Theory

Basic probability concepts

Random variables

Statistics is about the **extraction of information** from data that contain an unpredictable component.

Random variables (r.v.) are the mathematical devices employed to build *models* of this variability.

A r.v. takes a different value at *random* each time is observed.

> The aim of the model is to find a mathematical function that describes the randomness of the phenomenon.

The main tools used to describe the distribution of values taken by a r.v. are:

1. **Probability** (mass) **functions** (**pmf**)
   - describes the distribution of discrete r.v.
2. (Probability) **density functions** (**pdf**)
   - descrbes the distribution of continuous r.v.
3. **Cumulative distribution functions** (**cdf**)
   - describes both continuous and discrete r.v.
4. **Quantile functions**
   - describes both continuous and discrete r.v.

Notation

- r.v. → capital letters $(X, Y, ...)$
- observed values → lower letters $(x, y, ...)$
- every probability p is such that: $0 \leq p \leq 1$
- sample space: $S_X = \{1, ..., 6\}$

Discrete distributions

What's a discrete distribution

Discrete r.v. take values in a discrete set.
The probability (mass) function of a discrete r.v. X is the function f (x)
such that

$$f(x) = Pr(X = x)$$

with $0 \leq f(x) \leq 1$ and $\sum_i f(x_i) = 1$

The probability function defines the distribution of X

Examples

$$S_X = \{1, ..., 6\}$$
$$f(x_i) = P(X = x_i) = \frac{1}{6} \forall i$$
$$\sum_i f(x_i) = 1$$

combination of variables

$$W = \{X, Y\}$$
$$S_W = \{(1,1), (1,2), ..., (6,6)\}$$

Sum of variables

$$Z = (X + Y)$$
$$S_Z = \{2, ..., 12\}$$

```
# calculate all the combination between X and Y
# (with X = Y = {1, 2, 3, 4, 5, 6}) and store it in W
W <- expand.grid(1:6, 1:6)
# compute the row sum of W and store it in Z
Z <- rowSum(W)
# calculate the occurencies of the values in Z
table(Z)
# calculate the probability of each value in Z
table(Z)/length(Z)
# check if the sum is 1
sum(table(Z)/length(Z)) # = 1
# plot the distribution
plot(table(Z)/length(Z))
```

For many purposes, the first two moments of a distribution provide a
useful summary.

- The **mean** (expected value) of a discrete r.v. X is

$$E(X) = \sum_i x_i f(x_i)$$

and the definition is extended to any function g of X

$$E\{g(X)\} = \sum_i g(x_i) f(x_i)$$

The second mooment is:

$$g(X) = X^2 \qquad \Rightarrow \qquad E(X^2) = \sum_i x_i^2 f(x_i)$$

The special case $g(X) = (X - \mu)^2$, with $\mu = E(X)$, is the second centered moment, orjust **variance** of X:

$$Var(X) = E\{(X - \mu)^2\} = E(X^2) - \mu^2.$$

The **standard deviation** (σ) is just given by:

$$\sigma = \sqrt{Var(X)}$$

▼ Examples

X	$f(x)$
-1	0.05
0	0.8
1	0.15

$$E(X) = -1 \cdot 0.05 + 0 \cdot 0.8 + 1 \cdot 0.15 = 0.1$$
$$E(X^2) = (-1)^2 \cdot 0.05 + 0^2 \cdot 0.8 + 1^2 \cdot 0.15 = 0.2$$

$$Var(X) = E(X^2) - E(X)^2 =$$
$$(-1 - 0.1)^2 \cdot 0.05 + (0 - 0.1)^2 \cdot 0.8 + (1 - 0.1) \cdot 0.15 = 0.2035$$

▼ Binomial (and Bernoulli) distribution

Consider n independent binary trials each with success probability p, $0 < p < 1$. The r.v. X that counts the **number of successes** has ***binomial distribution*** with probability mass function:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, ..., n$$

The notation is:

- $X \sim B_i(n, p)$,

We have

D, D.

Bernoulli distribution

The case when $n = 1$ is known as ***Bernoulli distribution*** and a single binary trial is called ***Bernoulli trial***. The formula becomes:

$$P(X = x) = \binom{1}{x} p^x (1-p)^{1-x}, \qquad x = 0, 1$$

We have

D D

▼ Example

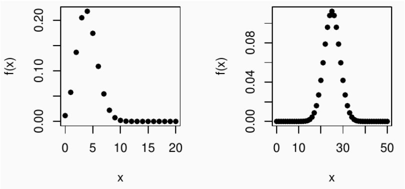
toss a coin 3 times, the probability of having 2 heads is:

$$P(X = 2) = \binom{3}{2} p^2 (1-p)$$

p is 0.5 so...

▼ R example

```
par(mfrow=c(1,2), pty="s", pch = 16)
plot(0:20, dbinom(0:20, 20, 0.2), xlab = "x", ylab = "f(x)")
plot(0:50, dbinom(0:50, 50, 0.5), xlab ="x", ylab = "f(x)")
```



the second one can be approximated to a normal distribution (we'll se later)

Poisson distribution

The special case the binomial distribution with $n \to \infty$ and $p \to 0$, while their product is held constant at $\lambda = np$, yields the Poisson distribution.

Used for counts of events that occur randomly over time when:

1. counts of events in disjoint periods are independent
2. it is essentially impossible to have two or more events simultaneously (disjoint)
3. the rate of occurrence is constant.

The probability function is

$$Pr(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \dots, \quad \lambda > 0$$

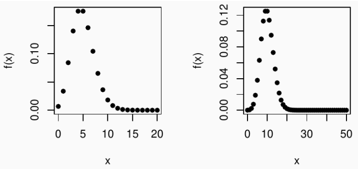
The notation is $X \sim P(\lambda)$.

We have:

$$E(X) = Var(X) = \lambda$$

Example

```
par(mfrow=c(1,2), pty="s", pch = 16)
plot(0:20, dpois(0:20, 5), xlab = "x", ylab = "f(x)")
plot(0:50, dpois(0:50, 10), xlab ="x", ylab = "f(x)")
```



Negative binomial (and Geometric) distribution

Let us consider a sequence of indipendent bernoulli trials with success probability p. Let X be the count of trials necessary to obtain the r^{th} success, then X has a ***Negative binomial*** (or ***Pascal***) distribution, with parameters p and r.

$$P(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-r} \qquad x = r, r+1, r+2, \dots$$

The notation is $X \sim \mathcal{NB}_i(p, r)$

we have:

It can also be defined with support the Natural number by simply considering the variable $Y = X - r$

Example

Disease with probability $p = 0.1$

We want to know how many people we need to meet so that we met $r = 10$ people with the disease.

$$E(X) = \frac{r}{p} = \frac{10}{0.1} = 100$$

Geometric distribution

the case for $r = 1$ is known as ***Geometric distribution:***

$$P(X = x) = (1-p)^{x-1}p$$

The notation is $X \sim \mathcal{G}(p)$

We have:

Continuous distributions

Key functions

Density function

Continuous r.v. take values from intervals on the real line.

The (**probability**) **density function** (p.d.f.) of a continuous r.v. X is the function
$f(x)$ such that, for any constants $a \le b$:

$$P(a \le X \le b) = \int_a^b f(x)dx$$

Note that $f(x) \ge 0$ and $\int_{-\infty}^{\infty} f(x)dx = 1$

The probability density function defines the **distribution** of X.

Mean and variance of a continuous r.v.

The definitions given in the discrete case are readily extended.

The mean (expected value) of a continuous r.v. X is

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

and the definition is extended to any function g of X

$$E\{g(X)\} = \int_{-\infty}^{\infty} g(x)f(x)dx$$

This includes the variance as a special case.

Two results, quite useful for continuous r.v., apply to a *linear transformation* $a + bX$, with a, b constants:

$$E(a + bX) = a + bE(X)$$
$$Var(a + bX) = b^2 \cdot Var(X)$$

Cumulative distribution function

The cumulative distribution function (C.D.F.) of a r.v. X is the function F(X) such that

$$F(X) = \Pr(X \le x)$$

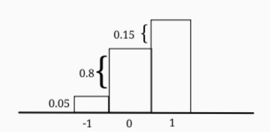
and it can be obtained from the probability function or the density function: the c.d.f. ***identifies*** the distribution.

