt{\frac{SS_u}{r(c-1)}} \sqrt{2 \frac{r - 1}{c}}$$

again with confidence level $\alpha$. 

---

Stefano Siboni
5.9.2 2-factor ANOVA

The set of random variables with two indices

\[ x_{ij} \quad i = 1, \ldots, r \quad j = 1, \ldots, c \]

describes the measured values of a quantity \( x \) concerning a material treated by a certain chemical reagent at a given temperature

- \( i \) identifies the treating chemical reagent
- \( j \) specifies the treatment temperature

**Question**

Have the chemical reagent or the treatment temperature a significant effect on the property \( x \) of the material?

**Model**

For any \( i = 1, \ldots, r \) and \( j = 1, \ldots, c \) the random variable \( x_{ij} \) is normal with mean \( \mu_{ij} \) and variance \( \sigma^2 \)

**Remark**

The mean of the variable \( x_{ij} \) may depend on both the indices \( i \) and \( j \) (on both the chemical reagent and the temperature), while the variance \( \sigma^2 \) is the same

**Hypothesis to be tested**

\[ H_0 : \mu_{ij} = \mu \forall i = 1, \ldots, r, \ j = 1, \ldots, c \] (equal means)

\[ H_1 : \mu_{ij} \neq \mu_{hj} \text{ for at least one } j \text{ and a pair } i, h \ (i \neq h) \]

or

\[ \mu_{ij} \neq \mu_{ik} \text{ for at least one } i \text{ and a pair } j, k \ (j \neq k) \]
Definitions

Mean of data

$$\bar{x} = \frac{1}{rc} \sum_{i=1}^{r} \sum_{j=1}^{c} x_{ij} = \frac{1}{rc} \sum_{i,j} x_{ij}$$

Mean over the index $j$

$$\bar{x}_{i.} = \frac{1}{c} \sum_{j=1}^{c} x_{ij}$$

Mean over the index $i$

$$\bar{x}_{.j} = \frac{1}{r} \sum_{i=1}^{r} x_{ij}$$

Residual

$$x_{ij} - \bar{x}$$

Residual “explained ” by the factor related to index $i$

$$\bar{x}_{i.} - \bar{x}$$

Residual “explained” by the factor related to index $j$

$$\bar{x}_{.j} - \bar{x}$$

“Unexplained” residual

$$x_{ij} - \bar{x} - (\bar{x}_{i.} - \bar{x}) - (\bar{x}_{.j} - \bar{x}) = x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}$$
Total variation

\[ SS_t = \sum_{i,j} (x_{ij} - \bar{x})^2 \]

Variation explained by the factor related to index \( i \)

\[ SS_1 = \sum_{i,j} (\bar{x}_i - \bar{x})^2 = c \sum_{i=1}^{r} (\bar{x}_i - \bar{x})^2 \]

Variation explained by the factor related to index \( j \)

\[ SS_2 = \sum_{i,j} (\bar{x}_j - \bar{x})^2 = r \sum_{j=1}^{c} (\bar{x}_j - \bar{x})^2 \]

Unexplained variation

\[ SS_u = \sum_{i,j} (x_{ij} - \bar{x}_i - \bar{x}_j + \bar{x})^2 \]
5.9.3 2-factor ANOVA without interaction

Model:

\[ x_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \quad i = 1, \ldots, r \quad j = 1, \ldots, c \]

with:

\[ \sum_i \alpha_i = 0 \quad \sum_j \beta_j = 0 \]

and

\[ \varepsilon_{ij} \text{ i.i.d. random variables } N(0, \sigma) \forall i, j \]

\( \sigma \) is assumed to be the same for all the variables!

The algebraic identity

\[
\sum_{i,j} (x_{ij} - \bar{x})^2 \quad \text{total variation } SS_t
\]

\[ = \sum_{i,j} (\bar{x}_i - \bar{x})^2 \quad \text{variation explained by the 1st factor, } SS_1 \]

\[ + \sum_{i,j} (\bar{x}_j - \bar{x})^2 \quad \text{variation explained by the 2nd factor, } SS_2 \]

\[ + \sum_{i,j} (x_{ij} - \bar{x}_j - \bar{x}_i + \bar{x})^2 \quad \text{unexplained variation } SS_u \]

is independent on the correctness of the model
If the model is correct we also have

\[
\frac{1}{\sigma^2} \sum_{i,j} (x_{ij} - \bar{x} - \alpha_i - \beta_j)^2 = \\
= \left\langle \frac{1}{\sigma} (\varepsilon_{ij}) \left| (I - P_1 P_2) \frac{1}{\sigma} (\varepsilon_{ij}) \right. \right\rangle = \chi^2_{rc-1}
\]

\[
\frac{1}{\sigma^2} \sum_{i,j} (\bar{x}_i - \bar{x} - \alpha_i)^2 = \\
= \left\langle \frac{1}{\sigma} (\varepsilon_{ij}) \left| P_2 (I - P_1) \frac{1}{\sigma} (\varepsilon_{ij}) \right. \right\rangle = \chi^2_{r-1}
\]

\[
\frac{1}{\sigma^2} \sum_{i,j} (\bar{x}_j - \bar{x} - \beta_j)^2 = \\
= \left\langle \frac{1}{\sigma} (\varepsilon_{ij}) \left| (I - P_2) P_1 \frac{1}{\sigma} (\varepsilon_{ij}) \right. \right\rangle = \chi^2_{c-1}
\]

\[
\frac{1}{\sigma^2} \sum_{i,j} (x_{ij} - \bar{x}_i - \bar{x}_j + \bar{x})^2 = \\
= \left\langle \frac{1}{\sigma} (\varepsilon_{ij}) \left| (I - P_2) (I - P_1) \frac{1}{\sigma} (\varepsilon_{ij}) \right. \right\rangle = \chi^2_{(r-1)(c-1)}
\]

with

\[
\chi^2_{r-1} \quad \chi^2_{c-1} \quad \chi^2_{(r-1)(c-1)}
\]

stochastically independent
Fundamental theorem

For any set $x_{ij}$ the total variation coincides with the sum of

- the variation explained by the factor related to $i$
- the variation explained by the factor associated to $j$
- the unexplained variation

$$SS_t = SS_1 + SS_2 + SS_u$$

If moreover the variables $x_{ij}$ are i.i.d. normal of mean $\mu$ and variance $\sigma^2$, we have that

$$SS_t/\sigma^2$$ is a $\chi^2$ variable with $rc - 1$ d.o.f.

$$SS_1/\sigma^2$$ is a $\chi^2$ variable with $r - 1$ d.o.f.

$$SS_2/\sigma^2$$ is a $\chi^2$ variable with $c - 1$ d.o.f.

$$SS_u/\sigma^2$$ is a $\chi^2$ variable with $(r - 1)(c - 1)$ d.o.f.

$$SS_1/\sigma^2, SS_2/\sigma^2$$ and $$SS_u/\sigma^2$$ are stochastically independent
Remarks

• the number of d.o.f. is balanced

\[ rc - 1 = (r - 1) + (c - 1) + (r - 1)(c - 1) \]

• we define the variances

\[ SS_t/(rc - 1) \] total variance

\[ SS_1/(r - 1) \] variance explained by the factor related to \( i \)

\[ SS_2/(c - 1) \] variance explained by the factor related to \( j \)

\[ SS_u/(r - 1)(c - 1) \] unexplained variance

• the ratio of the variance explained by the factor \( i \) over the unexplained one

\[ F = \frac{SS_1}{r - 1} \frac{(r - 1)(c - 1)}{SS_u} = (c - 1) \frac{SS_1}{SS_u} \]

follows a Fisher distribution with \((r - 1, (r - 1)(c - 1))\) d.o.f.

• the ratio of the variance explained by the factor \( j \) over the unexplained one

\[ F = \frac{SS_2}{c - 1} \frac{(r - 1)(c - 1)}{SS_u} = (r - 1) \frac{SS_2}{SS_u} \]

is a Fisher random variable with \((c - 1, (r - 1)(c - 1))\) d.o.f.
Test for the difference of the mean over the index $i$

Test variable

$$F = \frac{SS_1}{r-1} \frac{(r-1)(c-1)}{SS_u} = (c-1) \frac{SS_1}{SS_u}$$

If the mean $\mu_{ij}$ do not depend appreciably on the index $i$, there holds

$$\frac{SS_1}{r-1} \ll \frac{SS_u}{(r-1)(c-1)} \approx \sigma^2 \iff F \text{ small}$$

whereas in the opposite case we expect that

$$\frac{SS_1}{r-1} \gtrsim \frac{SS_u}{(r-1)(c-1)} \approx \sigma^2 \iff F \text{ large}$$

We define then a critical value of $F$

$$F_{\text{critical}} = F_{[1-\alpha](r-1,(r-1)(c-1))}$$

and an acceptance region of $H_0$ — means independent on the index $i$ —

$$F \leq F_{[1-\alpha](r-1,(r-1)(c-1))}$$

along with a corresponding rejection region — means $\mu_{ij}$ appreciably dependent on the index $i$ —

$$F > F_{[1-\alpha](r-1,(r-1)(c-1))}$$

both with given significance level $\alpha \in (0,1)$
Test for the difference of the means over the index $j$

Test variable

$$F = \frac{SS_2}{c-1} \cdot \frac{(r-1)(c-1)}{SS_u} = (r-1) \frac{SS_2}{SS_u}$$

If the means $\mu_{ij}$ do not depend appreciably on the index $j$ we have

$$\frac{SS_2}{c-1} \ll \frac{SS_u}{(r-1)(c-1)} \simeq \sigma^2 \iff F \text{ small}$$

while in the opposite case it seems reasonable to assume that

$$\frac{SS_2}{c-1} \gtrsim \frac{SS_u}{(r-1)(c-1)} \simeq \sigma^2 \iff F \text{ large}$$

We define then a critical value of $F$

$$F_{\text{critical}} = F_{[1-\alpha](c-1, (r-1)(c-1))}$$

and an acceptante region of $H_0$ — means independent on the index $j$ —

$$F \leq F_{[1-\alpha](c-1, (r-1)(c-1))}$$

along with a corresponding rejection region — means $\mu_{ij}$ appreciably dependent on the index $j$ —

$$F > F_{[1-\alpha](c-1, (r-1)(c-1))}$$

both with given significance level $\alpha \in (0, 1)$
5.9.4 2-factor ANOVA with interaction

Model:

\[ x_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk} \]

\[ i = 1, \ldots, r \quad j = 1, \ldots, c \quad k = 1, \ldots, s \]

with:

\[ \sum_i \alpha_i = 0 \quad \sum_j \beta_j = 0 \]

\[ \sum_i \alpha\beta_{ij} = 0 \quad \forall j \quad \sum_j \alpha\beta_{ij} = 0 \quad \forall i \]

and

\[ \varepsilon_{ijk} \text{ i.i.d. random variables } N(0, \sigma) \quad \forall i, j, k \]

The algebraic identity

\[
\sum_{ijk} (x_{ijk} - \overline{x})^2 \quad \text{total variation } SS_t
\]

\[
= \sum_{ijk} (\overline{x}_{i..} - \overline{x})^2 \quad \text{var.on explained by the 1st factor, } SS_1
\]

\[
+ \sum_{ijk} (\overline{x}_{.j} - \overline{x})^2 \quad \text{var.on explained by the 2nd factor, } SS_2
\]

\[
+ \sum_{ijk} (\overline{x}_{ij} - \overline{x}_{.j} - \overline{x}_{i..} + \overline{x})^2 \quad \text{var.on explained by the interaction between the factors 1 and 2, } S_{12}
\]

\[
+ \sum_{ijk} (x_{ijk} - \overline{x}_{ij..})^2 \quad \text{unexplained variation } SS_u
\]

holds also when the model is incorrect
Moreover, if the model is correct we have that

\[
\frac{1}{\sigma^2} \sum_{ijk} (x_{ijk} - \bar{x} - \alpha_i - \beta_j - \alpha \beta_{ij})^2 =
\]

\[
= \left\langle \frac{1}{\sigma} (\varepsilon_{ijk}) \right| (I - P_3P_2P_1) \frac{1}{\sigma} (\varepsilon_{ijk}) \right\rangle = \chi^2_{rcs-1}
\]

\[
\frac{1}{\sigma^2} \sum_{ijk} (\bar{x}_{i..} - \bar{x} - \alpha_i)^2 =
\]

\[
= \left\langle \frac{1}{\sigma} (\varepsilon_{ijk}) \right| P_3P_2(I - P_1) \frac{1}{\sigma} (\varepsilon_{ijk}) \right\rangle = \chi^2_{r-1}
\]

\[
\frac{1}{\sigma^2} \sum_{ijk} (\bar{x}_{..j} - \bar{x} - \beta_j)^2 =
\]

\[
= \left\langle \frac{1}{\sigma} (\varepsilon_{ijk}) \right| P_3(I - P_2P_1) \frac{1}{\sigma} (\varepsilon_{ijk}) \right\rangle = \chi^2_{c-1}
\]

\[
\frac{1}{\sigma^2} \sum_{ijk} (\bar{x}_{ij.} - \bar{x}_{.j} - \bar{x}_{i..} + \bar{x} - \alpha \beta_{ij})^2 =
\]

\[
= \left\langle \frac{1}{\sigma} (\varepsilon_{ijk}) \right| P_3(I - P_2)(I - P_1) \frac{1}{\sigma} (\varepsilon_{ijk}) \right\rangle = \chi^2_{(r-1)(c-1)}
\]

\[
\frac{1}{\sigma^2} \sum_{ijk} (x_{ijk} - \bar{x}_{i..})^2 =
\]

\[
= \left\langle \frac{1}{\sigma} (\varepsilon_{ijk}) \right| (I - P_3) \frac{1}{\sigma} (\varepsilon_{ijk}) \right\rangle = \chi^2_{rc(s-1)}
\]

with

\[
\chi^2_{r-1} \quad \chi^2_{c-1} \quad \chi^2_{(r-1)(c-1)} \quad \chi^2_{rc(s-1)}
\]

stochastically independent
5.10 Sign test for the median of a population

**Non-parametric** test: it does not require any hypothesis about the form of the probability distribution of the population.

Let us consider a discrete or continuous random variable $x$ of arbitrary distribution $p(x)$

**Hypothesis to be tested:**

$H_0$: the median of the distribution is $m$

$H_1$: the median of the distribution is different from $m$

If the median of the population holds $m$, then:

\[
\begin{align*}
    x - m &> 0 \text{ with probability } 1/2 \\
    x - m &\leq 0 \text{ with probability } 1/2
\end{align*}
\]

otherwise we will have a larger probability of obtaining positive or negative residuals

**Test variable**

For a set $x_1, x_2, \ldots, x_n$ of outcomes of $x$ we pose

\[
s_i = \begin{cases} 
1 & \text{if } x_i - m > 0 \\
0 & \text{if } x_i - m \leq 0 
\end{cases} \quad \forall \ i = 1, \ldots, n
\]

and introduce as a test variable the number of positive differences with respect to the median

\[
z = \sum_{i=1}^{n} s_i
\]
Rejection region

We are led to reject the hypothesis $H_0$ whenever the number of positive differences with respect to the median turns out to be too large or too small.

The rejection region will be two-sided and symmetrical

$$\{z \leq z_{\text{min}}\} \cup \{z \geq z_{\text{max}}\}$$

If the hypothesis $H_0$ holds true, the test variable $z$ follows a binomial distribution with $p = 1/2$

$$p_n(z) = \binom{n}{z} \frac{1}{2^n} \quad \forall z = 0, 1, \ldots, n$$

For a given significance level $\alpha$ of the test, the bounds $z_{\text{min}}$ and $z_{\text{max}}$ of the rejection region are therefore defined by

$$\sum_{z=0}^{z_{\text{min}}} \binom{n}{z} \frac{1}{2^n} = \frac{\alpha}{2} \quad \sum_{z=z_{\text{max}}}^{n} \binom{n}{z} \frac{1}{2^n} = \frac{\alpha}{2}$$
5.11 Sign test to check if two paired samples belong to the same population

The sign test is very useful to check whether two paired samples are drawn from the same statistical population, i.e. if two paired samples follow the same probability distribution.

The test relies on the remark that for any two independent continuous random variables \( x \) and \( y \) with the same probability distribution \( p(x) \) — or \( p(y) \) — the difference \( RV \)

\[
z = x - y
\]
follows an even probability distribution

\[
p_z(z) = p_z(-z) \quad \forall z \in \mathbb{R}
\]

no matter the common distribution \( p(x) \) is.