From the definition of F it follows that $F(-\infty) = 0, F(\infty) = 1$ and $F(X)$ is monotonic.

A useful propriety is that F is a continuous function then $U = F(X)$ has a uniform distribution.

Examples

x	p
-1	0.05
0	0.8
1	0.15



$$P(X \le 0) = P(X = -1) + P(X = 0) = 0.85$$

CDF of a an exponential

$X \sim \exp(\lambda)$

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \ge 0 \\ 0 & x < 0 \end{cases}$$

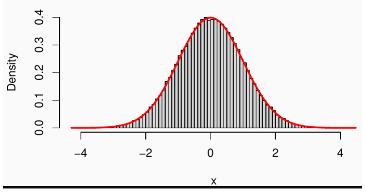
c.d.f:

$F(X \le x) = \int_0^x f(\xi)d\xi$

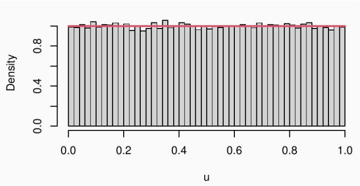
$$F(x) = \begin{cases} 1 - e^{-\lambda x} & x \ge 0 \\ 0 & x < 0 \end{cases}$$

R example

```
x <- rnorm(10^5) ### simulate values from N(0,1)
xx <- seq(min(x), max(x), l = 1000)
hist.scott(x, main = "") ### from MASS package
lines(xx, dnorm(xx), col = "red", lwd = 2)
```



```
u <- pnorm(x)  ### that's the uniform transformation
hist.scott(u, prob = TRUE, main="")
segments(0, 1, 1, 1, col = 2, lwd = 2)
```



$$Y = F_X(x)$$

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(F_X(X) \leq y) \\ &= P(F_x^{-1}(F_X(x)) \leq F_X^{-1}(y)) \\ &= P(X \leq F_X^{-1}(y)) \end{aligned}$$

▼ Quantile function

The inverse of the c.d.f. is defined as:

$$F^{-}(p) = \min(x|F(x) \geq p), \qquad 0 \leq p \leq 1$$

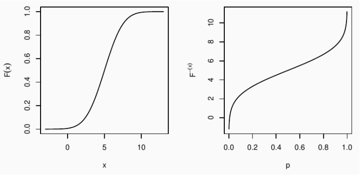
This is the usual inverse function of F when F is continuous.

Another useful property is that if $U \sim \mathcal{U}(0,1)$, namely it has a *uniform distribution* in $[0,1]$, then the r.v. $X = F^{-}(U)$ has c.d.f. F.

This provides a simple method to generate random numbers from a distribution with known quantile function: it is the **inversion sampling method,** that only requires the ability to simulate from a uniform distribution.

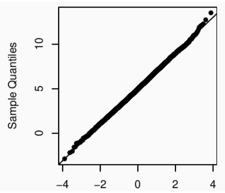
▼ Example

Let us consider the case of $X \sim \mathcal{N}(5, 2^2)$, with c.d.f. and quantile functions given by pnorm and qnorm



R lab

```
u <- runif(10^4); y <- qnorm(u, m = 5, s = 2)
par(pty = "s", cex = 0.8)
qqnorm(y, pch = 16, main = "")
qqline(y)
```



The previous slide demonstrated the usage of the quantile function to build a tool for **model goodness-of-fit**.

The quantile-quantile plot visualizes the plausibility of a theoretical distribution for a set of observations $y = (y_1, ..., y_n)$.

This is done by comparing the quantile function of the assumed model with the sample quantiles, which are the points that lie on the inverse of the **empirical distribution function**

$$F_n(t) = \frac{\text{number of elements of } y \leq t}{n}$$

If the agreement between the data and the theoretical distribution is good, the points on the plot would approximately lie on a line.

▼ The normal distribution

A r.v. has a ***normal*** distribution (or ***Gaussian***) distribution id it has p.d.f.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}, \quad -\infty < x < \infty$$

The notation is $X \sim N(\mu, \sigma^2)$,

D

$$\mu \in \mathbb{R}.$$

$$Var(X) = \sigma^2$$
$$\sigma^2 > 0,$$

An important property is that for any constants a, b:

$$a + bX \sim N(a + b\,\mu, b^2\sigma^2)$$

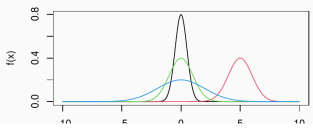
so that $Z = \dfrac{X - \mu}{\sigma} \sim N(0,1)$, the **standard normal distribution**.

Finally, $Y = e^X$ has a **lognormal distribution**, useful for asymmetric variables with occasional right-tail outliers.

▼ R lab

```
xx <- seq(-10, 10, l=1000)
plot(xx, dnorm(xx, 0, 0.5), xlab ="x", ylab ="f(x)", type ="l")
```

```
lines(xx, dnorm(xx, 5, 1), col = 2)
lines(xx, dnorm(xx, 0, 1), col = 3)
lines(xx, dnorm(xx, 0, 2), col = 4)
```



Gamma and exponential distribution

a r.v. X has Gamma distribution id it has the following pdf:

$$f(x) = \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}, \qquad x \geq 0$$

where $\lambda, \alpha > 0$ and $\Gamma(\alpha) = \int_0^\infty e^{-x} x^{\alpha-1} dx$.

The notation is $X \sim Ga(\alpha, \lambda)$

we have:

$$E(X) = \frac{\alpha}{\lambda} \qquad Var(X) = \frac{\alpha}{\lambda^2}$$

When α is an integer it is also called **Erlang distribution**.

Exponential distribution

When $\alpha = 1$ it is called **exponential distribution**. The exponential distribution is related to the Poisson r.v. since X **represents the waiting times between two arrivals in a Poisson process** (The process which generates the Poisson rv)

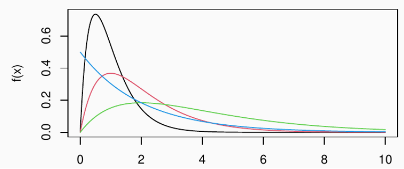
$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases}$$

We have:

$$E(X) = \frac{1}{\lambda} \qquad Var(X) = \frac{1}{\lambda^2}$$

R lab

```
xx <- seq(0, 10, l=1000)
plot(xx, dgamma(xx, 2, 2), xlab ="x", ylab ="f(x)", type ="l")
lines(xx, dgamma(xx, 2, 1), col = 2)
lines(xx, dgamma(xx, 2, .5), col = 3)
lines(xx, dgamma(xx, 1, .5), col = 4) # exponential distribution
```



Beta distribution

A r.v. X has a Beta distribution if it has the following pdf

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 < x < 1$$

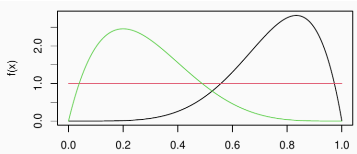
con $\alpha, \beta > 0$
The notation is
$X \sim Be(\alpha, \beta)$

$$E(X) = \frac{\alpha}{\alpha+\beta} \qquad Var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

The Uniform distribution on $[0, 1]$ is a special case when $\alpha = 1$ and $\beta = 1$.

R lab

```
xx <- seq(0, 1, l=1000)
plot(xx, dbeta(xx, 6,2), xlab ="x", ylab ="f(x)", type ="l")
lines(xx, dbeta(xx, 1,1), col = 2)
lines(xx, dbeta(xx, 2, 5), col = 3)
```



The χ^2 distribution

Let $Z_1, ..., Z_k$ be a set of independent $N(0, 1)$ r.v., then $X = \sum_{i=1}^{k} Z_i^2$ is a r.v. with a χ^2 distribution with k degrees of freedom.

The notation is X ~ χ^2

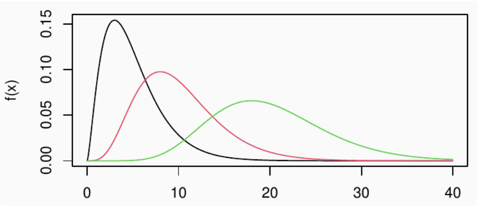
$$E(X) = k \qquad\qquad Var(X) = 2k$$

It is a special case of the Gamma distribution. In fact a χ^2 distribution with k degrees of freedom is a Gamma distribution with parameters $\alpha = k/2$ and $\lambda = 1/2$.