We have indeed

\[
1 = \int_{\mathbb{R}} \int_{\mathbb{R}} p(x) p(y) = \int_{\mathbb{R}} \int_{\mathbb{R}} dz \int_{\mathbb{R}} dy p(y + z) p(y) = \\
= \int_{\mathbb{R}} dz \int_{\mathbb{R}} dy p(y + z) p(y)
\]

so that the probability distribution of \( z \) is

\[
p_z(z) = \int_{\mathbb{R}} p(y + z) p(y) dy \quad \forall z \in \mathbb{R}.
\]
Consequently,
\[ p_z(-z) = \int_R p(y - z) p(y) \, dy \]
and by the change of variable \( y = \xi + z \) the integral becomes
\[ p_z(-z) = \int_R p(\xi) p(\xi + z) \, d\xi = \int_R p(\xi + z) p(\xi) \, d\xi = p_z(z). \]

Positive and (null or) negative differences have then the same probability to occur:
\[ p(z > 0) = \frac{1}{2} \quad p(z \leq 0) = \frac{1}{2}. \]

Thus we can test the null hypothesis

\( H_0 : \) the two samples are drawn from the same population

versus the alternative hypothesis

\( H_1 : \) the two samples are drawn from unlike populations

by using as a test variable the number \( z \) of the positive differences detected on the various data pairs of the sample

\( H_0 \) will be accepted if \( z \) is close to half the number of data pairs of the sample, while \( H_1 \) will be preferred whenever \( z \) is sufficiently low or too large (two-sided critical region)
The rejection region takes then the two-sided form

\[ \{ z \leq z_{\min} \} \cup \{ z \geq z_{\max} \} \]

If \( H_0 \) holds true, the variable \( z \) obeys a binomial distribution with \( p = 1/2 \) and the probability that \( z \) takes a value within the critical region must be identified with the probability \( \alpha \) of making a type I error.

At the significance level \( \alpha \), the limits \( z_{\min} \) and \( z_{\max} \) of the critical region are thus defined by

\[
\sum_{z=0}^{z_{\min}} \binom{n}{z} \frac{1}{2^n} = \frac{\alpha}{2} \quad \sum_{z=z_{\max}}^{n} \binom{n}{z} \frac{1}{2^n} = \frac{\alpha}{2}
\]

**Example: Sign test**

A company states that by adding an additive of own manufacture to the reservoir of a gasoline car the efficiency of the engine increases. To check this statement 16 cars are chosen and the mean distance that they cover by a litre of fuel, with and without additive, is measured (in Km).

<table>
<thead>
<tr>
<th>additive</th>
<th>no additive</th>
<th>additive</th>
<th>no additive</th>
</tr>
</thead>
<tbody>
<tr>
<td>17.35</td>
<td>15.70</td>
<td>13.65</td>
<td>13.10</td>
</tr>
<tr>
<td>14.15</td>
<td>13.60</td>
<td>16.05</td>
<td>15.70</td>
</tr>
<tr>
<td>9.80</td>
<td>10.20</td>
<td>14.80</td>
<td>14.40</td>
</tr>
<tr>
<td>12.55</td>
<td>12.30</td>
<td>11.20</td>
<td>11.55</td>
</tr>
<tr>
<td>7.85</td>
<td>7.45</td>
<td>12.65</td>
<td>12.00</td>
</tr>
<tr>
<td>12.25</td>
<td>11.15</td>
<td>14.05</td>
<td>13.65</td>
</tr>
<tr>
<td>14.35</td>
<td>13.40</td>
<td>12.15</td>
<td>11.45</td>
</tr>
<tr>
<td>11.75</td>
<td>12.05</td>
<td>11.95</td>
<td>12.40</td>
</tr>
</tbody>
</table>
Assuming that the drive conditions are the same, check if there is any difference in the performances of the engines due to the use of the additive, with a significance level of 5% and 1%.

**Solution**

We must test the null hypothesis $H_0$ that no difference of performances occurs between the cars fed with additivated fuel and the same vehicles filled up with normal gasoline, versus the alternative hypothesis $H_1$ that the cars filled up with additivated fuel have better performances.

The hypothesis $H_0$ is rejected in favour of $H_1$ only if the number of positive differences is sufficiently large: the test is one-sided.

The distances per litre covered by the additivated and the non-additivated cars show **12 positive differences** and **4 negative differences**.

If $H_0$ is correct, the probability of having at least 12 positive signs out of 16 differences writes

\[
\sum_{z=12}^{16} \binom{16}{z} \frac{1}{2^{16}} = \\
= \left[ \binom{16}{12} + \binom{16}{13} + \binom{16}{14} + \binom{16}{15} + \binom{16}{16} \right] \frac{1}{2^{16}} = 0.0384
\]

so that the observed number 12 of positive signs certainly belongs to the rejection region of $H_0$ if the significance level is taken at 5%. In contrast, if $\alpha = 1\%$ the number $z = 12$ of positive differences does not fall within the critical region of $H_0$, which then cannot be rejected.
5.12 Detection of outliers by Chauvenet criterion

Let a small sample of data $x_1, x_2, \ldots x_n$ be given

Assume that data are drawn from a normal population

Suppose that a datum $x_k$ is very far from the sample mean $\bar{x}$

**Problem:** does $x_k$ actually belong to the normal population, or is it an “intruder”?

**A possible strategy (Chauvenet criterion):**

Let us pose

$$z = \frac{|x_k - \bar{x}|}{s} \approx \frac{|x_k - \mu|}{\sigma}$$

where $\mu$ and $\sigma$ are replaced with the sample estimate $\bar{x}$ and $s$ respectively (*with fingers crossed!*)

$z$ is the distance of $x_k$ from the sample mean $\bar{x}$ in units of $s$

or, **approximately**, from the mean $\mu$ in units of $\sigma$
The probability $p(z)$ that a normal datum $x_k$ is at a distance $\geq z$ from the mean, in units of $\sigma$, is represented by the shaded area in the following graph of the $N(\mu, \sigma)$ distribution.

\[
p(z) = 1 - \int_{\mu - z\sigma}^{\mu + z\sigma} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} dx
\]

or, after some change of variables, in the equivalent forms

\[
p(z) = 1 - \frac{2}{\sqrt{2\pi}} \int_0^\infty e^{-\xi^2/2} d\xi = 1 - \text{erf} \left( \frac{z}{\sqrt{2}} \right)
\]

where erf denotes the **error function**:

\[
\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt
\]
The mean number of data at a distance $\geq z$ from $\mu$ in $\sigma$ units out of $n$ measurements is
\[
np(z)
\]

Let us choose $z$ so that the mean number of data as before is significantly smaller than 1. Typically we require that:
\[
np(z) = \frac{1}{2}
\]
i.e.
\[
1 - \text{erf}\left(\frac{z}{\sqrt{2}}\right) = \frac{1}{2n}
\tag{1}
\]

With the previous $z$, if
\[
\frac{|x_k - \bar{x}|}{s} \geq z
\]
it is quite unlikely that $x_k$ belongs to the normal population

In such a case, the datum should be rejected as meaningless
The graph shows the critical value $z$ versus the number $n$ of data
The following table collects the critical values $z$ versus the number $n$ of data, according to the previous equation (1):

<table>
<thead>
<tr>
<th>$n$</th>
<th>critical $z$</th>
<th>$n$</th>
<th>critical $z$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.674489750</td>
<td>27</td>
<td>2.355084009</td>
</tr>
<tr>
<td>2</td>
<td>1.150349380</td>
<td>28</td>
<td>2.368567059</td>
</tr>
<tr>
<td>3</td>
<td>1.382994127</td>
<td>29</td>
<td>2.381519470</td>
</tr>
<tr>
<td>4</td>
<td>1.534120544</td>
<td>30</td>
<td>2.393979800</td>
</tr>
<tr>
<td>5</td>
<td>1.644853627</td>
<td>40</td>
<td>2.497705474</td>
</tr>
<tr>
<td>6</td>
<td>1.731664396</td>
<td>50</td>
<td>2.575829304</td>
</tr>
<tr>
<td>7</td>
<td>1.802743091</td>
<td>60</td>
<td>2.638257273</td>
</tr>
<tr>
<td>8</td>
<td>1.862731867</td>
<td>70</td>
<td>2.690109527</td>
</tr>
<tr>
<td>9</td>
<td>1.914505825</td>
<td>80</td>
<td>2.734368787</td>
</tr>
<tr>
<td>10</td>
<td>1.959963985</td>
<td>90</td>
<td>2.772921295</td>
</tr>
<tr>
<td>11</td>
<td>2.000423569</td>
<td>100</td>
<td>2.807033768</td>
</tr>
<tr>
<td>12</td>
<td>2.036834132</td>
<td>150</td>
<td>2.935199469</td>
</tr>
<tr>
<td>13</td>
<td>2.069901831</td>
<td>200</td>
<td>3.023341440</td>
</tr>
<tr>
<td>14</td>
<td>2.100165493</td>
<td>250</td>
<td>3.090232306</td>
</tr>
<tr>
<td>15</td>
<td>2.128045234</td>
<td>300</td>
<td>3.143980287</td>
</tr>
<tr>
<td>16</td>
<td>2.153874694</td>
<td>350</td>
<td>3.188815259</td>
</tr>
<tr>
<td>17</td>
<td>2.177923069</td>
<td>400</td>
<td>3.227218426</td>
</tr>
<tr>
<td>18</td>
<td>2.200410581</td>
<td>450</td>
<td>3.260767488</td>
</tr>
<tr>
<td>19</td>
<td>2.221519588</td>
<td>500</td>
<td>3.290526731</td>
</tr>
<tr>
<td>20</td>
<td>2.241402728</td>
<td>550</td>
<td>3.317247362</td>
</tr>
<tr>
<td>21</td>
<td>2.260188991</td>
<td>600</td>
<td>3.34178956</td>
</tr>
<tr>
<td>22</td>
<td>2.277988333</td>
<td>650</td>
<td>3.363635456</td>
</tr>
<tr>
<td>23</td>
<td>2.294895209</td>
<td>700</td>
<td>3.384036252</td>
</tr>
<tr>
<td>24</td>
<td>2.310991338</td>
<td>800</td>
<td>3.420526701</td>
</tr>
<tr>
<td>25</td>
<td>2.326347874</td>
<td>900</td>
<td>3.452432937</td>
</tr>
<tr>
<td>26</td>
<td>2.341027138</td>
<td>1000</td>
<td>3.480756404</td>
</tr>
</tbody>
</table>
Example: Chauvenet criterion

Repeated measurements of a length \( x \) have provided the following results (expressed in mm):

\[
46, \ 48, \ 44, \ 38, \ 45, \ 47, \ 58, \ 44, \ 45, \ 43
\]

and can be assumed to be drawn from a normal population. By applying Chauvenet criterion, we want to check whether the anomalous result (outlier) \( x_{\text{sus}} = 58 \) must be rejected.

Solution

The sample mean and standard deviation are given by

\[
\bar{x} = \frac{1}{10} \sum_{i=1}^{10} x_i = 45.8 \quad s = \sqrt{\frac{1}{9} \sum_{i=1}^{10} (x_i - \bar{x})^2} = 5.1
\]

The distance of the suspect value from the mean, in units of \( s \), holds

\[
\frac{x_{\text{sus}} - \bar{x}}{s} = \frac{58 - 45.8}{5.1} = 2.4
\]

The probability that a measurement falls at a distance larger than 2.4 standard deviations from the mean can be calculated from the Table of the cumulative distribution function of the standard normal distribution:

\[
P(|x_{\text{sus}} - \bar{x}| \geq 2.4s) = 1 - P(|x_{\text{sus}} - \bar{x}| < 2.4s) =
\]

\[
= 1 - 2 \cdot P(\bar{x} \leq x_{\text{sus}} < \bar{x} + 2.4s) =
\]

\[
= 1 - 2 \cdot 0.49180 = 0.0164
\]

Out of 10 measurements we typically expect \( 10 \cdot 0.0164 = 0.164 \) “bad” results, at a distance larger than 2.4\( s \) from the mean. \textbf{Since 0.164 < 1/2 Chauvenet criterion suggests that the outlier} \( x_{\text{sus}} \) \textbf{should be rejected.}
6. PAIRS OF RANDOM VARIABLES

6.1 Linear correlation coefficient (Pearson r)

Let \((x, y)\) be a pair of random variables

Problem: are they stochastically dependent or independent?

The limit case of stochastic dependence is that where
one variable is a function of the other
\[ y = f(x) \]

The most interesting case is the linear one, when we assume
\[ y = ax + b, \text{ with given constants } a \text{ and } b \ (a \neq 0) \]

The joint probability distribution of \((x, y)\)
can be formally written as
\[ p(x, y) = p(x) \delta(ax + b - y) \]
in terms of the generalized function Dirac delta \(\delta\)
Means, variances and covariance of \((x, y)\) take then the form:

\[
\mu_x = \int_{\mathbb{R}^2} p(x) \delta(ax + b - y)x \, dy \, dx = \int_{\mathbb{R}} p(x)x \, dx
\]

\[
\mu_y = \int_{\mathbb{R}^2} p(x) \delta(ax + b - y)y \, dy \, dx = a\mu_x + b
\]

\[
\text{var}(x) = \int_{\mathbb{R}^2} p(x) \delta(ax + b - y)(x - \mu_x)^2 \, dy \, dx = \int_{\mathbb{R}} p(x)(x - \mu_x)^2 \, dx
\]

\[
= \int_{\mathbb{R}} p(x)(ax - a\mu_x)^2 \, dx = a^2 \int_{\mathbb{R}} p(x)(x - \mu_x)^2 \, dx
\]

\[
\text{cov}(x, y) = \int_{\mathbb{R}^2} p(x) \delta(ax + b - y)(x - \mu_x)(y - \mu_y) \, dy \, dx = \int_{\mathbb{R}} p(x)(x - \mu_x)(ax - a\mu_x) \, dx
\]

so that the correlation becomes

\[
\text{corr}(x, y) = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x)} \sqrt{\text{var}(y)}} = \frac{a}{|a|} = \begin{cases} +1 & \text{if } a > 0 \\ -1 & \text{if } a < 0 \end{cases}
\]

If \((x, y)\) are independent we have instead \(\text{corr}(x, y) = 0\)
Whenever $y$ is a **nonlinear** function $f(x)$ of the RV $x$, the correlation between $x$ and $y$

$$
corr(x, y) = \frac{\int_{\mathbb{R}} p(x)(x - \mu_x)(f(x) - \mu_y)dx}{\sqrt{\int_{\mathbb{R}} p(x)(x - \mu_x)^2dx \int_{\mathbb{R}} p(x)[f(x) - \mu_y]^2dx}}
$$

with

$$
\mu_y = \int_{\mathbb{R}} p(x)f(x)dx
$$

has no particular meaning

Nevertheless, the values of $corr(x, y)$ can be considered as an indication of:

- **direct linear dependence** ($a > 0$), if close to $+1$;

- **inverse linear dependence** ($a < 0$), when close to $-1$;

- **stochastic independence**, if close to 0
Let \((x_i, y_i) \in \mathbb{R}^2, i = 1, \ldots, n\) be a sample of the pair of random variables \((x, y)\), with sample means \(\bar{x}\) and \(\bar{y}\) respectively.

The correlation is then estimated by using the sample, by means of the (Pearson) **linear correlation coefficient**

\[
r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}
\]

Actually, from Cauchy-Schwarz inequality

\[
\sum_{i=1}^{n} a_i b_i \leq \sqrt{\sum_{i=1}^{n} a_i^2} \sqrt{\sum_{i=1}^{n} b_i^2} \quad \forall a_i, b_i \in \mathbb{R} \quad i = 1, \ldots, n
\]

it follows that the condition \(r = +1\) or \(r = -1\) is equivalent to have all the pairs \((x_i, y_i)\) of the sample arranged along the same straight line, of positive or negative slope respectively.
6.2 Statistics of the linear correlation coefficient
Once the stochastic dependence of the random variables \((x, y)\) has been established, the linear correlation coefficient expresses conventionally the “intensity” of such a dependence

But, generally speaking, the coefficient \(r\) is not a good indicator to test the stochastic dependence or independence of the variables \((x, y)\)

Indeed, even if the null hypothesis

\[H_0: \text{the variables } x \text{ and } y \text{ are independent}\]

holds true, the probability distribution of \(r\) is not determined, since those of \(x\) and \(y\) are unknown

\[\Downarrow\]

It is impossible to devise a test in order to check the hypothesis \(H_0\), with an appropriate significance level

It is necessary to refer to some noticeable cases, for which the distribution of \(r\) can be at least partially determined
6.2.1 Independent random variables \( (x, y) \)

For a pair of independent random variables \((x, y)\), whose probability distributions have a sufficiently large number of convergent moments, if \(n\) is large enough \((n > 20)\)

\[
\downarrow
\]

\(r\) follows approximately a normal distribution with zero average and variance \(1/n\)

\[
\downarrow
\]

The rejection region of \(H_0\) will be the union of intervals

\[
\{|r| \geq \eta\} = \{r < -\eta\} \cup \{r > \eta\}
\]

where \(\eta > 0\) is determined by the equation

\[
\frac{\alpha}{2} = \sqrt{\frac{n}{2\pi}} \int_{\eta}^{+\infty} e^{-z^2n/2}dz = \frac{1}{2} \text{erfc} \left( \eta \sqrt{\frac{n}{2}} \right)
\]

and \(\text{erfc}\) denotes the complementary error function

\[
\text{erfc}(x) = 1 - \text{erf}(x) = 1 - \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-\xi^2} d\xi = \frac{2}{\sqrt{\pi}} \int_{x}^{\infty} e^{-\xi^2} d\xi
\]

while \(\alpha\) is the probability of a type I error
6.2.2 Normal random variables \((x, y)\)

For normal variables \((x, y)\), with distribution

\[
p(x, y) = \frac{1}{2\pi} \sqrt{a_{11}a_{22} - a_{12}^2} e^{-\frac{1}{2}(a_{11}x^2 - 2a_{12}xy + a_{22}y^2)}
\]

\[
a_{11}a_{22} - a_{12}^2 > 0 \quad a_{11} + a_{22} > 0
\]

and hence correlation

\[
\rho = \text{corr}(x, y) = \frac{a_{12}}{\sqrt{a_{11}} \sqrt{a_{22}}} \in (-1, 1)
\]

the theoretical frequency distribution of \(r\) depends on \(\rho\) and on the number \(n\) of datapoints.