It plays an important role in the theory of hypothesis testing in statistics.

R lab

```
xx <- seq(0, 40, l=1000)
plot(xx, dchisq(xx, 5), xlab ="x", ylab ="f(x)", type ="l")
lines(xx, dchisq(xx, 10), col = 2)
lines(xx, dchisq(xx, 20), col = 3)
```



F distribution

Let $X \sim \chi^2$ and $Y \sim \chi^2$, independent, then the r.v.

$$F = \frac{X/n}{Y/m}$$

has an F distribution with n and m degrees of freedom.

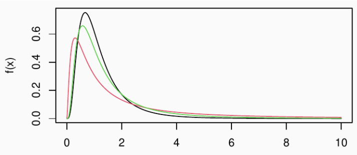
The notation is $F \sim \mathcal{F}_{n,m}$

$$E(F) = \frac{m}{m-2}, \qquad m > 2.$$

The distribution is almost never used as a model for observed data, but it has a central role in hypothesis testing involving linear models.

R lab

```
xx <- seq(0, 10, l=1000)
plot(xx, df(xx, 10, 10), xlab ="x", ylab ="f(x)", type ="l")
lines(xx, df(xx, 5, 2), col = 2)
lines(xx, df(xx, 10, 5), col = 3)
```



t and Cauchy distributions

Let $Z \sim \mathcal{N}(0, 1)$ and $X \sim \chi^2$, independent, then the r.v.

$$T = \frac{Z}{\sqrt{\frac{X}{n}}}$$

has an t distribution with n degrees of freedom.

The notation is $T \sim t_n$

$$Var(T) = \frac{n}{n-2}$$
$$n > 2$$

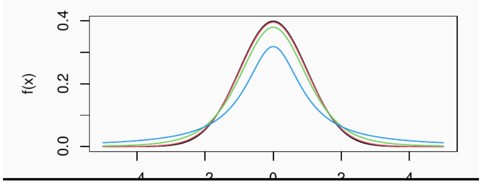
t_∞ is $\mathcal{N}(0, 1)$, while for n finite the distribution has heavier tails than the standard normal distribution.

The case t_1 is the **Cauchy distribution**.

The distribution has a central role in statistical inference; at times it is used for modelling phenomena presenting outliers.

▼ R lab

```
xx <- seq(-5, 5, l=1000)
plot(xx, dnorm(xx, 0, 1), xlab ="x", ylab ="f(x)", type ="l")
lines(xx, dt(xx, 30), col = 2)
lines(xx, dt(xx, 5), col = 3)
lines(xx, dt(xx, 1), col = 4)
```



▼ Random vectors

▼ Random vectors

$\mathbf{X_1}$	$\mathbf{X_2}$	$\cdots$
x_{11}	x_{21}	$\cdots$
x_{12}	x_{22}	$\cdots$
$\vdots$	$\vdots$	$\ddots$

In statistics, multiple variables are usually observed, and vectors of random variables (***random vectors***) are required. The two-dimensional case is useful to illustrate the main concepts and will be used here.

▼ Joint distribution

For continuous random variables (r.v.), the **joint (probability) density function** extends the one-dimensional case. It is the $f(x, y)$ function such that, for any $A \subseteq \mathbb{R}^2$:

$$\Pr\{(X, Y) \in A\} = \int\int_A f(x, y)\, dx\, dy$$

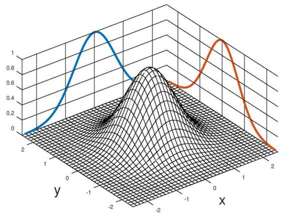
Note that:

- $f(x, y) \geq 0$
- $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1$

The probability density function defines the **joint distribution** of the random vector (X, Y).

▼ Example

$\mathbf{X}$	$\mathbf{Y}$
x_1	y_1
x_2	y_2
$\vdots$	$\vdots$



▼ Marginal distribution

The joint distribution embodies information about each components, so that the distribution of X, ignoring Y, can be obtained from $f(x, y)$.

The ***marginal density function*** of X is given by

$$f(x) = \int_{-\infty}^{\infty} f(x, y) dy$$

and similarly for the other variable.

> **Note**
>
> here and elsewhere we always use the symbol f for any p.d.f., identifying the specific case by the argument of the function).

▼ Conditional distribution

The conditional density function of Y given $X = x_0$ updates the distribution of Y by incorporating the information that $X = x_0$.

It is given by the important formula:

$$f(y|X = x_0) = \frac{f(x_0, y)}{f(x_0)}, \qquad \text{provide} \quad f(x_0) > 0$$

The simplified notation f (y |x0) is often employed.
The conditional p.d.f. is properly defined, since f (y |X = x0) ≥ 0 and
∞

R f (y |x0)dy = 1.
−∞

A symmetric definition applies to X given Y = y0 .

$$P(X|Y=y) = \left(\frac{P(X=x_1, Y=y_1)}{P(Y=y_1)}; \frac{P(X=x_2, Y=y_1)}{P(Y=y_1)}; \frac{P(X=x_3, Y=y_1)}{P(Y=y_1)}; \right)$$

Useful properties

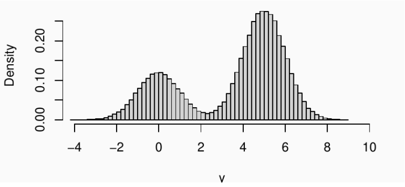
In the two dimensional case, it is readily possible to write:

Extensions to higher dimensions require some care:

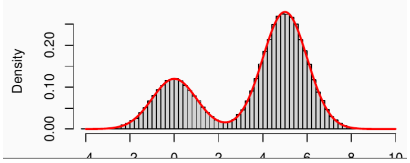
$$\begin{aligned} f(x,y,z) &= f(x,y|z)f(z) \\ f(x,y|z) &= f(x|z)f(y|x,z) \\ f(x,y,z) &= f(x|y,z)f(y,z) \\ f(x,y,z) &= f(x|y,z)f(y|z)f(z) \\ f(x_1,x_2,...,x_2) &= f(x_1)f(x_2|x_1)f(x_3|x_2,x_1)...f(x_n|x_{n-1},...,x_1) \end{aligned}$$

▼ R lab

```
x <- rbinom(10^5, size = 1, prob = 0.7)
y <- rnorm(10^5, m = x * 5, s = 1) ### Y| X = x ~ N(x * 5, 1)
hist.scott(y, main = "", xlim = c(-4, 10))
```



```
xx <- seq(-4, 10, l = 1000)
ff <- 0.3 * dnorm(xx, 0) + 0.7 * dnorm(xx, 5)
### This is a mixture of normal distributions
hist.scott(y, main = "", xlim = c(-4, 10))
lines(xx, ff, col = "red", lwd = 2)
```



▼ Bayes theorem

From the factorization of the joint distribution it readily follows that

$$f(x,y) = f(x)f(y|x) = f(y)f(x|y)$$

from which we obtain the **Bayes theorem**

$$f(x|y) = \frac{f(x)f(y|x)}{f(y)}$$

$\text{normalization factor} \uparrow$

This is a cornerstone of statistics, leading to an entire school of statistical modelling.

▼ Independence and conditional independence

When $f(y|x)$ does not depend on the value of x, the r.v. X and Y independent, and

$$f(x,y) = f(y)f(x)$$

More in general, n r.v. are independent if and only if:

$$f(x_1,x_2,...,x_n) = f(x_1)f(x_2)...f(x_n) = \prod_i f(x_i)$$

Conditional independence arises when two r.v. are independent given a third one:

$$f(y,x|z) = f(x|z)f(y|z)$$

An important part of statistical modelling exploits some sort of conditional independence.

The markov property

The general factorization defined above

$$f(x_1,x_2,...,x_n) = f(x_1)f(x_2|x_1)f(x_3|x_2,x_1)...f(x_n|x_{n-1},...,x_2,x_1)$$

will simplify considerably when the first order Markov property holds:

$$f(x_i|x_1,...,x_{i-1}) = f(x_i|x_{i-1})$$

which means that X_i is independent of $X_1,...,X_{i-2}$ given X_{i-1}. We get:

$$f(x_1,x_2,...,x_n) = f(x_1)\prod_{i=2}^{n} f(x_i|x_{i-1}).$$

When the variables are observed over time, this means that the process has short memory, a property quite useful in the statistical modelling of time series.