Defined for \(r \in (-1, 1)\), such a distribution can be written:

\[
p_{\rho,n}(r) = \frac{2^{n-3}}{\pi(n-3)!} (1 - \rho^2)^{(n-1)/2} (1 - r^2)^{(n-4)/2} a_{\rho,n}(r)
\]

with

\[
a_{\rho,n}(r) = \sum_{s=0}^{\infty} \Gamma^2 \left( \frac{n+s-1}{2} \right) \frac{(2\rho r)^s}{s!}
\]

The distribution is just too complex to be used directly.

Typically an approximation is used which holds true for large samples \((n \geq 10)\)

Or, alternatively, we refer to the case \(\rho = 0\) of uncorrelated \(\iff\) independent variables (what we are usually interested in)
6.2.2.1 Approximation for large $n$ of normal variables

**Fisher transformation**

For $n$ large enough ($n \geq 10$), the Fisher $z$ variable:

$$ z = \text{arctanh } r = \frac{1}{2} \ln \frac{1+r}{1-r} $$

follows approximately a normal distribution of mean

$$ \bar{z} = \frac{1}{2} \left[ \ln \frac{1+\rho}{1-\rho} + \frac{\rho}{n-1} \right] $$

and standard deviation

$$ \sigma(z) \simeq \frac{1}{\sqrt{n-3}} $$

The Fisher transformation is implemented in Excel by the function FISHER.

6.2.2.2 Independent (or uncorrelated) normal variables

For $\rho = 0$ the probability distribution of $r$ reduces to:

$$ p_{0,n}(r) = \frac{\Gamma[(n-1)/2]}{\sqrt{\pi n} \Gamma[(n-2)/2]} (1 - r^2)^{(n-4)/2} $$

in such a way that the random variable:

$$ t = \sqrt{n-2} \frac{r}{\sqrt{1-r^2}} $$

follows a Student’s distribution with $n - 2$ d.o.f.
In the case of normal variables \((x, y)\) it is then possible to test the hypothesis

\[ H_0 : x \text{ and } y \text{ are independent variables} \]

which will be regarded as acceptable when the test variable

\[ t = \sqrt{n-2} \frac{r}{\sqrt{1-r^2}} \]

is close enough to 0 (and also \(r\) almost vanishes)

The null hypothesis \(H_0\) will be rejected if in contrast \(t\) (and thus \(r\)) is significantly different from zero

We deduce a typical two-sided test, with a rejection region defined by

\[ \{t \leq t_{\frac{\alpha}{2}}(n-2)\} \cup \{t \geq t_{1-\frac{\alpha}{2}}(n-2)\} \]

i.e.

\[ \{t \leq -t_{1-\frac{\alpha}{2}}(n-2)\} \cup \{t \geq t_{1-\frac{\alpha}{2}}(n-2)\} \]

with significance level \(\alpha\)
Example: Linear correlation coefficient (Pearson r)
Let us consider the following table of data, concerning some repeated measurements of two quantities \(x\) and \(y\) (in arbitrary units), gathered in the same experimental conditions. The sample is assumed to be normal.

\[
\begin{array}{|c|c|c|}
\hline
i & x_i & y_i \\
\hline
1 & 1.8 & 1.1 \\
2 & 3.2 & 1.6 \\
3 & 4.3 & 1.1 \\
4 & 5.9 & 3.5 \\
5 & 8.1 & 3.8 \\
6 & 9.9 & 6.2 \\
7 & 11.4 & 5.6 \\
8 & 12.1 & 4.0 \\
9 & 13.5 & 7.5 \\
10 & 17.9 & 7.1 \\
\hline
\end{array}
\]

Use Pearson’s linear correlation coefficient to check whether the quantities \(x\) and \(y\) can be considered stochastically independent, with a significance level of 5% and 1%.

Solution
The sample means of the quantities are given by

\[
\bar{x} = \frac{1}{10} \sum_{i=1}^{10} x_i = 8.8100 \quad \bar{y} = \frac{1}{10} \sum_{i=1}^{10} y_i = 4.1500
\]

and allow us to reckon the following sums of residual products

\[
SS_{xy} = \sum_{i=1}^{10} (x_i - \bar{x})(y_i - \bar{y}) = 99.6050
\]
and squares

\[ SS_{xx} = \sum_{i=1}^{10} (x_i - \overline{x})^2 = 233.2690 \]

\[ SS_{yy} = \sum_{i=1}^{10} (y_i - \overline{y})^2 = 51.9050 \]

so that the linear correlation coefficient becomes

\[
r = \frac{SS_{xy}}{\sqrt{SS_{xx}} \sqrt{SS_{yy}}} = \frac{99.6050}{\sqrt{233.2690} \sqrt{51.9050}} = 0.9052
\]

Alternatively, the linear correlation coefficient can be calculated by using the functions CORREL or PEARSON of Excel.

The plot of the data suggests that \( x \) and \( y \) are dependent:
As $x$ and $y$ can be assumed to be normal, we can test the null hypothesis

$$H_0 : x \text{ and } y \text{ are stocastically independent}$$

versus the alternative hypothesis

$$H_1 : x \text{ and } y \text{ are stochastically dependent}$$

by means of the RV

$$t = \sqrt{n - 2} \frac{r}{\sqrt{1 - r^2}}$$

which, for $H_0$ true, follows a Student distribution with $n - 2$ d.o.f. In the present case we get

$$t = \sqrt{10 - 2} \frac{0.9052}{\sqrt{1 - 0.9052^2}} = 6.0247$$

The critical region takes the form

$$|t| > t_{[1-\frac{\alpha}{2}] (n-2)} = t_{[0.975]}(8) = 2.306$$

for a significance level $\alpha = 5\%$, and the form

$$|t| > t_{[1-\frac{\alpha}{2}] (n-2)} = t_{[0.995]}(8) = 3.355$$

when the required significance level is $\alpha = 1\%$.

In both cases $H_0$ must be rejected!
6.3 Remark

We have considered the circumstance that the linear correlation coefficient $r$ is calculated for a small sample of two random variables $(x, y)$

In many cases, however, the coefficient is calculated in a completely different situation, when the pairs $(x_i, y_i)$:

- cannot be regarded as individuals drawn from the same statistical population
- can be grouped into some subsets, each of which belonging to a proper statistical population

Example: experiments directed to detect the variations of a quantity $y$ due to the variations of another quantity $x$

The data $(x_i, y_i)$ cannot be regarded as the outcomes of one pair of random variables

\[ r \]

$r$ does not follow a known probability distribution and cannot be applied to test the stochastic dependence or independence of the quantities $x$ and $y$

$r$ provides an indication on the degree of alignment of data $(x_i, y_i)$, but it offers no statistical information about the significance of the suggested linear dependence

Such a significance must be verified in a different way: problem of the linear regression and of the goodness of fit
7. DATA MODELING

7.1 General setup of the problem

Measurements
\[ \downarrow \]
Data of the form \((x_i, y_i), \ i = 1, \ldots, n\)
\[ \downarrow \]
Synthesis of data into a “model”

“Model” is a mathematical function
which relates the observed quantities among themselves
and depends on suitable parameters to be “adjusted”

The procedure to adjust the value of the parameters is known as
“fitting”

As a rule, the fitting procedure consists:

(i) in the definition of a suitable **objective function** (or merit
function), which describes the agreement between data and
model for a given choice of the parameters;

(ii) in the subsequent calculation of the model parameters by
the **optimization** of the merit function (best fit)
We assume that

\[
\text{good agreement} \quad \iff \quad \text{small values of the merit function}
\]

so that

\[
\text{best fit} \quad \iff \quad \text{search for the minimum of the merit function}
\]

It is not, however, a simple optimization problem (search for the best-fit values of the parameters):

- **data** are typically affected by errors and must be regarded as the outcomes of appropriate RVs;

- the **best-fit parameters** of the model are then affected by an error and must be assumed in turn as RVs;

- **errors** and, eventually, the **probability distribution** of the **best-fit parameters** must be determined;

- appropriate **tests** are required in order to check whether the model is or is not adequate to describe data (goodness of fit)
Merit function

The choice of the merit function often relies on the so-called maximum likelihood method

• We assume that the parameters are determined in such a way that the data actually measured constitute an event of maximum probability: the set of measured data is the most probable one, whose occurrence is most likely

• Assumed that the variables \( x_i \) have negligible variance and can be considered almost certain, and that the \( y_i \) are independent normal variables of variance \( \sigma_i^2 \), the maximum likelihood leads to the chi-square method

• Whenever, in the same hypotheses, the variances of the variables \( y_i \) can be considered equal, the maximum likelihood approach reduces to a least-squares fitting

• The maximum likelihood method can be applied also in the case of non-normal variables

  It may happen that the probability distributions of \( y_i \) cannot be considered normal because of outcomes noticeably different from the means, whose occurrence would be almost impossible for normal variables (outliers, outcomes which differ from the mean for more than 3 standard deviations)

In such a case, the merit functions obtained by the maximum likelihood approach characterize the so-called “robust” fitting methods
Maximum Likelihood and Chi-square

Probability of obtaining the pairs of data \((x_i, y_i), i = 1, \ldots, n\)

\[
P(x_1, y_1, \ldots, x_n, y_n) \sim \prod_{i=1}^{n} e^{-\frac{1}{2\sigma_i^2}(y_i - R(x_i))^2} =
\]

\[
= e^{-\frac{1}{2} \sum_{i=1}^{n} \frac{1}{\sigma_i^2}(y_i - R(x_i))^2}
\]

in the assumption of variables \(x_i\) completely certain and \(y_i\) independent normal with mean \(R(x_i)\) and variance \(\sigma_i^2\)

The Maximum Likelihood method consists in requiring that

\[
P(x_1, y_1, \ldots, x_n, y_n) \quad \text{be maximum}
\]

or, equivalently, that

\[
\sum_{i=1}^{n} \frac{1}{\sigma_i^2}(y_i - R(x_i))^2 \quad \text{be minimum}
\]

i.e. that the chi-square of data \(y_i\) with respect to the regression function \(R(x)\) be minimum — **chi-square method**
7.2 Linear regression by the chi-square method (chi-square fitting)

Let

\[ \{(x_i, y_i) \in \mathbb{R}^2, \ 1 \leq i \leq n\} \]

be a set of data satisfying the following conditions:

(i) for each \( i = 1, \ldots, n \) the datum \( y_i \) is the outcome of a normal RV of known variance \( \sigma_i^2 \). The mean value of this variable is \( a \text{ priori} \) unknown, but it is assumed to depend exclusively on the abscissa \( x_i \) — of which it will be a function;

(ii) the set of data is described by the mathematical model

\[ \zeta_i = \sum_{j=1}^{p} \alpha_j \phi_j(x_i) + \varepsilon_i, \quad 1 \leq i \leq n, \quad (2) \]

in which \( \phi_j : \mathbb{R} \longrightarrow \mathbb{R}, 1 \leq j \leq p \), are \( p < n \) given functions and \( \alpha_j, 1 \leq j \leq p \), some constant parameters to be conveniently estimated, while the \( \varepsilon_i \)'s constitute a set of \( n \) independent normal variables of vanishing mean and variance \( \sigma_i^2 \)

\[ \mathbb{E}(\varepsilon_i) = 0 \quad \mathbb{E}(\varepsilon_i \varepsilon_k) = \delta_{ik} \sigma_k^2 \quad 1 \leq i, k \leq n. \]

The value \( y_i \) must be regarded as an outcome of the RV \( \zeta_i \), for each \( i \).
We define **linear regression** of the data set, by the chi-square method (**chi-square fitting**), the expression

\[ R(x) = \sum_{j=1}^{p} a_j \phi_j(x) \quad , \quad x \in \mathbb{R} , \]

where the \( a_j \)'s are the estimates of the parameters \( \alpha_j \) obtained by means of the sample \( \{(x_i, y_i), 1 \leq i \leq n\} \) by minimizing the quadratic functional \( \mathcal{L} : \mathbb{R}^p \longrightarrow \mathbb{R} \) defined by

\[ \mathcal{L}(\alpha_1, \ldots, \alpha_p) = \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \left[ -y_i + \sum_{j=1}^{p} \alpha_j \phi_j(x_i) \right]^2 \quad (3) \]

If the variances \( \sigma_i^2 \) of the RVs \( \varepsilon_i \) are all equal, their common value can be ignored in the minimization of the functional (3), which reduces therefore to

\[ \mathcal{L}(\alpha_1, \ldots, \alpha_p) = \sum_{i=1}^{n} \left[ -y_i + \sum_{j=1}^{p} \alpha_j \phi_j(x_i) \right]^2 \]

In such a case the procedure is known as **linear regression by the least-squares method** (**least-squares fitting**)

The more general chi-square linear regression method is also called **weighted least-squares fitting**

Stefano Siboni
7.2.1 Optimization. Normal equation and its solution

The critical points of the quadratic functional (3) are all and only the solutions $a$ of the normal equation

$$Fa = \Phi \sigma^{-1} y,$$  \hspace{1cm} (4)

where the matrices $F$ and $\Phi$ are defined by

$$F_{jk} = \sum_{i=1}^{n} \frac{\phi_j(x_i)\phi_k(x_i)}{\sigma_i^2} \quad 1 \leq j, k \leq p$$  \hspace{1cm} (5)

and

$$\Phi_{ki} = \frac{\phi_k(x_i)}{\sigma_i} \quad 1 \leq k \leq p, \ 1 \leq i \leq n$$  \hspace{1cm} (6)

while

$$a = \begin{pmatrix} a_1 \\ \vdots \\ a_p \end{pmatrix}, \quad \sigma^{-1} = \begin{pmatrix} 1/\sigma_1 & & \\ & 1/\sigma_2 & \emptyset \\ & & \ddots \\ \emptyset & & 1/\sigma_n \end{pmatrix}, \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

In particular, there holds $F = \Phi \Phi^T$, positive semidefinite real symmetric matrix. If the $p \times n$ matrix $\Phi$ has maximum rank $p$ (which is the typical situation), then $F$ is positive definite and the only solution of the normal equation (4) writes

$$a = F^{-1} \Phi \sigma^{-1} y.$$  \hspace{1cm} (7)
7.2.2 The parameters of the linear regression, estimated by chi-square fitting, as RVs.

- **The estimates $a_1, \ldots, a_p$ as normal RVs**
  If $\Phi$ has maximum rank $p$, the estimate (7) of the regression parameters constitutes a set of $p$ normal variables with means
  \[ \mathbb{E}(a_j) = \alpha_j, \quad 1 \leq j \leq p, \]
  and covariance matrix
  \[ C_a = \mathbb{E}\left[ (a - \mathbb{E}(a))(a - \mathbb{E}(a))^T \right] = F^{-1} \quad (8) \]
  and corresponding structure matrix $F$

**Remark**

\[ a_1, \ldots, a_p \text{ independent } \iff F \text{ diagonal} \]

- **Normalized sum of squares around regression**
  If $\Phi$ has maximum rank $p < n$, the normalized sum of squares around regression
  \[ \text{NSSAR} = \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \left[ -y_i + \sum_{j=1}^{p} \phi_j(x_i)a_j \right]^2 \quad (9) \]
  is a $\chi^2$ RV with $n - p$ d.o.f.

- **Stochastic independence of the estimated parameters**
  The NSSAR is a random variable stochastically independent on each of the estimated parameters $a_1, \ldots, a_p$ — defined by (7).
7.2.3 Goodness of fit

Too large values of NSSAR suggest that the model is presumably incorrect and should be rejected.

The null hypothesis

\[ H_0 : \text{the proposed model is correct} \]

can then be checked by a test based on the variable NSSAR and whose rejection region takes the form

\[ \{ \text{NSSAR} \geq \chi^2_{1-\alpha(n-p)} \} \]

where \( \alpha \) is, as usual, the probability of type I error.

In common practice we often prefer to reckon the probability \( Q \) that the variable NSSAR takes a value not smaller than the observed one

\[
Q = \int_{\text{NSSAR}}^{+\infty} p_{n-p}(\chi^2) \, d\chi^2 = \\
= \int_{\text{NSSAR}}^{+\infty} \frac{1}{\Gamma((n-p)/2) 2^{(n-p)/2}} e^{-\chi^2/2} (\chi^2)^{(n-p)/2-1} \, d\chi^2
\]

and use \( Q \) as an indicator of the goodness of fit of the model.

Very small values of \( Q \) induce to reject the model as unsatisfactory.
• **What happens if \( Q \) turns out to be very small?**

There are three possibilities:

(i) the model is incorrect, to be rejected

(ii) the variances of the normal variables \( y_i \) or \( \varepsilon_i \) have not been correctly calculated (the test assumes they are exactly known): NSSAR does not follow a \( \chi^2 \), or anyway a known, distribution

(iii) the random variables \( y_i \) or \( \varepsilon_i \) do not obey a normal distribution (or not all of them, at least): NSSAR is not a \( \chi^2 \) variable and its probability distribution appears unknown

The circumstance (iii) is quite common, as signalled by the occurrence of outliers

• **What must we do?**

(i) We may tolerate values of \( Q \) much smaller than those calculated from the cumulative distribution of \( \chi^2 \) without rejecting the model

(ii) If the (non-normal) distributions of the single variables \( y_i \) or \( \varepsilon_i \) are known, we can apply a Monte Carlo method:

we generate at random the values of \( y_i \)’s;

we determine the best-fit values \( a_j \)’s of the parameters \( \alpha_j \)’s

we calculate the value of NSSAR;

by repeating many times the previous procedure, we deduce the approximate probability distributions of NSSAR and of the best-fit coefficients \( a_j \)’s, by plotting the suitable histograms.

It is thus possible to estimate \( Q \) (goodness of fit), along with the error on the best-fit coefficients (very expensive!)
• **What happens if** $Q$ **is very large?**

Typically the problem *does not* arise because the variables $y_i$ are not normal.

Non-normal distributions generally determine a wider spread of data, a larger value of the calculated NSSAR and therefore a smaller $Q$.

The problem often originates from an *overestimate* of the standard deviations $\sigma_i$, so that the calculated value of NSSAR turns out to be smaller than due, and $Q$ larger.

• **How can we understand whether** $Q$ **is too small or too large?**

We simply have to recall that if the model is correct, the variables $y_i$’s normal and the variances $\sigma_i^2$ correct, the NSSAR follows a $\chi^2$ distribution with $\nu = n - p$ d.o.f.

The $\chi^2$ distribution with $\nu$ d.o.f. has mean $\nu$ and variance $2\nu$, and for large $\nu$ it can be approximated by a normal distribution with mean $\nu$ and variance $2\nu$.

As a consequence, typical values of the variable NSSAR belong to the interval

$$[\nu - 3\sqrt{2\nu}, \nu + 3\sqrt{2\nu}]$$

If the calculated value of NSSAR falls outside this interval, the result is suspect.
7.2.4 Remark. Standard deviations $\sigma_i$ unknown

If the standard deviations $\sigma_i$ of the variables $y_i$’s are unknown, the NSSAR cannot be determined

If however:

(i) the model is correct;

(ii) we can assume that all the variances $\sigma_i$ are equal to a common value $\sigma$ (i.e., the model is homoscedastic);

we can:

− pose $\sigma = 1$;

− calculate the best-fit solution for the model parameters;

− determine the relative NSSAR $\Rightarrow$ NSSAR$|_{\sigma^2=1}$;

− estimate a likely value of $\sigma$ by means of the relationship

$$\sigma^2 = \frac{1}{n-p}\text{NSSAR}|_{\sigma^2=1} = \frac{1}{n-p} \sum_{i=1}^{n} \left[ -y_i + \sum_{j=1}^{p} \phi_j(x_i)a_j \right]^2$$

since NSSAR is $\chi^2$ with $n - p$ d.o.f.

$$\mathbb{E}[\text{NSSAR}|_{\sigma^2=1}] = \sigma^2 \mathbb{E}[\text{NSSAR}] = \sigma^2(n - p)$$

Remark: this method cannot provide any information about the goodness of fit (the model is assumed a priori correct)!
7.3 $F$-test for the goodness of fit

Instead of using NSSAR — a $\chi^2$ variable with $n - p$ d.o.f. — the goodness of fit can be estimated by an appropriate Fisher variable ($F$-test)

- We consider models with an additive parameter

$$\zeta_i = \alpha_1 + \sum_{j=2}^{p} \alpha_j \phi_j(x_i) + \varepsilon_i \quad \forall i = 1, \ldots, n$$

and the estimated regression function is thus denoted with

$$R(x) = a_1 + \sum_{j=2}^{p} a_j \phi_j(x)$$

- If the model is correct, the variable

$$\text{NSSAR} = \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \left[ y_i - a_1 - \sum_{j=2}^{p} a_j \phi_j(x_i) \right]^2 = \chi^2_{n-p}$$

follows a $\chi^2$ distribution with $n - p$ d.o.f.

This is true also in the absence of additive parameters and for any values of the coefficients $\alpha_1, \ldots, \alpha_p$
(provided that $\Phi$ is of maximum rank $p$)
• The “equation of variations” is satisfied

\[
\sum_{i=1}^{n} \frac{1}{\sigma_i^2} (y_i - \bar{y})^2 = \sum_{i=1}^{n} \frac{1}{\sigma_i^2} [R(x_i) - \bar{y}]^2 + \sum_{i=1}^{n} \frac{1}{\sigma_i^2} [y_i - R(x_i)]^2
\]

where the following quantities appear:

\[
\bar{y} = \left(\sum_{i=1}^{n} \frac{1}{\sigma_i^2}\right)^{-1} \sum_{i=1}^{n} \frac{1}{\sigma_i^2} y_i \quad \text{weighted mean of } y_i \text{'s}
\]

\[
\sum_{i=1}^{n} \frac{1}{\sigma_i^2} (y_i - \bar{y})^2 \quad \text{total variation of data}
\]

\[
\sum_{i=1}^{n} \frac{1}{\sigma_i^2} [R(x_i) - \bar{y}]^2 \quad \text{variation “explained” by the regression}
\]

\[
\sum_{i=1}^{n} \frac{1}{\sigma_i^2} [y_i - R(x_i)]^2 \quad \text{residual variation (or NSSAR)}
\]

It is a purely algebraic relationship, which holds also if the regression model is incorrect provided that:

(i) the model is linear in the regression parameters
(ii) the parameters are evaluated by the chi-square method
(iii) an additive parameter is present!

The additive parameter lacking, the equation does not hold (a point seldom stressed in applied statistics textbooks!)
• If \( y \) actually depends on \( x \) \((y(x) \neq \text{constant})\) we expect a regression model of the form

\[
\zeta_i = \alpha_1 + \sum_{j=2}^{p} \alpha_j \phi_j(x_i) + \varepsilon_i \quad \forall i = 1, \ldots, n
\]

with parameters \( \alpha_2, \ldots, \alpha_p \) not all vanishing

If moreover the model is correct, the agreement between data and the regression function will be presumably good, so that

variation explained by the regression \( \gg \) residual variation

\[
\begin{align*}
\text{(\( \alpha_2, \ldots, \alpha_p \) not all vanishing)}
\end{align*}
\]

\[
\begin{align*}
(\alpha_2, \ldots, \alpha_p) \text{ all vanishing}
\end{align*}
\]

• If \( y \) does not depend on \( x \), then \( y(x) = \text{constant} \) and the regression model will still be the previous one, but with \( \alpha_2, \ldots, \alpha_p = 0 \). Therefore

\[0 = \text{variation explained by the regression} \leq \text{residual variation}\]
• For $y$ independent on $x$ (regression model correct with $\alpha_2, \ldots, \alpha_p = 0$) the following properties can be proved:

- the total variation is a $\chi^2$ variable with $n - 1$ d.o.f.

\[ \chi^2_{n-1} = \sum_{i=1}^{n} \frac{1}{\sigma_i^2} (y_i - \overline{y})^2 \]

- the variation explained by the regression constitutes a $\chi^2$ variable with $p - 1$ d.o.f.

\[ \chi^2_{p-1} = \sum_{i=1}^{n} \frac{1}{\sigma_i^2} [R(x_i) - \overline{y}]^2 \]

- explained variation and residual variation NSSAR are stochastically independent $\chi^2$ variables. The ratio:

\[ F = \frac{\text{explained variance}}{\text{residual variance}} = \frac{\text{explained variation}/(p - 1)}{\text{residual variation}/(n - p)} = \]

\[ = \frac{n - p}{p - 1} \frac{\chi^2_{p-1}}{\chi^2_{n-p}} = \frac{n - p}{p - 1} \frac{\chi^2_{p-1}}{\text{NSSAR}} \]

so thus follows a Fisher distribution with $(p - 1, n - p)$ d.o.f.

**The additive parameter is needed!**

If this is not the case, the variables $\chi^2_{n-1}$ and $\chi^2_{p-1}$ do not follow the above distributions, and neither does $F$
• We test the **null hypothesis**

\[ H_0 : \text{correct regression model} \quad \alpha_1 + \sum_{j=2}^{p} \alpha_j \phi_j \]

with \( \alpha_2 = \ldots = \alpha_p = 0 \)

that is

\[ H_0 : y \text{ independent on } x \]

versus the **alternative hypothesis**

\[ H_1 : \text{correct regression model} \quad \alpha_1 + \sum_{j=2}^{p} \alpha_j \phi_j \]

with \( \alpha_2, \ldots, \alpha_p \) not all vanishing

The test variable is

\[ F = \frac{\text{explained variance}}{\text{residual variance}} = \frac{n - p}{p - 1} \frac{\chi^2_{p-1}}{\text{NSSAR}} \]

and typically implies

- large \( F \) \( \iff \) good fit with \( \alpha_2, \ldots, \alpha_p \) not all vanishing (\( H_0 \) false)

- small \( F \) \( \iff \) good fit with \( \alpha_2, \ldots, \alpha_p \) all vanishing (\( H_0 \) true)

The hypothesis \( H_0 \) is rejected if

\[ F > F_{[1-\alpha](p-1,n-p)} \]

with significance level \( \alpha \in (0,1) \)
7.4 *F*-test with repeated measurements on \( y \)

We suppose that for each \( x_i, \ i = 1, \ldots, n \), more values of the corresponding variable \( y \) are available

\[
y_{ik} \quad k = 1, \ldots, n_i
\]

for a total number of observations given by

\[
N = \sum_{i=1}^{n} n_i
\]

These values must be regarded as the outcomes of independent normal variables of mean \( \mu_i \) and variance \( \sigma_i^2 \)

The hypotheses of the regression are satisfied: moreover, we assume that the \( n_i \) observations of \( y \) at a given \( x = x_i \) are **independent**

The regression model (to be tested) is the usual one

\[
\zeta_i = \sum_{j=1}^{p} \alpha_j \phi_j(x_i) + \varepsilon_i \quad \forall \ i = 1, \ldots, n
\]

with regression parameters \( \alpha_1, \ldots, \alpha_p \) and \( \varepsilon_i \) variable \( N(0, \sigma_i) \)

If the model is correct, the mean value of \( \zeta_i \) holds

\[
\mu_i = \sum_{j=1}^{p} \alpha_j \phi_j(x_i) \quad \forall \ i = 1, \ldots, n
\]

We distinguish two cases:

- known variances \( \sigma_i^2 \)
- variances \( \sigma_i^2 \) unknown but equal \( (\sigma_i^2 = \sigma^2 \ \forall \ i) \)
7.4.1 Known variances $\sigma_i^2$

We calculate two normalized sums of squares

$$\text{NSS}_{m.i.} = \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \sum_{k=1}^{n_i} (y_{ik} - \overline{y}_i)^2$$

$$\text{NSS}_{m.d.} = \sum_{i=1}^{n} \frac{1}{\sigma_i^2 / n_i} \left[ \overline{y}_i - \sum_{j=1}^{p} a_j \phi_j(x_i) \right]^2$$

where $\overline{y}_i$ is the mean of $y_{ik}$'s at a given $x = x_i$

$$\overline{y}_i = \frac{1}{n_i} \sum_{k=1}^{n_i} y_{ik}$$

and $a_1, \ldots, a_p$ are the chi-square estimates of the parameters $\alpha_1, \ldots, \alpha_p$, obtained by minimizing the functional

$$\mathcal{L}(\alpha_1, \ldots, \alpha_p) = \sum_{i=1}^{n} \frac{1}{\sigma_i^2 / n_i} \left[ \overline{y}_i - \sum_{j=1}^{p} \alpha_j \phi_j(x_i) \right]^2$$

$\text{NSS}_{m.i.}$ describes the data spread in a way independent on the regression model

$\text{NSS}_{m.d.}$ describes the data spread around the regression model: it expresses the variation “unexplained” by the regression

In a heuristic way, we may assume the correspondence

\[ \text{incorrect regression model} \quad \Longleftrightarrow \quad \text{NSS}_{m.d.} \gg \text{NSS}_{m.i.} \]
More precisely:

- \( NSS_{m.i.} \) is a \( X^2 \) variable with \( N - n \) d.o.f., **no matter the regression model is correct or is not**

- \( NSS_{m.d.} \) can be proved to be a \( X^2 \) variable with \( n - p \) d.o.f. **if the regression model is correct**

- In such a case \( NSS_{m.i.} \) and \( NSS_{m.d.} \) are **stochastically independent**

For a correct model the ratio of the reduced \( X^2 \) variables

\[
\frac{NSS_{m.d.}/(n - p)}{NSS_{m.i.}/(N - n)} = \frac{N - n}{n - p} \frac{NSS_{m.d.}}{NSS_{m.i.}}
\]

follows a Fisher distribution with \((n - p, N - n)\) d.o.f.

The null hypothesis

\[ H_0 : \text{the regression model is correct} \]

will be then rejected when

\[
\frac{N - n}{n - p} \frac{NSS_{m.d.}}{NSS_{m.i.}} > F_{1-\alpha}(n-p,N-n)
\]

with a significance level \( \alpha \in (0, 1) \)
7.4.2 Equal (although unknown) variances

If the variances $\sigma_i^2$ take the same value $\sigma^2$, the previous discussion is still valid and we reject the regression model if

$$\frac{N - n}{n - p} \frac{\text{NSS}_{m.d.}}{\text{NSS}_{m.i.}} > F_{[1-\alpha]}(n-p,N-n)$$

In this case the ratio of the reduced chi-square variables is independent on $\sigma^2$, which might also be unknown.

We simply have to write

$$\frac{N - n}{n - p} \frac{\text{NSS}_{m.d.}}{\text{NSS}_{m.i.}} = \frac{N - n}{n - p} \frac{\text{SS}_{m.d.}}{\text{SS}_{m.i.}}$$

in terms of the sums of squares

$$\text{SS}_{m.i.} = \sum_{i=1}^{n} \sum_{k=1}^{n_i} (y_{ik} - \bar{y}_i)^2$$

$$\text{SS}_{m.d.} = \sum_{i=1}^{n} n_i \left[ \bar{y}_i - \sum_{j=1}^{p} a_j \phi_j(x_i) \right]^2$$

which do not contain $\sigma^2$.

An estimate of $\sigma^2$ is given by

$$\frac{\text{SS}_{m.i.}}{N - n} = \frac{1}{N - n} \sum_{i=1}^{n} \sum_{k=1}^{n_i} (y_{ik} - \bar{y}_i)^2$$

or, if the regression model is correct, by

$$\frac{\text{SS}_{m.d.}}{n - p} = \frac{1}{n - p} \sum_{i=1}^{n} n_i \left[ \bar{y}_i - \sum_{j=1}^{p} a_j \phi_j(x_i) \right]^2$$
7.5 Incremental $F$-test

For a given data set \( \{(x_i, y_i), 1 \leq i \leq n\} \) let us consider two linear regression models which differ for an additional parameter

\[
\sum_{j=1}^{p} \alpha_j \phi_j(x) \quad \text{and} \quad \sum_{j=1}^{p+1} \alpha_j \phi_j(x)
\]

If the first model is correct, so can be considered the second (we only have to assume \( \alpha_{p+1} = 0 \))

In such a case the chi-square best-fit procedure leads to the \( \chi^2 \) variables

\[
\text{NSSAR}_p \quad \text{and} \quad \text{NSSAR}_{p+1}
\]

with \( n - p \) and \( n - p - 1 \) d.o.f. respectively

It is clear that for any choice of the data \((x_i, y_i)\) there holds

\[
\text{NSSAR}_p \geq \text{NSSAR}_{p+1}
\]

It is possible to deduce that the difference

\[
\text{NSSAR}_p - \text{NSSAR}_{p+1}
\]

constitutes a \( \chi^2 \) variable with 1 d.o.f.

and is \textbf{stochastically independent} on \( \text{NSSAR}_{p+1} \)

This allows us to introduce the Fisher variable

\[
F_A = \frac{(\text{NSSAR}_p - \text{NSSAR}_{p+1})/1}{\text{NSSAR}_{p+1}/(n - p - 1)}
\]

with 1, \( n - p - 1 \) d.o.f.
The ratio $F_A$ measures the improvement of the reduced chi-square of the model due to the additional term in the regression function.

$F_A$ will be **small** if the $p+1$ parameter model does not improve significantly the fitting of the $p$ parameter model.

In contrast, for **large** values of $F_A$ we will have to deduce that the additional parameter has significantly improved the fitting and that therefore the $p+1$ parameter model **must be preferred** to the $p$ parameter one.