Mean and Variance of Linear Transformations

For two r.v. X and Y and two constants a, b, we get:

$$E(aX+bY) = aE(X)+bE(Y)$$

The result follows from the more general one:

$$E\{g(X,Y)\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)dxdy$$

For variances, we first need to introduce the covariance between X and Y:

$$\text{cov}(X,Y) = E\{(X-\mu_X)(Y-\mu_Y)\} = E(XY)-\mu_X\mu_Y$$

Where $\mu_X = E(X)$ and $\mu_Y = E(Y)$. Then:

$$\text{var}(aX+bY) = a^2\text{var}(X)+b^2\text{var}(Y)+2ab\,\text{cov}(X,Y)$$

> **Note**
>
> For X, Y independent, it follows that $\text{cov}(X,Y)=0$. The reverse is not true unless the joint distribution of X and Y is multivariate normal.

Mean Vector

For a random vector$X = (X_1, X_2, \ldots, X_n)^T$, the mean vector is:

$$E(X) = \begin{pmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_n) \end{pmatrix}$$

The mean vector has the same properties as the scalar case, so that, for example $E(X+Y) = E(X)+E(Y)$ and for A and b a $n \times n$ matrix and a $n \times 1$ vector, respectively, it follows that:

$$E(AX+b) = AE(X)+b$$

Variance-Covariance Matrix

The variance-covariance matrix of the random vector X collects all the variances (on the main diagonal) and all the pairwise covariances (off the main diagonal), being the $n \times n$ symmetric semi-definite matrix:

$$\Sigma = E\{(X-\mu_X)(X-\mu_X)^T\} = \begin{pmatrix} \text{var}(X_1) & \text{cov}(X_1,X_2) & \ldots & \text{cov}(X_1,X_n) \\ \text{cov}(X_1,X_2) & \text{var}(X_2) & \ldots & \text{cov}(X_2,X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(X_1,X_n) & \ldots & \ldots & \text{var}(X_n) \end{pmatrix}$$

Useful properties:

$$\begin{aligned} \Sigma_{AX+b} &= A\Sigma A^T \qquad (1) \\ \Sigma_{X^TAX} &= \mu_X^T A\mu_X + \text{tr}(A\Sigma) \qquad (2) \end{aligned}$$

Transformation of Random Variables and Random Vectors

Given a continuous r.v. X and a transformation $Y=g(X)$, with g an invertible function, it readily follows that:

$$f_Y(y) = f_X(g^{-1}(Y))\left|\frac{dx}{dy}\right|$$

The result is extended to two continuous random vectors with the same dimension:

$$f_Y(y) = f_X(g^{-1}(Y))|J|$$

With $J_{ij} = \frac{\partial x_i}{\partial y_j}$.

For discrete r.v., the results are simpler, with no need of including the Jacobian of the transformation.

The multivariate normal distribution

Start from a set of n i.i.d. (***Independent and identically distributed***) $Z_i \sim \mathcal{N}(0,1)$, so that $E(z)=0$ and covariance matrix I_n. If B is $m \times n$ matrix of coefficients and μ a m-vector of coefficients, then the m-dimensional random vector X:

$$X = BZ+\mu$$

Has a **multivariate normal distribution** with covariance matrix:

$$\Sigma = BB^T$$

The notation is:

$$X \sim \mathcal{N}(\mu, \Sigma)$$

Joint p.d.f.

Using basic results on transformation of random vectors, starting from the joint p.d.f of $Z_1, Z_2, ..., Z_n$ we obtain

$$f_X(X) = \frac{1}{\sqrt{(2\pi)^m |\Sigma|}} exp\left\{-\frac{1}{2}(X-\mu)^\top \Sigma^{-1}(X-\mu)\right\}, \qquad \text{for } X \in \mathbb{R}^m$$

provide that Σ has full rank m. The result can be extended to singular Σ by recourse to the pseudo-inverse of Σ: this is used, for example, in the analysis of compositional data.

A useful property which holds only for this distribution: two r.v. with multivariate normal distribution and zero covariance are independent.

Basic statistics

Random sample

The collection of r.v. $X_1, X_2, ..., X_n$ is said to be a **random sample** of size n if they are independent and identically distributed, that is

- $X_1, X_2, ..., X_n$ are independent r.v.
- They have the same distribution, namely the same c.d.f.

The concept is central in statistics, and it is the suitable mathematical model for the outcome of sampling units from a very large population. The definition is, however, more general.

(For more details: *https://www.probabilitycourse.com/chapter8/8_1_1_random_sampling.php*)

A statistic is a r.v. defined as a function of a set of r.v.
Obvious examples are the sample mean and variance of data $y_1, y_2, ..., y_n$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

($\bar{y}$ is also a variable which follows the normal distribution)

Consider a random vector Y with p.d.f. $f_\theta(Y)$ depending on a vector θ (which is the parameter, as we will see).

If a statistic $t(Y)$ is such that $f_\theta(Y)$ can be written as

(factorization theorem) where h does not depend on θ, and g depends on Y only through $t(Y)$, then t is a **sufficient statistic** for θ: all the information available on θ contained in Y is supplied by $t(Y)$.

$$if \;\; Y_1, \dots, Y_n \;\text{ are }\; i.i.d \quad \Rightarrow \quad f_\theta(Y) = \prod_{i=1}^{n} f_\theta(y_i)$$

The concepts of information and sufficiency are central in statistical inference.

Example: sufficient statistic for the normal distribution

Given a vector of independent normal r.v. $Y_i \sim N(\mu, \sigma^2)$, it follows that $\theta = (\mu, \sigma^2)$. We have:

$$\begin{gathered} f_\mu(Y) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu)^2\right\} = \\ = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}[(y_i - \bar{y}) - (\mu - \bar{y})]^2\right\} = \\ = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}[(y_i - \bar{y})^2 + (\mu - \bar{y})^2 - (y_i - \bar{y})(\mu - \bar{y})]\right\} = \\ = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i - \bar{y})^2 + \sum_{i=1}^{n}(\mu - \bar{y})^2\right]\right\} = \end{gathered}$$

Complements

Moment generating function

The moment generating function (m.g.f.) characterises the distribution of a r.v. X , and it is defined as:

$$M_X(t) = E(e^{tX}), \qquad t \in \mathbb{R}$$

The name derives from the fact the k^{th} th derivative of the m.g.f. at $t = 0$ gives the k^{th} uncentered moment:

$$\left.\frac{d^k M_X(t)}{dt^k}\right|_{t=0} = E(X^k)$$

Two useful properties:

- If $M_X(t) = M_Y(t)$ for some small interval around $t = 0$, then X and Y have the same distribution.
- If X and Y are independent, $M_{X+Y}(t) = M_X(t)M_Y(t)$.

The central limit theorem

For i.i.d. r.v. $X_1, X_2, ..., X_n$ with mean μ and finite variance σ^2, the central limit theorem states that for large n, the distribution of the r.v. $\overline{X}_n = \sum_{i=1}^{n} X_i/n$ is approximately

$$\overline{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$$

More formally, the theorem says that for any $x \in \mathbb{R}$ the c.d.f. of $Z_n = (\overline{X}_n - \mu)/\sqrt{\sigma^2/n}$ satisfies

$$\lim_{n\to\infty} F_{Z_n}(X) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$

The proof is simple, and it uses the m.g.f.

The theorem generalizes to multivariate and non-identical settings.

It has a central importance in statistics, since it supports normal approximation to the distribution of a r.v. that can be viewed as the sum of other r.v.

▼ The low of large numbers

Consider i.i.d. (independent and identically distributed) r.v. $X_1, ..., X_n$, with mean µ and $(E|X_i|) < \infty$.

The **strong law of large numbers** states that, for any positive $\mathcal{E}$

$$\Pr\left(\lim_{n\to\infty} |\overline{X}_n - \mu| < \mathcal{E}\right) = 1$$

namely $\overline{X}_n$ *converges almost surely* to µ.