Therefore:

\[
\text{small } F_A \iff p \text{ parameter model more satisfactory } \quad (H_0) \\
\text{large } F_A \iff p + 1 \text{ parameter model more adequate } \quad (H_1)
\]

As a conclusion:

if $F_A \leq F_{[1-\alpha](1,n-p-1)}$ we accept the hypothesis $H_0$ that the $p$ parameter model is correct;

if $F_A > F_{[1-\alpha](1,n-p-1)}$ we reject such hypothesis, by accepting the $p+1$ parameter model with $\alpha_{p+1}$ significantly different from zero.
7.6 Confidence intervals for the regression parameters

For a general linear regression model of the form

\[ \zeta_i = \sum_{j=1}^{p} \alpha_j \phi_j(x_i) + \varepsilon_i \quad i = 1, \ldots, n \]

the confidence intervals of the parameter estimates \( a_1, \ldots, a_p \) are given by

\[ \alpha_j = a_j \pm t_{\left[1-\frac{\eta}{2}\right]}(n-p) \sqrt{(F^{-1})_{jj} \frac{\text{NSSAR}}{n-p}} \quad j = 1, \ldots, p \]

with confidence level \( 1 - \eta \in (0, 1) \).

7.7 Confidence interval for predictions

We assume that the value of the dependent variable \( y \) at a given abscissa \( x = x_0 \) is represented by the random variable

\[ \zeta_0 = \sum_{j=1}^{p} \alpha_j \phi_j(x_0) + \varepsilon_0 \]

where \( \varepsilon_0 \) is a normal RV of zero mean and variance \( \sigma_0^2 \), independent on the \( \varepsilon_i \)'s.

The prediction of \( y \) at \( x = x_0 \), based on the linear regression, is defined as the RV

\[ y_0 = \sum_{j=1}^{p} a_j \phi_j(x_0) + \varepsilon_0 \]
It constitutes a normal RV of mean

$$\mathbb{E}(y_0) = \mathbb{E}(\zeta_0) = \sum_{j=1}^{p} \alpha_j \phi_j(x_i)$$

and variance

$$\text{var}(y_0) = \sum_{j,k=1}^{p} \phi_j(x_0) \phi_k(x_0) \text{cov}(a_j, a_k) + \sigma_0^2 =$$

$$= \sum_{j,k=1}^{p} \phi_j(x_0) \phi_k(x_0) \left(F^{-1}\right)_{jk} + \sigma_0^2,$$

so that the confidence interval of $\mathbb{E}(y_0)$ takes the form

$$\mathbb{E}(y_0) = y_0 \pm t_{\left[1-\frac{\alpha}{2}\right]}(n-p) \sqrt{\text{var}(y_0)} \sqrt{\frac{\text{NSSAR}}{n-p}}$$

with confidence level $1 - \alpha \in (0, 1)$

**In the homoscedastic case** the confidence interval is independent on the common standard deviation $\sigma$, since

$$\text{var}(y_0) \frac{\text{NSSAR}}{n-p} = \left[ \sum_{j,k=1}^{p} \phi_j(x_0) \phi_k(x_0) \frac{1}{\sigma^2} \left(F^{-1}\right)_{jk} + 1 \right] \frac{\text{SSAR}}{n-p}$$

on having introduced the sum of squares around regression

$$\text{SSAR} = \sum_{i=1}^{n} \left[ -y_i + \sum_{j=1}^{p} a_j \phi_j(x_i) \right]^2$$
7.8 Important case. Regression straight line

The linear model reduces to

\[ \zeta_i = \alpha + \beta x_i + \varepsilon_i \]

with regression parameters \( \alpha \) and \( \beta \)

The related chi-square estimates \( a \) and \( b \) minimize the sum

\[
NSSAR = \sum_{i=1}^{n} \left( \frac{a + bx_i - y_i}{\sigma_i} \right)^2,
\]

a \( \chi^2 \) variable with \( n - 2 \) d.o.f. independent on \( a \) and \( b \)

The normal equations

\[
\begin{align*}
    aS + bS_x &= S_y \\
    aS_x + bS_{xx} &= S_{xy}
\end{align*}
\]

have the solution

\[
\begin{align*}
    a &= \frac{S_{xx}S_y - S_xS_{xy}}{\Delta} \\
    b &= \frac{SS_{xy} - S_xS_y}{\Delta}
\end{align*}
\]

with covariance matrix and correlation coefficient

\[
C = \frac{1}{\Delta} \begin{pmatrix} S_{xx} & -S_x \\ -S_x & S \end{pmatrix} \quad \text{corr}(a, b) = -\frac{S_x}{\sqrt{SS_{xx}}}
\]

where:

\[
\begin{align*}
    S &= \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \\
    S_x &= \sum_{i=1}^{n} \frac{x_i}{\sigma_i^2} \\
    S_y &= \sum_{i=1}^{n} \frac{y_i}{\sigma_i^2} \\
    S_{xx} &= \sum_{i=1}^{n} \frac{x_i^2}{\sigma_i^2} \\
    S_{xy} &= \sum_{i=1}^{n} \frac{x_iy_i}{\sigma_i^2} \\
    \Delta &= SS_{xx} - S_x^2 > 0
\end{align*}
\]

Remark The random variables \( a \) and \( b \) are not independent!
Regression straight line. Alternative model

We consider the 2 parameter model

\[ \zeta_i = \mu + \kappa (x_i - \bar{x}) + \varepsilon_i \]

where

\[ \bar{x} = \left( \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \right)^{-1} \sum_{i=1}^{n} \frac{1}{\sigma_i^2} x_i \]

while \( \mu \) and \( \kappa \) are the regression parameters.

The related chi-square estimates \( m \) and \( q \) minimize the sum

\[ \text{NSSAR} = \sum_{i=1}^{n} \left( \frac{m + q(x_i - \bar{x}) - y_i}{\sigma_i} \right)^2, \]

a \( \chi^2 \) variable with \( n - 2 \) d.o.f. independent on \( m \) and \( q \).

The normal equations provide the solution

\[ m = \frac{\sum_{i=1}^{n} \frac{1}{\sigma_i^2} y_i}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}} \quad q = \frac{\sum_{i=1}^{n} \frac{(x_i - \bar{x}) y_i}{\sigma_i^2}}{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\sigma_i^2}} \]

with a diagonal covariance matrix

\[ C = \begin{pmatrix} \left( \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \right)^{-1} & 0 \\ 0 & \left( \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\sigma_i^2} \right)^{-1} \end{pmatrix} \]

and therefore a correlation coefficient equal to zero.

The random variables \( m \) and \( q \) are independent!

(while \( a \) and \( b \) are not in the previous case)
Regression straight line
Confidence intervals for the parameters $\alpha$ and $\beta$

The 2 parameter linear regression model

$$\zeta_i = \alpha + \beta x_i + \varepsilon_i \quad i = 1, \ldots, n$$

being given, the confidence intervals for the estimates $a$ and $b$ of the parameters $\alpha$ and $\beta$ can be written as

$$\alpha = a \pm t_{[1-\frac{\eta}{2}]}(n-2) \sqrt{\frac{S_{xx}}{\Delta} \frac{NSSAR}{n-2}}$$

$$\beta = b \pm t_{[1-\frac{\eta}{2}]}(n-2) \sqrt{\frac{S}{\Delta} \frac{NSSAR}{n-2}}$$

with confidence level $1 - \eta \in (0, 1)$.

We have again

$$a = \frac{S_{xx}S_y - S_xS_{xy}}{\Delta} \quad b = \frac{SS_{xy} - S_xS_y}{\Delta}$$

In the case of equal variances $\sigma_i^2 = \sigma^2$ the confidence intervals are independent on $\sigma$

Remark
The probability that $\alpha$ and $\beta$ belong simultaneously to the respective confidence intervals is not $(1-\eta)^2 = (1-\eta)(1-\eta)$
The variables $a$ and $b$ are not independent!
Regression straight line
Confidence intervals for the parameters $\mu$ and $\kappa$

For the 2 parameter linear regression model

$$\zeta_i = \mu + \kappa(x_i - \bar{x}) + \varepsilon_i \quad i = 1, \ldots, n$$

the confidence intervals of the estimates $m$ and $q$ of the parameters $\mu$ and $\kappa$ can be written

$$\mu = m \pm t_{[1-\frac{\eta}{2}]\,(n-2)} \sqrt{\left( \frac{1}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}} \right)^{-1} \frac{\text{NSSAR}}{n-2}}$$

$$\kappa = q \pm t_{[1-\frac{\eta}{2}]\,(n-2)} \sqrt{\left( \frac{1}{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\sigma_i^2}} \right)^{-1} \frac{\text{NSSAR}}{n-2}}$$

with confidence level $1 - \eta \in (0, 1)$

In the case of equal variances $\sigma_i^2 = \sigma^2$ the confidence intervals are again independent on $\sigma$

**Remark**
The probability that $\mu$ and $\kappa$ belong simultaneously to the respective confidence intervals is $(1-\eta)^2 = (1-\eta)(1-\eta)$
The variables $m$ and $q$ are independent!
Regression straight line
Confidence interval for predictions

We consider the linear regression model

$$\zeta_i = \mu + \kappa(x_i - \bar{x}) + \varepsilon_i$$

We assume that the value of the variable $y$ corresponding to an abscissa $x = x_0$ is described by a random variable

$$\zeta_0 = \mu + \kappa(x_0 - \bar{x}) + \varepsilon_0$$

with $\varepsilon_0$ normal variable of zero mean and variance $\sigma^2_0$, independent on $\varepsilon_i$’s.

We define the prediction of the value of $y$ in $x_0$ based on the linear regression as the random variable

$$y_0 = m + q(x_0 - \bar{x}) + \varepsilon_0$$

The prediction $y_0$ is a normal variable of mean $\mathbb{E}(y_0) = \mathbb{E}(\zeta_0) = \mu + \kappa(x_0 - \bar{x})$ — the “true” prediction — and variance

$$\text{var}(y_0) = \left(\sum_{i=1}^{n} \frac{1}{\sigma^2_i}\right)^{-1} + \left(\sum_{i=1}^{n} \frac{1}{\sigma^2_i}(x_i - \bar{x})^2\right)^{-1}(x_0 - \bar{x})^2 + \sigma^2_0$$

The confidence interval of $\mathbb{E}(y_0)$ takes then the form

$$\mathbb{E}(y_0) = y_0 \pm t_{[1-\frac{\alpha}{2}](n-2)} \sqrt{\text{var}(y_0)} \sqrt{\frac{\text{NSSAR}}{n-2}}$$

with confidence level $1 - \alpha \in (0, 1)$
If the variance is independent on the abscissa
\[ \sigma_i = \sigma \quad \forall i = 1, \ldots, n \quad \text{and} \quad \sigma_0 = \sigma, \]
the previous confidence interval is independent on \( \sigma \) too
\[
E(y_0) = m + q(x_0 - \bar{x}) \pm t_{[1 - \frac{\alpha}{2}],(n-2)} \sqrt{V} \sqrt{\frac{\text{SSAR}}{n - 2}}
\]
with
\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \text{SSAR} = \sum_{i=1}^{n} [ -y_i + m + q(x_i - \bar{x}) ]^2 \\
V = \frac{1}{n} + \frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})^2} (x_0 - \bar{x})^2 + 1
\]

**Remark**
The coefficient \( V \) decreases when the spread of \( x_i \)'s around their mean value \( \bar{x} \) grows

**Remark**
By varying \( x_0 \in \mathbb{R} \), the confidence interval describes a region of the \((x,y)\) plane of the form
\[
\{ (x,y) \in \mathbb{R}^2 : |m + q(x - \bar{x}) - y| \leq c \sqrt{1 + g(x - \bar{x})^2} \}
\]
with \( c \) and \( g \) suitable positive constants
This **confidence region** constitutes a neighborhood of the regression straight line and its qualitative trend is illustrated by the shadowed area in the following picture.

![Confidence Region Diagram](image)

For values of $x_i$'s more dispersed around the mean $\bar{x}$ the confidence region shrinks.

**Remark**

In the case of the linear regression model

$$
\zeta_i = \alpha + \beta x_i + \varepsilon_i \quad i = 1, \ldots, n
$$

the prediction at $x = x_0$ is defined in an analogous way

$$
y_0 = a + bx_0 + \varepsilon_0
$$

with $\varepsilon_0$ normal variable of zero mean and variance $\sigma_0^2$, independent on $\varepsilon_i$'s. The confidence interval of $\mathbb{E}(y_0)$ has the same
formal expression as before

\[ \mathbb{E}(y_0) = y_0 \pm t_{1-\frac{\alpha}{2}, (n-2)} \sqrt{\text{var}(y_0)} \sqrt{\frac{\text{NSSAR}}{n-2}} \]

although the variance of \( y_0 \) takes now a different form

\[
\text{var}(y_0) = (1 \quad x_0) \frac{1}{\Delta} \begin{pmatrix} S_{xx} & -S_x \\ -S_x & S \end{pmatrix} \begin{pmatrix} 1 \\ x_0 \end{pmatrix} + \sigma_0^2 = \frac{1}{\Delta} \left( S_{xx} - 2S_x x_0 + S x_0^2 \right) + \sigma_0^2
\]

due to the different structure of the covariance matrix of the coefficient estimates \( a, b \)

\[
C = \frac{1}{\Delta} \begin{pmatrix} S_{xx} & -S_x \\ -S_x & S \end{pmatrix}
\]

For homoscedastic systems the CI is independent on the common standard deviation \( \sigma \). Indeed

\[
\text{var}(y_0) \frac{\text{NSSAR}}{n-2} = \left[ \frac{1}{\Delta} \left( S_{xx} - 2S_x x_0 + S x_0^2 \right) + \sigma^2 \right] \frac{1}{\sigma^2} \frac{\text{SSAR}}{n-2} = \left[ \frac{1}{\sigma^2 \Delta} \left( S_{xx} - 2S_x x_0 + S x_0^2 \right) + 1 \right] \frac{\text{SSAR}}{n-2}
\]

with

\[
\text{SSAR} = \sum_{i=1}^{n} [ -y_i + a + bx_i ]^2
\]
Example: Regression straight line

Experimental measurements of a quantity $y$ versus another quantity $x$ are collected in the table below

<table>
<thead>
<tr>
<th>$k$</th>
<th>$x_k$</th>
<th>$y_{k1}$</th>
<th>$y_{k2}$</th>
<th>$y_{k3}$</th>
<th>$y_{k4}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.6</td>
<td>0.55</td>
<td>0.80</td>
<td>0.75</td>
<td>$\times$</td>
</tr>
<tr>
<td>2</td>
<td>1.0</td>
<td>1.10</td>
<td>0.95</td>
<td>1.05</td>
<td>$\times$</td>
</tr>
<tr>
<td>3</td>
<td>1.5</td>
<td>1.40</td>
<td>1.30</td>
<td>1.50</td>
<td>$\times$</td>
</tr>
<tr>
<td>4</td>
<td>2.0</td>
<td>1.60</td>
<td>1.90</td>
<td>1.85</td>
<td>1.95</td>
</tr>
<tr>
<td>5</td>
<td>2.5</td>
<td>2.25</td>
<td>2.30</td>
<td>2.40</td>
<td>2.45</td>
</tr>
<tr>
<td>6</td>
<td>2.9</td>
<td>2.65</td>
<td>2.70</td>
<td>2.75</td>
<td>$\times$</td>
</tr>
<tr>
<td>7</td>
<td>3.0</td>
<td>2.65</td>
<td>2.70</td>
<td>2.80</td>
<td>2.85</td>
</tr>
</tbody>
</table>

The random error on the abscissas $x_k$’s is negligible, while the data $y_k$’s are independent random variables with the same variance $\sigma^2$ (homoscedastic system).

Determine:

(i) the regression straight line by the least-squares method, in the form

$$y = \mu + \kappa (x - \bar{x})$$

(ii) the 95%-confidence interval for the intercept $\mu$ and that for the slope $\kappa$;

(iii) the 95%-confidence region for predictions;

(iv) the 95%-confidence interval for the predicted value of $y$ at $x = 2.3$;

(v) the goodness of fit of the regression model if $\sigma = 0.08$. 