With the further assumption $Var(X_i) = \sigma^2$, the **weak law of large numbers** follows:

$$\lim_{n\to\infty} \Pr(|\overline{X}_n - \mu| \geq \mathcal{E}) = 0$$

Proof

First we recall the Chebyshev's inequality:

given a r.v. X such that $E(X^2) < \infty$ and a constant $a > 0$, then

$$\Pr(|X| \geq a) \leq \frac{E(X^2)}{a^2}$$

We apply the inequality to the case of interest, so that

$$\Pr(|\overline{X}_n - \mu| \geq \mathcal{E}) \leq \frac{E\{(\overline{X}_n \text{-} \mu)^2\}}{\mathcal{E}^2} = \frac{Var(\overline{X}_n)}{\mathcal{E}^2} = \frac{\sigma^2}{n\mathcal{E}^2}$$

which tends to zero when $n \to \infty$.

The result may hold also for non-i.i.d. cases, provided $Var(\overline{X}_n) \to 0$ for large n.

▼ Jensen's inequality

Another useful result states that for a r.v. X and a **concave** function g:

$$g\{E(X)\} \geq E\{g(X)\}$$

> **Note**
>
> A concave function is such that
>
> $$g\{\alpha\, x_1 + (1-\alpha)\, x_2\} \geq \alpha\, g(x_1) + (1-\alpha)\, g(x_2)$$
>
> for any x_1, x_2 , and $0 \leq \alpha \leq 1$).

An example is $g(x) = -x^2$, so that

$$-E(X)^2 \geq -E(X^2) \quad \Rightarrow \quad E(X)^2 \leq E(X^2)$$

which is obviously true since $E(X^2) = Var(X) + E(X)^2$.

▼ Statistical models

▼ The concept of statistical model

Statistics aims to **extract information from data**, and in particular on the process that generated the data.

Two intrinsic difficulties:

- It may be hard to infer what we wish to know from the data available;
- Most data contain some **random variability**: by replicating the data-gathering process several times we would obtain different data on each occasion.

We search for conclusions drawn from a single data set that are **generally valid**, and not the result of random peculiarities of that data set.

▼ Example

Since the exponential distribution is such that $E(X) = \dfrac{1}{\lambda}$, we can estimate λ of a random vector using the sample mean $\overline{X}$ calculated on the data: $\lambda = \dfrac{1}{\overline{X}}$

Statistics is able to draw conclusions from random data mainly though the use of **statistical models**.

A statistical model can be thought as a mathematical cartoon describing how our data might have been generated, if the unknown features of the data-generating process were actually known.

If the unknowns were known, a good model *can generate data* resembling the main features of observed data.

The purpose of **statistical inference** is to use the statistical model to infer the model unknowns that are consistent with the observed data.

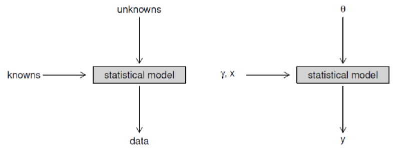
Notation

- $\mathbf{y}$ random vector containing the **observed data**
- $\boldsymbol{\theta}$ vector of parameters of **unknown value**

We assume that knowing the parameters would answer the question of interest about the process generating the data.

The model specifies how data akin to y may be simulated, implicitly defining the **distribution** of y and how it depends on θ.

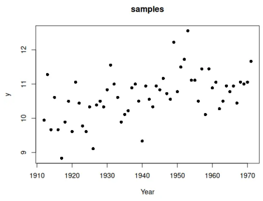
Moreover, a statistical model may depend on some known parameters γ and some further data x, treated as known and denoted as *covariates* or *predictor variables*.

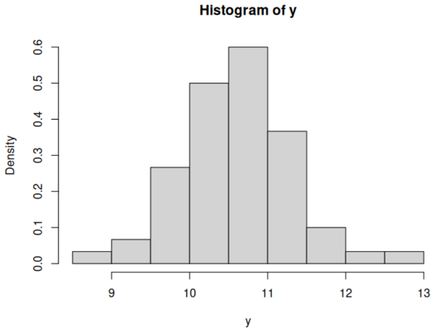


▼ **Example**

Consider the following record of 60 mean annual temperatures in New Haven, expressed in °C:

```
y <- (nhtemp - 32) / 1.8
plot(1912:1971, y, pch = 16, xlab = "Year", ylab = "y")
# let's draw de histogram and try to guess its distribution
hist(y, prob = T)
```

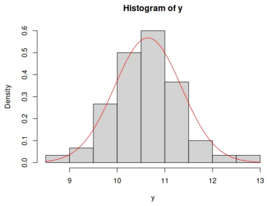




▼ **First model**

A first model simply assumes that the data are a random sample from a normal distribution namely they are the observation of i.i.d. r.v. from $\mathcal{N}(\mu, \sigma^2)$.

```
# look at the plot, it seams a normal distribution
# let's try to draw the curve of anormal over the histogram
curve(dnorm(x, mean(y), sqrt(var(y))), add = T, col = "red")
```



It follows that the distribution for the entire data vector is the product of the single contributions:

$$(y) = \prod_{i=1}^{n} \frac{1}{\sigma} \phi \left\{ \frac{(y_i - \mu)}{\sigma} \right\}$$

where ϕ is the $\mathcal{N}(0, 1)$ p.d.f.

▼ **Second model**

A second model retains the random sample assumption, but replaces the normal distribution with a heavier-tailed t5 distribution, assuming

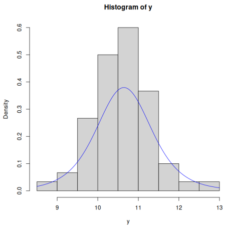
$$\frac{Y_i - \mu}{\sigma} \sim t_5$$

The distribution of the data becomes:

$$(y) = \prod_{i=1}^{n} \frac{1}{\sigma} f_{t_5} \left\{ \frac{(y_i - \mu)}{\sigma} \right\}$$

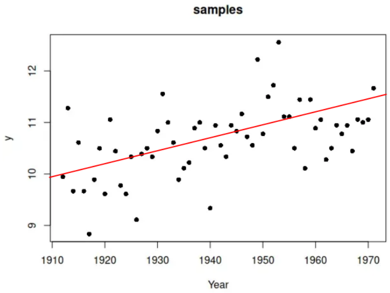
where f_{t_5} is the t_5 p.d.f.

```
curve(dt((x - mean(y))/sd(y), 5), add = T, col = "blue")
```



▼ Third model

The third model relaxes the assumption of identical distribution, assuming a linear trend over time:



after setting $t_i = year_i - 1911, i = 1, ..., 60$; we then take:

$$y_i = \beta_0 + \beta_1 t_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2)$$

The independence between observations still holds, so that:

$$(y) = \prod_{i=1}^{n} \frac{1}{\sigma} \phi \left\{ \frac{(y_i - \beta_o - \beta_1 t_i)}{\sigma} \right\}$$

▼ Fourth model

The last model maintains the trend assumption, but also includes autocorrelation for the error term, meaning that we assume

$$y_i = \beta_0 + \beta_1 t_i + \varepsilon_i, \qquad \varepsilon_i = \rho \varepsilon_{i-1} + v_i$$

with $v_i \sim N(0, \sigma^2)$, and the autocorrelation $\rho \in (-1, 1)$.

The model also requires to specify the distribution for the first observation, here taken as $Y_1 \sim N\{\beta_0, \sigma^2/(1-\rho^2)\}$, so that all the variables in the sample have the same variance.

The model is an instance of a linear regression model with autocorrelated errors. The r.v. of the sample are not longer independent, yet the distribution of Y can be found with some algebra.

It is possible to verify that Y is multivariate normal, with mean vector given by the linear trend

$$E(Y_i) = \mu_i = \beta_0 + \beta_1 t_i$$

and covariance matrix

$$\Sigma = \frac{\sigma^2}{(1-\rho^2)} \begin{pmatrix} 1 & \rho & \cdots & \rho^{n-1} \\ \rho & 1 & \cdots & \rho^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \cdots & 1 \end{pmatrix}$$

so that $f(y) = \phi_n(y; \mu, \Sigma)$, being ϕ_n the multivariate normal p.d.f.

It is useful to write down the vector parameters θ for each of the four model specifications proposed:

- Model 1: $\theta = (\mu, \sigma^2)$
- Model 2: $\theta = (\mu, \sigma^2)$
- Model 3: $\theta = (\beta_0, \beta_1, \sigma^2)$
- Model 4: $\theta = (\beta_0, \beta_1, \rho, \sigma^2)$

Note that the meaning of each parameter depends on the chosen model:

$\sigma^2 = Var(Y_i)$ in Model 1, but $\sigma^2 = 0.6\ Var(Y_i)$ in Model 2.

▼ Simulation from a statistical model

A decent model would allow to simulate data sets reproducing some of the features of the observed data, with better models providing more realistic results.

Simulation is an essential part of modern statistical inference. Its role is not only for the assessment of a candidate statistical model, but also to obtain **predictions** based on a chosen model.

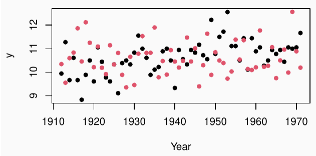
Simulation requires that a value for the model parameters θ is chosen beforehand. This task is accomplished by **parameter estimation.**

▼ Example (continue)

For Model 1, the parameters μ and σ^2 are readily estimated by $\bar{y}$ and s^2.

Then, a further dataset can be simulated using such values:

```
set.seed(2018); ysim <- rnorm(length(y), m = mean(y), s = sd(y))
plot(1912:1971, y, pch = 16, xlab = "Year", ylab = "y")
points(1912:1971, ysim, col = 2, pch = 16)
```



In order to evaluate whether the simulated dataset is similar to the observed one, we should focus on some important features.

For example, climate changes over time may suggest that the temperature of a given year may be ***positively* correlated** with the temperature of the subsequent year.

We can quantify this point by computing the sample autocorrelation:

$$r_1 = \frac{\sum_{i=1}^{n-1}(y_i - \bar{y})(y_{i+1} - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

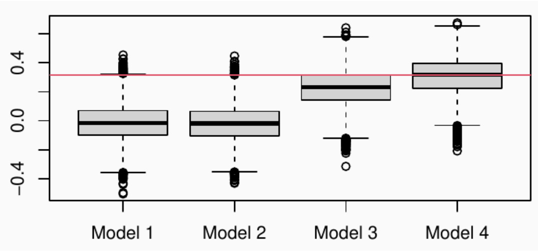
which is computed by the R function `acf`.

▼ Example (continue)

For the original data set $r_1 = 0.31$, whereas for the simulated data from Model 1 $r_1 = -0.12$.

This is just a single data set, though.

We simulate 10,000 samples from each of the four models, and each time we compute the r_1 coefficient. The sample distributions obtained are displayed in the plot below.



Model 4 is better at reproducing autocorrelation, as expected.

The example shows that no model gives a perfect fit for this data set, a fact that we ought to accept in broad generality.

Model 3 and Model 4 both provide an acceptable fit, with the latter slightly better in reproducing some of the autocorrelation observed in data.

Model diagnostic

Model diagnostics, a basic tool for model checking, it also has a role for simple models.

A basic tool is given by quantile-quantile plots, which can be used to verify whether the data y are consistent with an assumed model.

This is straightforward for i.i.d. models, (like Model 1 and 2 in the example), where the fact that the assumed distribution for y_i depends on μ and σ is rather inconsequential.

For more complex settings (such as Model 3 and 4 in the example), the general idea is as follows.

Assume that according to the fitted model the expected value and covariance matrix of y are $\mu_{\hat{\theta}}$ and $\Sigma_{\hat{\theta}}$.

Then the **standardized residuals** are:

$$\hat{\mathcal{E}} = \Sigma_{\hat{\theta}}^{-1/2}(y - \mu_{\hat{\theta}})$$

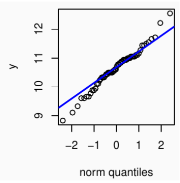
where $\Sigma_{\hat{\theta}}^{-1/2}$ is any matrix square root of $\Sigma_{\hat{\theta}}^{-1}$, such as its Choleski factor.

If the model is correct, ε should appear approximately independent, with zero mean and unit variance, and roughly normal if the model assumes normality.

▼ Example (continue)

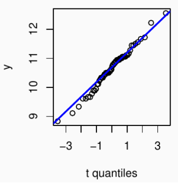
- **Model 1**

```
par(pty="s")
library(car)
qqPlot(y, dist="norm", envelope=FALSE, grid=FALSE, id=FALSE)
```



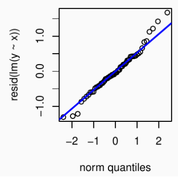
- **Model 2**

```
par(pty="s")
qqPlot(y, dist="t", df=5, envelope=FALSE, grid=FALSE, id=FALSE)
```



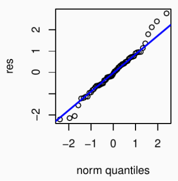
- **Model 3**

```
par(pty="s")
x <- 1912:1971-1911
qqPlot(resid(lm(y~x)), envelope=FALSE, grid=FALSE, id=FALSE)
```



- **Model 4**

```
par(pty="s")
qqPlot(res, dist="norm", envelope=FALSE, grid=FALSE, id=FALSE)
```



> More sophisticated models may give better results, but simple models conform to the **Occam's Razor principle**, that for statistical modelling argues in favor of ***simple models for simple problems***, moving to more complex models when simple models are inappropriate.

The problem of statistical inference

Inferential questions

Given a statistical model for data y, with model parameters θ, there are some basic questions to ask:

1. What values of θ are most consistent with y? [Point estimation]
2. What range of values of θ are consistent with y? [Interval estimation]
3. Is some prespecified restriction on θ consistent with y? [Hypothesis testing]
4. Is the model consistent with the data for any values of θ at all? [Model checking]

Question 4 can be enlarged to include which of several alternative models is most consistent with y? This is point of model selection, which partially overlaps with model checking.

The main challenge is to recognize the **intrinsic uncertainty** involved in attempting to understand θ.

For settings where some control over the data-gathering process is possible, a further question arises:

5. How might the data-gathering process be organized to produce data that enables answers to the preceding questions to be as accurate and precise as possible?

This is the core of *experimental and survey design methods*.

It represents an often neglected question, of central importance in many traditional fields where statistics is routinely applied (medical sciences, industrial research, biosciences ...). It is also very relevant for business and web analytics, like in A/B testing.

Approaches to statistical inference

There two classes of methods providing an answer to questions 1-4, namely the **frequentist** and **Bayesian** approach.

The main difference between the two approaches is in the role of model parameters θ, which are treated as fixed constants in the first approach and as a random variable in the latter one.

The difference may appear remarkable, and there has been controversy over the years about the merits of each approach.

Yet, from a a practical perspective the two approaches have much in common, and tend to give similar answers when properly applied, especially when compared to approaches that are not based on a statistical model.

Estimation

Point estimation

Given a model for the data y, with parameter θ, **point estimation** is concerned with finding a reasonable parameter estimate from the data.

There are several methods for doing this, and the problem can be simply stated as finding the parameter value most consistent with the data, a definition that leads to the method of **maximum likelihood estimation**.

We will delve into the details of maximum likelihood estimation in due time, but here we focus on some general aspects of point estimation.

- **Example (sample mean and variance)**

  A simple model assumes that the data are a random sample from a normal distribution namely they are the observations of i.i.d. r.v. from $\mathcal{N}(\mu, \sigma^2)$.

  Straightforward estimates of μ and σ^2 are given by the **sample mean**:

$$\hat{\mu} = \hat{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

  and by the **sample variance**:

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

  Such estimates are actually sensible anytime we are interested in estimating the mean and variance of an i.i.d. sample.

Properties

To figure out what could be a good estimate, we need to consider repeated estimation under repeated replication of the data-generating process.

This makes fully sense whenever the available data are a random sample obtained from a large population, like in industrial or social surveys, so that it would perfectly possible to iterate the sampling and obtain further data with the same structure of y.

However, we apply the same logic even when repetition is just the result of an idealization. The point is: what do we expect to find if we repeat the same analysis to many data sets generated from the same model?

- **Unbiasedness**

  If we replicate the random data and we repeat the estimation process, the result will be a different value of $\hat{\theta}$ for each replicate.

  The values are observations of a random vector, the **estimator** of θ, which is usually also denoted by $\hat{\theta}$ (the context will make clear whether we are referring to the estimator or to the estimate for a given sample).

  Since, the estimator is a r.v., it makes fully sense to compute its mean.

  For an **unbiased** estimator

$$E(\hat{\theta}) = \theta$$

  Unbiasedness is a desirable property, and we would also like the estimator to have **low variance**.