Stefano Siboni 203
Solution

(i) Since the standard deviations are equal, the chi-square fitting reduces to the usual least-squares fitting and the best-fit estimates of parameters $m$ and $q$ can be written as:

$$m = \frac{1}{n} \sum_{i=1}^{n} y_i = 1.8833 \quad q = \frac{\sum_{i=1}^{n} (x_i - \bar{x})y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = 0.87046$$

with $n = 24$ and $\bar{x} = 2.000$. The regression straight line, as calculated with the least-squares method, writes therefore:

$$y = m + q(x - \bar{x}) = 1.8833 + 0.87046(x - 2.000) = 0.1424 + 0.87046x$$

(ii) The sum of squares around regression holds:

$$SSAR = \sum_{i=1}^{n} [m + q(x_i - \bar{x}) - y_i]^2 = 0.22704376$$

At a confidence level of $1 - \alpha \in (0, 1)$ the confidence interval of the intercept $\mu$ and that of the slope $\kappa$ take the form:

$$\mu = m \pm t_{\frac{1-\alpha}{2}}(n-2) \sqrt{\frac{1}{n} \frac{SSAR}{n}}$$

$$\kappa = q \pm t_{\frac{1-\alpha}{2}}(n-2) \sqrt{\left[ \sum_{i=1}^{n} (x_i - \bar{x})^2 \right]^{-1} \frac{SSAR}{n-2}}$$
In the present case we have $\alpha = 0.05$, $n = 24$ and the confidence intervals become therefore:

$$
\mu = m \pm t_{[0.975](22)} \sqrt{\frac{1}{24} \frac{\text{SSAR}}{22}}
$$

$$
\kappa = q \pm t_{[0.975](22)} \sqrt{\left[ \sum_{i=1}^{24} (x_i - \bar{x})^2 \right]^{-1} \frac{\text{SSAR}}{22}}
$$

with:

- $m = 1.8833$
- $q = 0.87046$
- $\text{SSAR} = 0.22704376$
- $\sum_{i=1}^{24} (x_i - \bar{x})^2 = 17.060000$
- $t_{[0.975](22)} = 2.074$

As a conclusion:

- the 95%-confidence interval for the intercept $\mu$ is
  $$
  1.8833 \pm 0.0430 = [1.840, 1.926]
  $$

- the 95%-confidence interval for the slope $\kappa$ holds
  $$
  0.87046 \pm 0.05101 = [0.819, 0.921]
  $$
(iii) Since the model is assumed to be homoscedastic, the confidence interval (at a confidence level of $1 - \alpha$) for the prediction of $y = y_0$ at a given abscissa $x = x_0$ can be calculated by the general formula:

$$E(y_0) = m + q(x_0 - \bar{x}) \pm t_{[1-\frac{\alpha}{2}] (n-2)} \sqrt{V} \sqrt{\frac{SSAR}{n-2}}$$

where:

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

$$SSAR = \sum_{i=1}^{n} [ -y_i + m + q(x_i - \bar{x}) ]^2$$

$$V = 1 + \frac{1}{n} + \frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \cdot (x_0 - \bar{x})^2 .$$

In the present case we have $\bar{x} = 2.000$, so that:

$$y_0 = 1.8833 + 0.87046(x_0 - 2.000) \pm t_{[0.975](22)} \cdot \sqrt{1 + \frac{1}{24} + \frac{1}{17.060000}(x_0 - 2.000)^2} \sqrt{0.22704376 \over 22}$$

and the confidence interval for the prediction of $y$ at $x = x_0$ reduces to:

$$y_0 = 1.8833 + 0.87046(x_0 - 2.000) \pm 0.21069\sqrt{1.04167 + 0.058617(x_0 - 2.000)^2}$$
In the following picture the regression straight line is superimposed to the experimental data (dots):

\[ y = m + q(x - \bar{x}) \]

\[ m = 1.883 \pm 0.043 \]
\[ q = 0.870 \pm 0.051 \]
\[ \bar{x} = 2.000 \]

The confidence region for predictions (at a confidence level of 95\%) is evidenced in the figure below (factor V exaggerated)
(iv) the 95%-confidence interval for the prediction of $y$ at $x = 2.3$ can be obtained by replacing $x_0 = 2.3$ in the previous formula

$$y_0 = 1.8833 + 0.87046(x_0 - 2.000) \pm 0.21069\sqrt{1.04167 + 0.058617(x_0 - 2.000)^2}$$

We have therefore:

$$y_{2.3} = [1.903, 2.385] = 2.144 \pm 0.241$$

In the following graph the confidence interval is the intersection of the 95%-confidence region with the vertical line $x = 2.3$: 

![Graph showing the confidence interval for $y_{2.3}$]
(v) the goodness of fit $Q$ of the regression model is defined by the relationship

$$Q = \int_{\text{NSSAR}}^{+\infty} \rho_{n-p}(\chi^2) \, d\chi^2$$

where $\rho_{n-p}$ denotes the $\chi^2$ distribution with $n - p$ d.o.f.

This is because, if the regression model holds true, the normalized sum of squares around regression

$$\text{NSSAR} = \sum_{i=1}^{n} \frac{1}{\sigma^2} [m + q(x_i - \bar{x}) - y_i]^2 \frac{\text{SSAR}}{\sigma^2}$$

is a $\chi^2$ random variable with $n - p$ d.o.f.

In order to make calculations it is crucial to know the common value of the standard deviation $\sigma = 0.08$, since NSSAR and not simply SSAR is needed.

In the case under scrutiny we have $n = 24$ data and the regression model is based on two parameters, $\mu$ and $\kappa$; as a consequence, $p = 2$ and the NSSAR follows a $\chi^2$ distribution with $n - p = 22$ d.o.f.

For the given sample the normalized sum of squares holds

$$\text{NSSAR} = \frac{\text{SSAR}}{\sigma^2} = \frac{0.22704376}{0.08^2} = 35.47558$$
The table of upper critical values of $\chi^2$ with $\nu = 22$ d.o.f. gives

<table>
<thead>
<tr>
<th>Probability{$\chi^2 \geq 33.924$}</th>
<th>Probability{$\chi^2 \geq 36.781$}</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>0.025</td>
</tr>
</tbody>
</table>

so that a simple linear interpolation scheme:

\[
\begin{array}{c|c}
33.924 & 0.05 \\
35.476 & Q \\
36.781 & 0.025 \\
\end{array}
\]

\[
\frac{35.476 - 33.924}{36.781 - 33.924} = \frac{Q - 0.05}{0.025 - 0.05}
\]

provides the required estimate of $Q$:

\[
Q = 0.05 + (0.025 - 0.05) \frac{35.476 - 33.924}{36.781 - 33.924} = 0.0364
\]

A more accurate $Q$ follows via a numerical integration

\[
Q = \text{Probability}\{\chi^2 \geq 35.47558\} = \int_{35.47558}^{+\infty} p_{22}(\chi^2) \, d\chi^2
\]

for instance by the Maple 11 command line

1 - stats[statevalf, cdf, chisquare[22]](35.475587);

or the Excel function

\[
\text{CHIDIST}(35, 47558; 22)
\]

which lead to $Q = 0.0345$.

If the regression model were rejected, $Q$ would express the probability of a type I error — about 3.5% in the present case.
7.9 Multiple linear regression by the chi-square method

The linear regression model

\[ \zeta_i = \sum_{j=1}^{p} \alpha_j \phi_j(x_i) + \varepsilon_i \quad i = 1, \ldots, n \]

does not require that the abscissa \( x \) is a scalar variable: \( x \) may stand for a list of independent scalar variables.

If this is the case the model is known as a multiple linear regression.

The chi-square estimates of the regression coefficients are calculated exactly as in the case of a scalar \( x \), and so are the goodness of fit, the confidence intervals of the coefficients and the CI of predictions.

**Example: Multiple linear regression**

It is known from van der Waals theory that the molar enthalpy of a real gas can be approximately expressed by a relationship of the form

\[ h(T, P) = \alpha_1 T + \alpha_2 \frac{P}{T} + \alpha_3 P \]

where \( T \) denotes the absolute temperature, \( P \) stands for the pressure and \( \alpha_1, \alpha_2, \alpha_3 \) are appropriate constants related to the molar specific heat at constant pressure \( c_P \), the gas constant \( R \) and the van der Waals parameters \( a, b \):

\[ \alpha_1 = c_P \quad \alpha_2 = -\frac{2a}{R} \quad \alpha_3 = b. \]
Repeated enthalpy measurements carried out on a gas at different temperatures and pressures have produced the results listed in the table below:

<table>
<thead>
<tr>
<th>( i )</th>
<th>( T_i ) (K)</th>
<th>( P_i ) (Pa)</th>
<th>( h_i ) (J mol(^{-1}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>200</td>
<td>( 5 \cdot 10^5 )</td>
<td>5000</td>
</tr>
<tr>
<td>2</td>
<td>250</td>
<td>( 5 \cdot 10^5 )</td>
<td>6622</td>
</tr>
<tr>
<td>3</td>
<td>300</td>
<td>( 5 \cdot 10^5 )</td>
<td>8142</td>
</tr>
<tr>
<td>4</td>
<td>350</td>
<td>( 5 \cdot 10^5 )</td>
<td>9704</td>
</tr>
<tr>
<td>5</td>
<td>400</td>
<td>( 5 \cdot 10^5 )</td>
<td>11240</td>
</tr>
<tr>
<td>6</td>
<td>200</td>
<td>( 10 \cdot 10^5 )</td>
<td>4174</td>
</tr>
<tr>
<td>7</td>
<td>250</td>
<td>( 10 \cdot 10^5 )</td>
<td>5898</td>
</tr>
<tr>
<td>8</td>
<td>300</td>
<td>( 10 \cdot 10^5 )</td>
<td>7618</td>
</tr>
<tr>
<td>9</td>
<td>350</td>
<td>( 10 \cdot 10^5 )</td>
<td>9257</td>
</tr>
<tr>
<td>10</td>
<td>400</td>
<td>( 10 \cdot 10^5 )</td>
<td>10819</td>
</tr>
<tr>
<td>11</td>
<td>200</td>
<td>( 15 \cdot 10^5 )</td>
<td>3366</td>
</tr>
<tr>
<td>12</td>
<td>250</td>
<td>( 15 \cdot 10^5 )</td>
<td>5297</td>
</tr>
<tr>
<td>13</td>
<td>300</td>
<td>( 15 \cdot 10^5 )</td>
<td>7077</td>
</tr>
<tr>
<td>14</td>
<td>350</td>
<td>( 15 \cdot 10^5 )</td>
<td>8754</td>
</tr>
<tr>
<td>15</td>
<td>400</td>
<td>( 15 \cdot 10^5 )</td>
<td>10424</td>
</tr>
<tr>
<td>16</td>
<td>200</td>
<td>( 20 \cdot 10^5 )</td>
<td>2530</td>
</tr>
<tr>
<td>17</td>
<td>250</td>
<td>( 20 \cdot 10^5 )</td>
<td>4653</td>
</tr>
<tr>
<td>18</td>
<td>300</td>
<td>( 20 \cdot 10^5 )</td>
<td>6549</td>
</tr>
<tr>
<td>19</td>
<td>350</td>
<td>( 20 \cdot 10^5 )</td>
<td>8329</td>
</tr>
<tr>
<td>20</td>
<td>400</td>
<td>( 20 \cdot 10^5 )</td>
<td>10022</td>
</tr>
</tbody>
</table>

Each result is a mean of many independent measurements, whose distribution can be assumed normal with a standard deviation \( \sigma = 25 \text{ J} \cdot \text{mol}^{-1} \). We want to determine:

(i) a regression model by the least-squares method;
(ii) the CI at 95% for the parameters $\alpha_1$, $\alpha_2$ and $\alpha_3$;

(iii) the 95%-confidence region for predictions;

(iv) the CI at 95% for the predicted value of $h(T, P)$ at $T = 345$ K and $P = 8.5 \cdot 10^5$ Pa;

(v) the goodness of fit of the model.

**Solution**

(i) The **linear regression model** involves two independent variables, $T$ and $P$. Thus we are about to deal with a multiple linear regression problem. By posing $x = (T, P)$, the base functions to be considered are

\[
\phi_1(x) = T \quad \phi_2(x) = \frac{P}{T} \quad \phi_3(x) = P
\]

and the sample values of the independent variables can be denoted with

\[
x_i = (T_i, P_i) \quad i = 1, \ldots, 20
\]

For a homoscedastic system the common variance $\sigma^2$ can be neglected and the chi-square method simply reduces to a least-squares fitting. The representation matrix of the normal equations takes the form

\[
F = \begin{pmatrix}
F_{11} & F_{12} & F_{13} \\
F_{12} & F_{22} & F_{23} \\
F_{13} & F_{23} & F_{33}
\end{pmatrix}
\]

with the entries

\[
F_{11} = \frac{1}{\sigma^2} \sum_{i=1}^{20} \phi_1(x_i)\phi_1(x_i) = \frac{1}{\sigma^2} \sum_{i=1}^{20} T_i^2 = 1.9 \cdot 10^6 \frac{1}{\sigma^2}
\]
\( F_{12} = \frac{1}{\sigma^2} \sum_{i=1}^{20} \phi_1(x_i)\phi_2(x_i) = \frac{1}{\sigma^2} \sum_{i=1}^{20} P_i = 2.5 \cdot 10^7 \frac{1}{\sigma^2} \)
\( F_{13} = \frac{1}{\sigma^2} \sum_{i=1}^{20} \phi_1(x_i)\phi_3(x_i) = \frac{1}{\sigma^2} \sum_{i=1}^{20} P_i T_i = 7.5 \cdot 10^9 \frac{1}{\sigma^2} \)
\( F_{22} = \frac{1}{\sigma^2} \sum_{i=1}^{20} \phi_2(x_i)\phi_2(x_i) = \frac{1}{\sigma^2} \sum_{i=1}^{20} \frac{P_i^2}{T_i^2} = 4.98933 \cdot 10^{-8} \frac{1}{\sigma^2} \)
\( F_{23} = \frac{1}{\sigma^2} \sum_{i=1}^{20} \phi_2(x_i)\phi_3(x_i) = \frac{1}{\sigma^2} \sum_{i=1}^{20} \frac{P_i^2}{T_i} = 1.32679 \cdot 10^{11} \frac{1}{\sigma^2} \)
\( F_{33} = \frac{1}{\sigma^2} \sum_{i=1}^{20} \phi_3(x_i)\phi_3(x_i) = \frac{1}{\sigma^2} \sum_{i=1}^{20} P_i^2 = 3.750 \cdot 10^{13} \frac{1}{\sigma^2} \)

which can be easily calculated by means of an Excel Worksheet. Therefore:

\[
F = \frac{1}{\sigma^2} \begin{pmatrix}
1.9 \cdot 10^6 & 2.5 \cdot 10^7 & 7.5 \cdot 10^9 \\
2.5 \cdot 10^7 & 4.98933 \cdot 10^{-8} & 1.32679 \cdot 10^{11} \\
7.5 \cdot 10^9 & 1.32679 \cdot 10^{11} & 3.750 \cdot 10^{13}
\end{pmatrix}
\]

and the covariance matrix \( F^{-1} \) of the parameter estimates becomes

\[
\begin{pmatrix}
3.12439 \cdot 10^{-6} & 1.62631 \cdot 10^{-7} & -1.20028 \cdot 10^{-9} \\
1.62631 \cdot 10^{-7} & 4.23597 \cdot 10^{-8} & -1.82399 \cdot 10^{-10} \\
-1.20028 \cdot 10^{-9} & -1.82399 \cdot 10^{-10} & 9.12067 \cdot 10^{-13}
\end{pmatrix} \sigma^2
\]

The known terms of the normal equations are finally given by

\[
c_1 = \frac{1}{\sigma^2} \sum_{i=1}^{20} \phi_1(x_i)h_i = \frac{1}{\sigma^2} \sum_{i=1}^{20} T_i h_i = \frac{1}{\sigma^2} 4.70647 \cdot 10^7
\]
\[ c_2 = \frac{1}{\sigma^2} \sum_{i=1}^{20} \phi_2(x_i) h_i = \frac{1}{\sigma^2} \sum_{i=1}^{20} \frac{P_i}{T_i} h_i = \frac{1}{\sigma^2} 5.63486 \cdot 10^8 \]

\[ c_3 = \frac{1}{\sigma^2} \sum_{i=1}^{20} \phi_2(x_i) h_i = \frac{1}{\sigma^2} \sum_{i=1}^{20} P_i h_i = \frac{1}{\sigma^2} 1.74663 \cdot 10^{11} \]

The estimates of the regression parameters hold then

\[
\begin{pmatrix}
  a_1 \\
  a_2 \\
  a_3
\end{pmatrix} = F^{-1}\begin{pmatrix}
  c_1 \\
  c_2 \\
  c_3
\end{pmatrix} = \begin{pmatrix}
  29.04383301 \\
  -0.335066751 \\
  3.44115 \cdot 10^{-5}
\end{pmatrix}
\]

and the desired model writes therefore

\[ h(T, P) = 29.04383301 T - 0.335066751 \frac{P}{T} + 3.44115 \cdot 10^{-5} P \]

A graphical representation of the result follows.
where the graph of the model is the green 2-surface immersed in the space $\mathbb{R}^3$, while the black dots represent the experimental data.