  The sample mean is an unbiased estimator

$$E(\bar{Y}) = E\frac{(\sum_i Y_i)}{n} = \sum_i \frac{E(Y_i)}{n} = \sum \frac{\mu}{n} = \mu$$

  Explicit bias calculation:

$$Bias = E(\bar{Y}) - \mu = \mu - \mu = 0$$

  The estimated variance is not an unbiased estimator

$$E(\hat{\sigma}^2) = E\left[\frac{\sum(y_i - \bar{y})^2}{n}\right] = E\left[\frac{\sum[(y_i - \mu) - (\bar{y} - \mu)]^2}{n}\right] = E\left[\frac{\sum[(y_i - \mu)^2 + (\bar{y} - \mu)^2 - 2(y_i - \mu)(\bar{y} - \mu)]}{n}\right] = E\left[\frac{\sum(y_i - \mu)^2}{n} + \frac{\sum(\bar{y} - \mu)^2}{n} - \frac{2\sum(y_i - \mu)(\bar{y} -}{n}$$

  The last term can be written as:

$$\frac{2(\bar{y} - \mu)}{n}\left(\sum y_i - n\mu\right) = -2(\bar{y} - \mu)^2$$

  So we get:

$$E(\hat{\sigma}^2) = E\left[\frac{\sum(y_i - \mu)^2}{n} - (\bar{y} - \mu)^2\right]$$

  we want to express $E(\hat{\sigma}^2)$ in terms of σ^2 so that we can adjust for bias.

$$\begin{gathered} = E\left[\frac{\sum(y_i - \mu)^2}{n}\right] - E(\bar{y} - \mu)^2 = \\ \frac{\sum Var(Y_i)}{n} - Var(\bar{Y}) \quad = \quad \cancel{n} \cdot \frac{\sigma^2}{\cancel{n}} \cdot + \frac{\sigma^2}{n} = \\ \sigma^2\left(1 - \frac{1}{n}\right) \end{gathered}$$

  Since $E(\hat{\sigma}^2)$ is biased by a factor of $\left(1 - \frac{1}{n}\right)$, we can correct this bias by multiplying $\hat{\sigma}^2$ by $\frac{n}{n-1}$. The unbiased estimator of the variance, often denoted by s^2, is therefore:

$$\boxed{s^2 = \tfrac{n}{n-1}\hat{\sigma}^2 = \tfrac{1}{n-1}\ \textstyle\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

  Thus, s^2 is an unbiased estimator of the population variance σ^2.

- **Consistency**

  Mean Squared Error

  There is tradeoff between unbiasedness and low variance, so we usually seek to get both (to some extent): ideally we would target a small Mean Squared Error (MSE):

$$MSE(\hat{\theta}) = E\{(\hat{\theta} - \theta)^2\}$$

With some algebra, we obtain:

$$MSE(\hat{\theta}) = \{E(\hat{\theta}) - \theta\}^2 + Var(\hat{\theta}) = \text{Squared bias} + \text{Variance}$$

▼ **Example: normal random sample**

For a normal random sample, it is straightforward to verify that

- $E(\bar{Y}) = \mu$
- $\text{Var}(\bar{Y}) = \dfrac{\sigma^2}{n} = \text{MSE}(\bar{Y})$

For the sample variance, we use the property that:

$$\frac{(n-1)}{\sigma^2} S^2 \sim \chi^2_{n-1}$$

to obtain:

- $E(S^2) = \sigma^2$
- $\text{Var}(S^2) = \dfrac{2\sigma^4}{n-1} = \text{MSE}(S^2)$

The unbiasedness of the sample mean and variance is a general property, holding also for non-normal samples.

Consistency

A (scalar) estimator is said to be **(weakly) consistent** if, for any $\epsilon > 0$:

$$\Pr(|\hat{\theta}_n - \theta| > \epsilon) \to 0 \text{ as } n \to \infty.$$

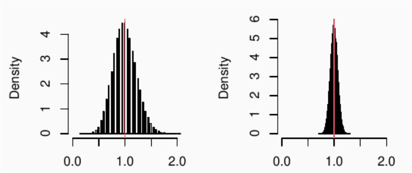
A sufficient condition for this is that $\text{MSE}(\hat{\theta}) \to 0$ for large samples, which requires that both bias and variance become negligible.

The law of large samples implies that the sample mean is a consistent estimator for the true mean in random samples.

▼ **Example: R lab**

```
M <- 100000; n1 <- 20; n2 <- 200; y1 <- y2 <- rep(NA, M)
for(i in 1:M) {y1[i] <- mean(rpois(n1, 1))
                             y2[i] <- mean(rpois(n2, 1))}

par(mfrow=c(1,2))
hist.scott(y1, xlim=c(0,2), main="", xlab=""); abline(v=1,col=2)
hist.scott(y2, xlim=c(0,2), main="", xlab=""); abline(v=1,col=2)
```



▼ **Efficiency**

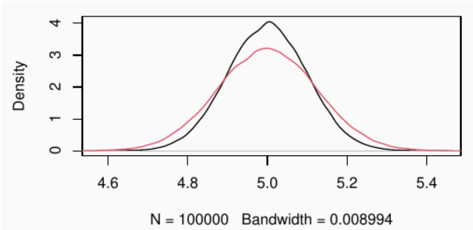
An **efficient estimator** is an estimator that estimates the parameter of interest in some optimal manner.

Among estimators with negligible bias, efficiency is associated to small variance. Since this is the case of consistent estimators, they are usually compared in terms of their variance.

▼ **Example: R lab**

For a normal random sample, both the sample mean and sample median are consistent estimators of μ. The mean is more efficient.

```
M <- 100000; n <- 100; mat.y <- matrix(NA, nrow = M, ncol = 2)
for(i in 1:M) {y <- rnorm(n, 5)
                             mat.y[i,] <- c(mean(y), median(y))}
plot(density(mat.y[,1]), type="l", main="")
lines(density(mat.y[,2]), col=2)
```



▼ **Standard error**

An important quantity defined for a (scalar) estimator is given by its **standard error**, defined as:

$$\text{SE}(\hat{\theta}) = \sqrt{\text{Var}(\hat{\theta})}$$

Once a sample is observed, and a numerical estimate of θ is obtained, the estimated standard error is obtained by replacing θ with $\hat{\theta}$.

An example is the **standard error of the mean** $\mathrm{SE}(\bar{Y}) = \frac{\sigma}{\sqrt{n}}$, which is estimated by $\frac{s}{\sqrt{n}}$.

In applications, the estimated standard error is routinely reported along with the estimate, since it quantifies the **estimation precision**.

The Delta Method

The **Delta Method** is a powerful tool in statistical inference for approximating the standard error of a function of an estimator.

Basic Setup

Suppose we are interested in a parameter that is a function of an underlying parameter θ. Specifically, let:

$$\psi = g(\theta)$$

where g is a continuous and differentiable function. If we have a consistent estimator $\hat{\theta}$ for θ, the **continuous mapping theorem** guarantees that $g(\hat{\theta})$ will also be a consistent estimator for $g(\theta)$, meaning that $g(\hat{\theta})$ will converge in probability to $g(\theta)$ as the sample size grows.

Approximation of the Standard Error

The **delta method** is used to approximate the standard error of $g(\hat{\theta})$. Using a first-order Taylor expansion around θ, we can approximate $g(\hat{\theta})$ by:

$$g(\hat{\theta}) \approx g(\theta) + g'(\theta)(\hat{\theta} - \theta)$$

This linear approximation allows us to derive the standard error of $g(\hat{\theta})$ in terms of the standard error of $\hat{\theta}$, denoted as $\mathrm{SE}(\hat{\theta})$, giving:

$$\mathrm{SE}(\hat{\psi}) \approx \mathrm{SE}(\hat{\theta}) \cdot |g'(\hat{\theta})|$$

This approximation becomes more accurate as the sample size increases, as larger samples reduce the approximation error in the Taylor expansion.