The same model can be obtained by the below sequence of Maple 11 commands:

\[\text{with(Statistics)};\]

to load the Maple package "Statistics",

\[X := \text{Matrix}([[T_1, P_1], [T_2, P_2], \ldots [T_{20}, P_{20}]], \text{datatype } = \text{float});\]

to define the sampled set of independent variables $(T, P)$,

\[Y := \text{Array}([h_1, h_2, \ldots h_{20}], \text{datatype } = \text{float});\]

to introduce the the corresponding values of the dependent variable $h$,

\[V := \text{Array}([1, 1, 1, \ldots 1, 1]);\]

to assign all the data the same statistical weight 1 in the application of the least-squares procedure, and finally

\[\text{LinearFit}([T, P/T, P], X, Y, [T, P], \text{weights } = V);\]

to compute and display the regression model.

\[(ii)\) The confidence intervals of the parameters $\alpha_1$, $\alpha_2$, $\alpha_3$ with a confidence level of $1 - \eta = 0.95$ can be determined through the general formula

\[\alpha_j = a_j \pm t_{[1-\frac{\eta}{2}](n-p)}\sqrt{(F^{-1})_{jj} \frac{\text{NSSAR}}{n-p}}\]

\[j = 1, 2, 3,\]
by posing:
\[ \eta = 0.05 \]
\[ n - p = 20 - 3 = 17 \]

and inserting the numerical estimates

\[ t_{[1-\frac{2}{2}](n-p)} = t_{[0.975](17)} = 2.109815578 \]

\[
\text{NSSAR} = \frac{1}{\sigma^2} \sum_{i=1}^{20} \left[ -h_i + a_1 T_i + a_2 \frac{P_i}{T_i} + a_3 P_i \right]^2 = 7119.774073911 \frac{1}{\sigma^2}
\]

\[
(F^{-1})_{11} = 3.12439 \cdot 10^{-6} \sigma^2
\]

\[
(F^{-1})_{22} = 4.23597 \cdot 10^{-6} \sigma^2
\]

\[
(F^{-1})_{33} = 9.12067 \cdot 10^{-13} \sigma^2
\]

along with the estimates \(a_1, a_2, a_3\) of the parameters

\[ a_1 = 29.04383301 \quad a_2 = -0.335066751 \quad a_3 = 3.44115 \cdot 10^{-5} \]

Whence we obtain, for \(j = 1, 2, 3\),

\[ \alpha_j = a_j \pm 2.109815578 \sqrt{(F^{-1})_{jj} \frac{7119.774073911}{17}} \frac{1}{\sigma^2}, \]

and consequently

\[ \alpha_1 = 29.04383301 \pm 0.0763195 \]
\[ \alpha_2 = -0.335066751 \pm 0.00888647 \]
\[ \alpha_3 = (3.44115 \pm 4.12350) \cdot 10^{-5} \]
Notice the large uncertainty on the last parameter $\alpha_3$

(iii) The **CI for the prediction** $h_0$ of $h$ at an arbitrary point $x_0 = (T_0, P_0)$ of the independent variables is defined by the relationship

$$
\mathbb{E}(h_0) = h(T_0, P_0) \pm t_{[1-\frac{\eta}{2}]}(n-p) \sqrt{\text{var}(h_0) \frac{\text{NSSAR}}{n-p}}
$$

where the variance of the prediction $h_0$ can be expressed as

$$
\text{var}(h_0) = \sum_{j,k=1}^{3} \phi_j(x_0) \phi_k(x_0) (F^{-1})_{jk} + \sigma^2
$$

so that

$$
t_{[1-\frac{\eta}{2}]}(n-p) \sqrt{\text{var}(h_0) \frac{\text{NSSAR}}{n-p}} =
$$

$$
= t_{[1-\frac{\eta}{2}]}(n-p) \left[ \sum_{j,k=1}^{3} \phi_j(x_0) \phi_k(x_0) \frac{1}{\sigma^2} (F^{-1})_{jk} + 1 \right]^{1/2} \cdot \sqrt{\frac{\text{NSSAR} \sigma^2}{n-p}} =
$$

$$
= \left[ 1.86425914 \cdot 10^3 + 2.43942982 T_0^2 + 0.253954938 P_0 - 1.87428869 \cdot 10^{-3} T_0 P_0 + 3.30732012 \cdot 10^{-2} \frac{P_0^2}{T_0^2} - 2.84823259 \cdot 10^{-4} \frac{P_0^2}{T_0} + 7.12115332 \cdot 10^{-7} P_0^2 \right]^{1/2}
$$
is independent on the variance $\sigma^2$, as expected. In the figure below the lower and the upper limits of the 95%-CI for predictions at an arbitrary point $(T, P)$ are represented as a blue and a red surface, while the green surface is the graph of the model (as before).

The domain between the lower and the upper surface constitutes the confidence region of the regression model, at the assigned confidence level $1 - \eta = 95\%$. Notice that the two surfaces tend to divaricate at points $(T, P)$ far from the sampled domain, which means that the CI width increases: *we find again the same phenomenon already described for the regression straight line* (for better clarity, in the picture the CI half-width has been enlarged by a factor 5).
(iv) The 95%-CI for the prediction of the molar enthalpy \( h \) at \((T, P) = (345, 8.5 \cdot 10^5)\) can be calculated immediately by the previous formula, which leads to

\[
h(345, 8.5 \cdot 10^5) = 9223.845365 \pm 277.9010056
\]

or, equivalently, to

\[
8945.944359 \leq h(345, 8.5 \cdot 10^5) \leq 9501.746371
\]

(v) The goodness of fit of the model can be determined by means of the NSSAR:

\[
NSSAR = 7119.774073911 \frac{1}{\sigma^2} =
\]

\[
= 7119.774073911 \frac{1}{25^2} = 11.39163852
\]

from which we deduce the very satisfactory value

\[
Q = \lim_{\nu \to \infty} \int_{NSSAR}^{+\infty} p_{n-p}(\chi^2) d\chi^2 =
\]

\[
= \int_{11.39163852}^{+\infty} p_{17}(\chi^2) d\chi^2 = 0.8354957481
\]

The integral is easily calculated by the Maple 11 command

\[
Q := 1 - statevalf[cdf, chisquare[17]](11.39163852);
\]
7.10 Chi-square fitting by SVD

The chi-square method consists in determining the minimum of the quadratic functional (3)

$$L(\alpha_1, \ldots, \alpha_p) = \sum_{i=1}^{n} \left[ -\frac{1}{\sigma_i} y_i + \sum_{j=1}^{p} \alpha_j \frac{1}{\sigma_i} \phi_j(x_i) \right]^2$$

which, by means of the auxiliary matrices we have already introduced

$$\alpha = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_p \end{pmatrix} \quad \sigma^{-1} = \begin{pmatrix} 1/\sigma_1 & & \\ & 1/\sigma_2 & 0 \\ & 0 & \ddots \\ 0 & \cdots & 1/\sigma_n \end{pmatrix} \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

and

$$\Phi_{ih} = \Phi_{hi} = \frac{\phi_h(x_i)}{\sigma_i} \quad 1 \leq h \leq p, \ 1 \leq i \leq n,$$

takes the equivalent form

$$L(\alpha) = |\sigma^{-1}y - \Phi^T \alpha|^2$$

We have then to calculate the vector $$a \in \mathbb{R}^n$$ for which

$$|\sigma^{-1}y - \Phi^T a|^2 = \min_{\alpha \in \mathbb{R}^n} |\sigma^{-1}y - \Phi^T \alpha|^2$$

that is the solution in the sense of the least-squares of the linear set

$$\sigma^{-1}y = \Phi^T a$$

The problem can be solved, besides using the normal equation (7), by the **Singular Value Decomposition (SVD)** of the matrix $$\Phi^T$$
7.10.1 Singular Value Decomposition of a $m \times n$ matrix

The SVD of a matrix $A$ is the expression of $A$ as a product of the form

$$A = V \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix} UT$$

where:

- $U$ is a $n \times n$ matrix whose columns $u_1, \ldots, u_n$ constitute an orthonormal basis of eigenvectors of the $n \times n$, symmetric positive semidefinite matrix $A^T A$

$$U = (u_1 \ldots u_n) \quad A^T A u_i = \sigma_i^2 u_i \quad \forall i = 1, \ldots, n$$

$$\sigma_1^2 \geq \sigma_2^2 \geq \ldots \geq \sigma_r^2 > \sigma_{r+1}^2 = \ldots = \sigma_n^2 = 0$$

- $\Sigma$ is a diagonal $r \times r$ matrix, with $r \leq \min(m, n)$,

$$\Sigma = \begin{pmatrix} \sigma_1 \\ \sigma_2 \\ \vdots \\ \sigma_r \end{pmatrix}$$

whose diagonal elements are the arithmetical square roots of the positive eigenvalues of of $A^T A$ (so-called singular values of the matrix $A$)

- $V$ is a $m \times m$ matrix whose first $r$ columns are given by the orthonormal vectors

$$v_i = \frac{1}{|Au_i|^2} Au_i \quad \forall i = 1, \ldots, r$$

while the eventual residual $m - r$ columns are chosen in such a way that they form a orthonormal basis of $\mathbb{R}^m$
Remark: Any $m \times n$ matrix admits a SVD!

The SVD of $A$

$$A = V \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix} U^T$$

allows us to define the so-called pseudo-inverse of $A$ (Moore-Penrose pseudo-inverse)

$$A^+ = U \begin{pmatrix} \Sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix} V^T$$

Remark: $U^{-1} = U^T$ and $V^{-1} = V^T$, the matrices $U$ and $V$ are orthogonal by definition

If $A$ is invertible:

$$A^+ = A^{-1}$$

If not, $A^+$ is however defined

Application

Given the linear set

$$Ax = b,$$

the vector

$$x^+ = A^+ b$$

provides the solution in the sense of the least-squares

$$|Ax^+ - b|^2_2 = \min_{x \in \mathbb{R}^n} |Ax - b|^2_2$$

For an invertible $A$, $x^+$ is the solution in the usual sense
7.10.2 SVD and covariance matrix of the estimated best-fit parameters

The matrix \( \Phi_{ki} = \frac{\phi_k(x_i)}{\sigma_i}, \; k = 1, \ldots, p, \; i = 1, \ldots, n \), also admits a SVD

\[
\Phi = V \begin{pmatrix}
\Sigma & 0 \\
0 & 0
\end{pmatrix} U^T
\]

\[ p \times n \quad p \times p \quad p \times n \quad n \times n \]

and the structure matrix of the estimated best-fit parameters — a multivariate normal random variable — can be written

\[
F = \Phi \Phi^T = V \begin{pmatrix}
\Sigma^2 & 0 \\
0 & 0
\end{pmatrix} V^T
\]

where all the matrices on the right-handed side are \( p \times p \)

If \( \Phi \) has the maximum rank — \( p \) — there holds

\[
\begin{pmatrix}
\Sigma^2 & 0 \\
0 & 0
\end{pmatrix} = \Sigma^2
\]

where all the diagonal entries of \( \Sigma^2 \) are positive

Therefore the covariance matrix of the estimated best-fit parameters turns out to be

\[
F^{-1} = V(\Sigma^2)^{-1} V^T
\]
7.10.3 Example of SVD by Maple 11

The calculation is performed by the \textit{linalg} Maple Package:

\texttt{with(linalg);} \\

The matrix whose SVD we want to compute is defined by:

\[ A := \text{linalg[matrix]}(2, 3, [1, 2, 3, 4, 5, 6]); \]

The following line

\texttt{evalf(Svd(A));} \\

provides the singular values of \( A \):

\[ [9.508032001, 0.7728696357] \]

while the matrices of the left (\( V \)) and right (\( U \)) singular vectors are given by:

\texttt{evalf(Svd(A, V, U));} \\

The result is displayed by

\texttt{evalm(V);} \\

\[ \begin{bmatrix} -0.3863177031 & -0.9223657801 \\ -0.9223657801 & 0.3863177031 \end{bmatrix} \]

and

\texttt{evalm(U);} \\

\[ \begin{bmatrix} -0.4286671335 & 0.8059639086 & 0.4082482905 \\ -0.5663069188 & 0.1123824141 & -0.8164965809 \\ -0.7039467041 & -0.5811990804 & 0.4082482905 \end{bmatrix} \]
To check the result we simply have to define the $2 \times 3$ matrix of the singular values

$$S := array(1..2, 1..3, [[9.508032001, 0, 0], [0, 0.7728696357, 0]]);$$

$$S := \begin{bmatrix} 9.508032001 & 0 & 0 \\ 0 & 0.7728696357 & 0 \end{bmatrix}$$

and compute then the product $V S U^T$:

$$multiply(V, S, transpose(U));$$

$$\begin{bmatrix} 0.9999999999 & 2.000000000 & 3.000000000 \\ 3.9999999999 & 5.000000000 & 6.000000000 \end{bmatrix}$$

The matrix we obtained coincides with $A$ up to roundoff errors, due to floating point arithmetic. The verification is successful.
7.11 Nonlinear models

They are models where the dependence on the fitting parameters is **nonlinear**

Unlike the linear models, the determination of the best-fit parameters **cannot be obtained analytically**, by the solution of a set of **linear algebraic equations** (normal equations)

The best-fit parameters must be calculated by an appropriate numerical optimization algorithm (for the search of minima) The algorithm in of a recursive type

Optimization algorithms of current application (for search of minima):

- steepest descent

- inverse Hessian method

- Levenberg-Marquardt method

- conjugated gradient method

- relaxation method (Barbasin-Krasovskii)
For the probability distribution of NSSAR and the joint probability distribution of the best-fit parameters \( a_1, \ldots, p \) the same conclusions hold we have derived for linear models provided that the \( \sigma_i \)'s are small and that therefore the best-fit parameters vary only a little

If the probability distributions of the \( y_i \)'s are non-normal and well known, the previous results are no more applicable and it is convenient to use Monte-Carlo methods

As already stressed, the troublesome change is not from linear to nonlinear models, but from normal variables to non-normal \( y_i \)'s

This remark leads to the concept of “robust fitting”
7.11.1 Robust fitting

Generally speaking, a parameter estimate or a fitting are known as “robust” if they are a little sensible to “small” deviations from the hypotheses on which estimate or fitting are based.

These “small deviations” may be actually “small” and concern all data.

Or they may be “large” but concern only a small portion of the data we use for the estimate or fitting.

There are many kinds of “robust” estimates.

The most important ones are the so-called local $M$-estimates, based on the maximum likelihood principle.

These provide the method for the calculation of the best-fit parameters in the “robust fitting” algorithms.
Estimate of the best-fit parameters of a model by local $M$-estimates

If in a sequence of measurements the random errors are not normal, the maximum likelihood method for the estimate of the parameters $\alpha$ of a model $y(x, \alpha)$ requires that we maximize the quantity

$$ P = \prod_{i=1}^{n} e^{-\rho[y_i, y(x_i, \alpha)\]} $$

where the function $\rho$ is the natural logarithm — changed in sign — of the probability distribution of the variable $y_i$, with mean $y(x_i, \alpha)$

This is equivalent to minimize the sum

$$ \sum_{i=1}^{n} \rho[y_i, y(x_i, \alpha)] $$

The function often $\rho[y_i, y(x_i, \alpha)]$ depend only on the difference $y_i - y(x_i, \alpha)$, possibly scaled by a suitable factor $\sigma_i$ which can be assigned to each experimental point

In that case the $M$-estimate is called local

The maximum likelihood condition becomes then

$$ \min_{\alpha \in \mathbb{R}^p} \sum_{i=1}^{n} \rho\left(\frac{y_i - y(x_i, \alpha)}{\sigma_i}\right) $$
The minimum in $\alpha$ (best-fit estimate $a$ of the parameters $\alpha$) can be determined directly (the mosto common procedure)
or it can be characterized as a critical point, solution of the set of equations

$$
\sum_{i=1}^{n} \frac{1}{\sigma_i} \psi \left( \frac{y_i - y(x_i, \alpha)}{\sigma_i} \right) \frac{\partial y(x_i, \alpha)}{\partial a_j} = 0 \quad j = 1, \ldots, p
$$

where we have posed

$$
\psi(z) = \frac{d\rho(z)}{dz} \quad \forall z \in \mathbb{R}
$$

For **normal variables** the function involved in the best-fit procedure are

$$
\rho(z) = \frac{z^2}{2} \quad \text{and} \quad \phi(z) = z
$$

For **variables with bilateral Gaussian distribution** (or Laplace distribution)

$$
\text{Prob}[y_i - y(x_i, \alpha)] \sim \exp\left(-\frac{|y_i - y(x_i, \alpha)|}{\sigma_i}\right)
$$

we have instead

$$
\rho(z) = |z| \quad \text{and} \quad \phi(z) = \text{sgn}(z)
$$
If the variables follow a **Lorentz distribution** (or Cauchy distribution)

\[
\text{Prob}[y_i - y(x_i, \alpha)] \sim \frac{1}{1 + \frac{1}{2} \left( \frac{y_i - y(x_i, \alpha)}{\sigma_i} \right)^2}
\]

the functions useful for the best-fit estimate become

\[
\rho(z) = \ln \left( 1 + \frac{z^2}{2} \right) \quad \text{e} \quad \phi(z) = \frac{z}{1 + \frac{z^2}{2}}
\]

In the analysis of data sets with moderate outliers it may be convenient to make use of weight functions \( \psi \) which do not correspond to any typical probability distribution

**Example I. Andrew’s sine**

\[
\psi(z) = \begin{cases} 
\sin(z/c) & |z| < c\pi \\
0 & |z| \geq c\pi 
\end{cases}
\]

**Example II. Tukey’s biweight**

\[
\psi(z) = \begin{cases} 
z(1 - z^2/c^2)^2 & |z| < c \\
0 & z \geq c
\end{cases}
\]

If the distributions are approximately normal the best values of the constants \( c \) can be proved to be \( c = 2.1 \) e \( c = 6.0 \), in the two cases
7.12 Multidimensional cases

The problems of modeling in more, dependent or independent, variables can be formally treated as in the case of a unique independent variable and a single dependent variable.

For problems with **more independent variables**, the $k$-ples of independent variables are regarded as single variables.

The sums which define the merit functions are extended over all the available $k$-ples.

For problems with **more dependent variables**, the calculation of merit functions is performed by treating each $h$-ple of dependent variables as a vector of $\mathbb{R}^h$, whose module is considered (typically the usual Euclidean norm).

**An important case is that of large data matrices**

It is often investigated by the so-called **Principal Component Analysis (PCA)**.
7.12.1 Principal Component Analysis (PCA)

Let $X$ be a matrix of data, with $n$ rows and $p$ columns

We can assume that each of its rows corresponds to a point in the space $\mathbb{R}^p$ of vectors with $p$ components, one per each column.

The matrix $X$ specifies then a set of $n$ points in $\mathbb{R}^p$

$$X = \begin{pmatrix}
    x_1^T \\
    x_2^T \\
    \vdots \\
    x_n^T
\end{pmatrix} \quad x_i \in \mathbb{R}^p \quad \forall \ i = 1, \ldots, n$$

It may happen that the points $x_i$ are not distributed at random in $\mathbb{R}^p$, but along a particular subset of $\mathbb{R}^p$
In particular, let us assume that the data are described by a vector space $S_k$, of dimension $k \leq p$, spanned by an orthonormal set of vectors in $\mathbb{R}^p$

$$\{h_1, h_2, \ldots, h_k\}$$

$$h_j^T h_q = \delta_{jq} \forall j, q = 1, \ldots, k$$

Such vectors will form the columns of a matrix $p \times k$

$$H = \begin{pmatrix} h_1 & h_2 & \ldots & h_k \end{pmatrix}$$
The orthogonal projections of the vectors $x_i$’s on $\mathcal{S}_k$ are given by the column vectors

$$HH^T x_i \quad \forall i = 1, \ldots, n$$

or, equivalently, by the row vectors

$$x_i^T HH^T \quad \forall i = 1, \ldots, n$$

and the matrix of orthogonal projections holds therefore

$$XHH^T$$

so that

orthogonal projections of $x_i$’s on $\mathcal{S}_k$ \iff rows of $XHH^T$
The distances of the points $x_i$ from $S_k$ — the “residuals” — can be written

$$\delta_i = x_i - HH^T x_i = (I - HH^T) x_i \quad i = 1, \ldots, n$$

and the sum of their squares turns out to be

$$SS = \sum_{i=1}^{n} \delta_i^2 = \text{tr}[X(\mathbb{I} - HH^T)X^T] =$$

$$= \text{tr}[XX^T] - \text{tr}[XHH^TX^T] =$$

$$= \text{tr}[XX^T] - \text{tr}[H^TX^TXH]$$

**How can we determine the matrix $H$ — or $S_k$?**

**Least-squares:**
we require that the matrix $H$ — and therefore $S_k$ — is chosen in such a way that the residual sum of squares is minimum

$$SS \quad \text{minimum}$$

$$\updownarrow$$

$$\text{tr}[H^TX^TXH] \quad \text{maximum}$$

with the condition that the columns $h_1, \ldots, h_k$ of $H$ are orthonormal
Result
Let us consider an orthonormal basis of eigenvectors

\[ h_1, h_2, \ldots, h_p \]

doing the real, symmetric and semipositive defined matrix

\[ X^T X \]

with

\[ X^T X h_i = \lambda_i h_i \quad \forall i = 1, \ldots, p \]

and eigenvalues arranged in decreasing order

\[ \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0 \]

Then

- we can pose

\[ H = \left( \begin{array}{c|c|c} h_1 & h_2 & \ldots & h_k \end{array} \right) \]

- the corresponding (residual) sum of squares holds

\[
\text{tr}[X(\mathbb{I} - HH^T)X^T] = \sum_{i=k+1}^{p} \lambda_i
\]
If we conjecture that the data \( x_i \) are principally located along \( \mathbb{S}_k \), we replace the points \( x_i \) with their orthogonal projections on \( \mathbb{S}_k \)

\[
x_i \longrightarrow H H^T x_i
\]

so that the data matrix \( X \) is replaced with the \textit{k-principal component model}

\[
X H H^T = \sum_{j=1}^{k} (X h_j) h_j^T
\]

(so-called \textit{calibration} of the \textit{k}-principal component model)

The sum of squares of data ("deviance of the sample")

\[
\sum_{i=1}^{n} \sum_{j=1}^{p} x_{ij}^2 = \text{tr}(X^T X) = \sum_{j=1}^{p} \lambda_j
\]

is composed by a "deviance explained by the model"

\[
\text{tr}[X H H^T X^T] = \text{tr}[H^T X^T X H] = \sum_{j=1}^{k} \lambda_j
\]

and a "unexplained deviance" or (residual) sum of squares

\[
SS = \text{tr}[X (I - H H^T) X^T] = \sum_{j=k+1}^{p} \lambda_j
\]
If the $k$-principal component model is correct

all the data vectors $x \in \mathbb{R}^p$ can be represented as linear combinations of the orthonormal eigenvectors $h_1, \ldots, h_k$:

$$x = \sum_{j=1}^{k} \alpha_j h_j$$

with $\alpha_1, \ldots, \alpha_k \in \mathbb{R}$

This allows us to express $p - k$ components of $x$ as linear functions of the remaining $k$ components

For instance, the first $p - k$ as a function of the last $k$:

$$x_1 = f_1(x_{p-k+1}, \ldots, x_p)$$
$$x_2 = f_2(x_{p-k+1}, \ldots, x_p)$$
$$\vdots$$
$$x_{p-k} = f_{p-k}(x_{p-k+1}, \ldots, x_p)$$

The $k$-principal component model allows us to predict the values of $p - k$ components of the datum $x \in \mathbb{R}^p$, whenever the remaining $k$ components are known!

(use of the $k$-principal component model for prediction)
By comparison, we can easily verify that the $k$-principal component model is no more than a truncation to the largest $k$ singular values of the SVD of $X$

$$X = V \left( \sum_{0}^{\infty} \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \right) U^T = \sum_{j=1}^{r} \sigma_j v_j u_j^T$$

with:

- $r = \text{Rank}(X) = \text{Rank}(X^T X) \leq \min(n, p)$
- $U = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_p \end{pmatrix}$

$u_1, u_2, \ldots, u_p$ orthonormal basis of eigenvectors of $X^T X$

$$X^T X u_i = \sigma_i^2 u_i \quad i = 1, \ldots, p$$

according to the eigenvalues of $X^T X$ arranged in a decreasing order

$$\sigma_1^2 \geq \sigma_2^2 \geq \ldots \geq \sigma_p^2$$

- and

$$V = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}$$

$v_1, v_2, \ldots, v_n$ orthonormal basis of $\mathbb{R}^n$ such that

$$v_i = \frac{1}{\sigma_i} Xu_i \quad \forall i = 1, \ldots, r$$

Therefore, if $k \leq r$ — the only interesting case — we have

$$\sum_{j=1}^{k} \sigma_j v_j u_j^T = \sum_{j=1}^{k} \sigma_j \frac{1}{\sigma_j} (X u_j) u_j^T = \sum_{j=1}^{k} (X u_j) u_j^T$$

and to obtain the $k$-principal component model it is enough to pose $h_j = u_j \quad \forall j = 1, \ldots, k$. Moreover, there holds $\lambda_j = \sigma_j^2 \quad \forall j = 1, \ldots, k$. 

Stefano Siboni
The $k$-principal component model of the matrix $X$ is thus given by

$$X H H^T = \sum_{j=1}^{k} (X h_j) h_j^T$$

with:

- $h_i$ $i$-th principal component (loading)

and

- $X h_i$ $i$-th score

**Remark** The PCA is nothing else than a truncated SVD

The algorithms for the PCA are therefore the same of the SVD!
7.12.2 PCA for non-centred data (General PCA)

We can also attempt to model the data matrix $X$

by a **linear manifold of dimension** $k$

(or $k$-dimensional hyperplane of $\mathbb{R}^p$)

which **does not pass through the origin** $O$

and thus does not constitute a vector subspace of $\mathbb{R}^p$, like $S_k$

It can be proved that the result is still obtained by the previous procedure

provided that, however, the appropriate means are subtracted to all the points $x_i$

$$x_i - \overline{x} = x_i - \frac{1}{n} \sum_{q=1}^{n} x_q$$

or, equivalently, that the analysis is performed on the “centred” data matrix

$$X - NX = (I - N)X$$

where $N$ denotes the $n \times n$ matrix of constant entries

$$N_{ij} = \frac{1}{n} \quad \forall i, j = 1, \ldots, n$$

(The proof follows an argument similar to that of Huygens-Steiner theorem in theoretical mechanics)
8. Summary of useful Maple 11 and Excel commands

(•) Calculation of $t_{[\alpha]}(n)$
Maple 11: \[ \text{statevalf[icdf, students[n]]}(\alpha); \]
Excel: \[ \text{TINV}(2 - 2\alpha; n) \]

(•) Calculation of $\chi^2_{[\alpha]}(n)$
Maple 11: \[ \text{statevalf[icdf, chisquare[n]]}(\alpha); \]
Excel: \[ \text{CHIINV}(1 - \alpha; n) \]

(•) Calculation of $z_{[\alpha]}$
Maple 11: \[ \text{statevalf[icdf, normald[0, 1]]}(\alpha); \]
Excel: \[ \text{NORMINV}(\alpha; 0; 1) \]

(•) Calculation of $F_{[\alpha]}(n_1, n_2)$
Maple 11: \[ \text{statevalf[icdf, fratio[n_1, n_2]]}(\alpha); \]
Excel: \[ \text{FINV}(1 - \alpha; n_1; n_2) \]

(•) Goodness of fit $Q$ for a regression model with $n - p$ d.o.f.
Maple 11: \[ 1 - \text{stats[statevalf, cdf, chisquare[n - p]]}(\text{NSSAR}); \]
Excel: \[ \text{CHIDIST}(\text{NSSAR}; n - p) \]
# List of topics

1. Experimental errors ........................................... 1
2. Random variables ............................................. 4
2.1 General definitions and properties ......................... 4
2.2 Examples of discrete RVs .................................... 7
2.3 Examples of continuous RVs ................................ 8
2.4 Remark. Relevance of normal variables .................... 16
2.5 Probability distributions in more dimensions .......... 18
2.5.1 (Stochastically) independent RVs ....................... 19
2.5.2 (Stochastically) dependent RVs ......................... 19
2.5.3 Mean, covariance, correlation of a set of RVs .......... 20
2.5.4 Multivariate normal (or Gaussian) RVs ................. 22
3. Indirect measurements. Functions of RVs .................. 25
3.1 Generalities .................................................... 25
3.2 Linear combination of RVs .................................. 27
3.3 Linear combination of multivariate normal RVs ......... 28
3.4 Quadratic forms of standard variables ..................... 30
3.4.1 Craig theorem ............................................. 30
3.4.2 Characterization theorem of chi-square RVs .......... 30
3.4.3 Fisher-Cochran theorem .................................. 32
3.4.4 Two-variable cases ....................................... 33
3.5 Error propagation in indirect measurements .............. 34
3.5.1 Gauss law .................................................. 35
3.5.2 Logarithmic differential .................................. 36
3.5.3 Error propagation in solving a set
       of linear algebraic equations ............................ 38
3.5.4 Probability distribution of a function of RVs
       estimated by Monte-Carlo methods ....................... 44
4. Sample theory. Sample estimates of $\mu$ and $\sigma^2$ .... 48
4.1 Sample estimate of the mean $\mu$ .......................... 49
4.2  Sample estimate of the variance $\sigma^2$ .......................... 52
4.3  Normal RVs. Confidence intervals for $\mu$ and $\sigma^2$ .... 54
4.3.1 Confidence interval for the mean $\mu$ ......................... 56
4.3.2 Confidence interval for the variance $\sigma^2$ ................... 58
4.4  Large samples of an arbitrary population.
     Approximate confidence intervals of $\mu$ and $\sigma^2$ ....... 62
4.4.1 Confidence interval for the mean $\mu$ ......................... 62
4.4.2 Confidence interval for the variance $\sigma^2$ ................... 62
5.  Hypothesis testing .............................................. 70
5.1  $\chi^2$-test for adapting a distribution to a sample ..... 78
5.2  Kolmogorov-Smirnov test ................................. 86
5.2.1 Remark. Calculation of the KS variable ............... 94
5.2.2 Remark. The KS probability distribution is
     independent on the cumulative distribution $P(x)$ of
     the random variable on which the test is performed . 95
5.3  $t$-test on the mean of a normal population .......... 96
5.4  $\chi^2$-test on the variance of a normal population ...... 101
5.5  $z$-test to compare the means of two independent
     normal populations (of known variances) ............... 106
5.6  $F$-test to compare the variances of two independent
     normal populations ........................................ 107
5.7  Unpaired $t$-test to compare the means of two
     independent normal populations (unknown variances) 114
5.7.1 Case of equal variances: $\sigma_1^2 = \sigma_2^2 = \sigma^2$ ........ 114
5.7.2 Case of unequal variances: $\sigma_1^2 \neq \sigma_2^2$ .......... 115
5.7.3 Remark for the case $\sigma_1^2 = \sigma_2^2$. Proof that the test
     variable actually follows a Student’s distribution
     with $p + q - 2$ d.o.f. ................................. 116
5.8  Paired $t$-test to compare the means of two
     normal populations (unknown variances) ........... 120
5.9  $F$-test for the comparison of means ...................... 124
5.9.1 1-factor ANOVA .................................................. 128
5.9.2 2-factor ANOVA .................................................. 135
5.9.3 2-factor ANOVA without interaction ....................... 138
5.9.4 2-factor ANOVA with interaction .......................... 144
5.10 Sign test for the median of a population .................... 146
5.11 Sign test to check if two paired samples 
belong to the same population ................................. 148
5.12 Detection of outliers by Chauvenet criterion .......... 152
6. Pairs of random variables ........................................ 157
6.1 Linear correlation coefficient (Pearson r) ................. 157
6.2 Statistics of the linear correlation coefficient ............ 161
6.2.1 Independent random variables \((x, y)\) ................. 162
6.2.2 Normal random variables \((x, y)\) ......................... 163
6.2.2.1 Approximation for large \(n\). Fisher transformation ... 164
6.2.2.2 Independent (or uncorrelated) normal variables .. 164
6.3 Remark. .............................................................. 169
7. Data modelling ....................................................... 170
7.1 General setup of the problem .................................. 170
7.2 Linear regression by the chi-square method 
(chi-square fitting) .................................................. 174
7.2.1 Optimization. Normal equation and its solution .... 176
7.2.2 The parameters of the linear regression, 
estimated by chi-square fitting, as RVs .................... 177
7.2.3 Goodness of fit .................................................... 178
7.2.4 Remark. Standard deviations \(\sigma_i\) unknown .......... 181
7.3 \(F\)-test for the goodness of fit .............................. 182
7.4 \(F\)-test with repeated measurements on \(y\) ............ 187
7.4.1 Known variances \(\sigma^2_i\) ................................. 188
7.4.2 Equal (although unknown) variances \(\sigma^2_i\) ...... 190
7.5 Incremental \(F\)-test .............................................. 191
7.6 Confidence intervals for the regression parameters .. 193
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.7</td>
<td>Confidence intervals for predictions</td>
<td>193</td>
</tr>
<tr>
<td>7.8</td>
<td>Important case. Regression straight line</td>
<td>195</td>
</tr>
<tr>
<td>7.10</td>
<td>Chi-square fitting by Singular Value Decomposition</td>
<td>221</td>
</tr>
<tr>
<td>7.10.1</td>
<td>Singular Value Decomposition of a $m \times n$ matrix</td>
<td>222</td>
</tr>
<tr>
<td>7.10.2</td>
<td>SVD and covariance matrix of the estimated best-fit parameters</td>
<td>224</td>
</tr>
<tr>
<td>7.10.3</td>
<td>Example of SVD by Maple 11</td>
<td>225</td>
</tr>
<tr>
<td>7.11</td>
<td>Nonlinear models</td>
<td>227</td>
</tr>
<tr>
<td>7.11.1</td>
<td>Robust fitting</td>
<td>229</td>
</tr>
<tr>
<td>7.12</td>
<td>Multidimensional cases</td>
<td>233</td>
</tr>
<tr>
<td>7.12.1</td>
<td>Principal Component Analysis (PCA)</td>
<td>234</td>
</tr>
<tr>
<td>7.12.2</td>
<td>PCA for non-centred data (General PCA)</td>
<td>243</td>
</tr>
<tr>
<td>8</td>
<td>Summary of useful Maple 11 and Excel commands</td>
<td>244</td>
</tr>
</tbody>
</table>