Extension to Multiple Parameters

The delta method can be extended to functions of multiple parameters. Suppose we are estimating a vector parameter $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)^\top$ and are interested in a function $g(\boldsymbol{\theta})$. If $\hat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}$, then the gradient (Jacobian) vector $\nabla g(\boldsymbol{\theta})$ allows us to approximate the variance of $g(\hat{\boldsymbol{\theta}})$. The standard error of $g(\hat{\boldsymbol{\theta}})$ is then approximately:

$$\mathrm{Var}(g(\hat{\boldsymbol{\theta}})) \approx \nabla g(\boldsymbol{\theta})^\top \, \mathrm{Var}(\hat{\boldsymbol{\theta}}) \, \nabla g(\boldsymbol{\theta})$$

where $\mathrm{Var}(\hat{\boldsymbol{\theta}})$ is the covariance matrix of $\hat{\boldsymbol{\theta}}$.

Robust Estimation

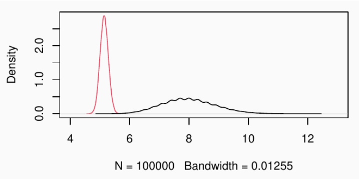
A **robust** estimator has good performances across a wide range of statistical models for the data.

The **sample median** is a robust estimation of location, not affected by possible outlying data, quite the opposite of the sample mean.

Robust estimation trades some efficiency with resistance to outliers, and they are often a sensible choice for semi-automatic data analyses.

Example: R Lab (Robustness of the Sample Median)

```
M <- 100000; n <- 100; mat.y <- matrix(NA, nrow = M, ncol = 2)
for(i in 1:M) {
    x <- rbinom(n, size = 1, prob = 0.9)
    y <- x * rnorm(n, 5) + (1 - x) * rnorm(n, 35)
    mat.y[i, ] <- c(mean(y), median(y))
}
plot(density(mat.y[, 2]), type = "l", main = "", xlim = c(4, 13), col = 2)
lines(density(mat.y[, 1]), col = 1)
```



> Unbiasedness is not a crucial property of the estimator (like consistency is), becouse for large samples, the bias becomes insignificant.

Interval estimation

The Aim of Interval Estimation

Confidence intervals provide a more comprehensive approach to estimation than relying on point estimates alone. Instead of giving a single value, they offer a **range of plausible values** for the model parameter, thereby reflecting the uncertainty in the estimation process.

Typically, confidence intervals focus on a single parameter, although it's possible to extend the concept to ***multidimensional confidence regions***. However, these extensions are rarely used in practice due to their complexity and limited interpretability.

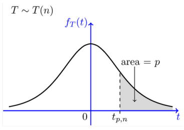
Pivots in Confidence Intervals

A key concept in constructing confidence intervals is the use of **pivots**, special functions that combine the observed data with the unknown parameter in a way that results in a known probability distribution.

If we have a random sample from a $N(\mu, \sigma^2)$ distribution, with σ^2 **unknown**, the following pivot can be used to estimate the mean μ:

Let X_1, X_2, ..., X_n be i.i.d. $N(\mu, \sigma^2)$ random variables. Also, let S^2 be the sample variance for this random sample. Then, the random variable T defined as:

$$T(\mu) = \frac{\overline{Y} - \mu}{\sqrt{\frac{S^2}{n}}} \sim t_{n-1}, \qquad \forall \mu \in \mathbb{R},\ \sigma^2 > 0$$



so $T(\mu)$ has a t-distribution with *n* − 1 degrees of freedom.

This pivot has a known distribution (t_{n-1}), which allows us to derive the confidence interval for μ.

Confidence Interval

Using the pivot for the normal random sample, we can derive a confidence interval for μ by setting up the probability statement:

$$\Pr\left(t_{n-1;\frac{\alpha}{2}} \leq T(\mu) \leq t_{n-1;1-\frac{\alpha}{2}}\right) = 1-\alpha$$

After a few algebraic manipulations, this leads to:

$$\Pr\left(\bar{Y} - t_{n-1;1-\alpha/2}\sqrt{\frac{S^2}{n}} \leq \mu \leq \bar{Y} + t_{n-1;1-\alpha/2}\sqrt{\frac{S^2}{n}}\right) = 1-\alpha$$

This probability statement implies that the interval:

$$\left(\bar{Y} - t_{n-1;1-\alpha/2}\sqrt{\frac{S^2}{n}}, \quad \bar{Y} + t_{n-1;1-\alpha/2}\sqrt{\frac{S^2}{n}}\right)$$

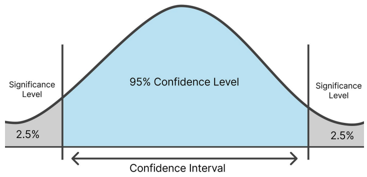
contains the true mean μ with probability $(1-\alpha)$. This interval is called a $(1-\alpha)\times 100\%$ **confidence interval**.

> Common choices for confidence levels being 95% and 99%.

Interpreting

For a specific dataset $\{y_1, \ldots, y_n\}$, we calculate the confidence interval by substituting $\bar{Y}$ and S^2 with their sample values, $\bar{y}$ and s^2:

$$\left(\bar{y} - t_{n-1;1-\alpha/2}\sqrt{\frac{s^2}{n}}, \quad \bar{y} + t_{n-1;1-\alpha/2}\sqrt{\frac{s^2}{n}}\right)$$



This interval either contains the true parameter μ or it does not. However, the probability interpretation is tied to a **hypothetical repetition** of data samples: across numerous samples, we expect that $(1-\alpha)\times 100\%$ of the confidence intervals constructed will contain μ.

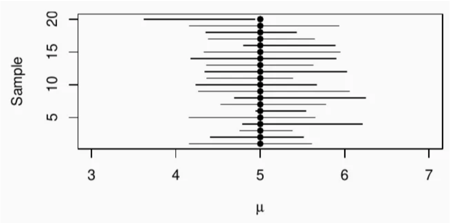
R Lab: Simulating Confidence Intervals

In R, we can simulate confidence intervals to explore their properties. The following code snippet generates 100,000 confidence intervals and calculates the proportion that contains the true mean of 5:

```
M <- 100000; n <- 10; mat.ci <- matrix(NA, nrow = M, ncol = 2)
for(i in 1:M) {
  y <- rnorm(n, 5)
  se_t <- sqrt(var(y) / n) * qt(0.975, n-1)
  mat.ci[i,] <- mean(y) + se_t * c(-1, 1)
}
mean(mat.ci[,1] < 5 & mat.ci[,2] > 5)
## [1] 0.94909
```

We can visualize the first 20 simulated confidence intervals, expecting that (on average) 19 out of 20 will include the true μ

```
plot(rep(5, 20), 1:20, pch = 16, ylab="Sample", xlab=expression(mu))
for(i in 1:20) segments(mat.ci[i,1],i,mat.ci[i,2],i)
```



One-Sided Confidence Intervals

Standard confidence intervals are usually **two-sided**, but it's possible to create **one-sided intervals** by adjusting the tail probabilities. For example:

$$\Pr\left(\bar{Y} - t_{n-1;1-\alpha_1}\sqrt{\frac{S^2}{n}} \leq \mu \leq \bar{Y} + t_{n-1;1-\alpha_2}\sqrt{\frac{S^2}{n}}\right) = 1-\alpha$$

where $\alpha_1 + \alpha_2 = \alpha$. Setting $\alpha_1 = 0$ makes the lower bound $-\infty$, and $\alpha_2 = 0$ makes the upper bound ∞, providing **one-sided confidence intervals** useful in certain applications.

Approximate Confidence Intervals & Coverage Probability

In many cases, an **exact pivot** is difficult to find, but approximate pivots are widely applicable. For large samples, a common approach is the **Wald pivot** for a generic parameter of interest ψ, based on a consistent estimator, which is approximately normally distributed for large samples:

$$Z(\hat{\psi}) = \frac{\hat{\psi} - \psi}{SE(\hat{\psi})} \dot{\sim} \mathcal{N}(0,1), \quad \forall \psi \in \Psi$$

This leads to the confidence interval:

$$\left(\hat{\psi} - z_{1-\alpha/2} SE(\hat{\psi}), \quad \hat{\psi} + z_{1-\alpha/2} SE(\hat{\psi})\right)$$

The Central Limit Theorem provides such a solution for random samples, when ψ corresponds to the mean of each variable.

▼ R Lab: Approximate Confidence Interval Simulation

In the following R code, we simulate approximate confidence intervals for the mean, examining the interval's accuracy as sample size increases:

```
M <- 100000; n <- 10; mat.ci <- matrix(NA, nrow = M, ncol = 2)
for(i in 1:M) {
  y <- rnorm(n, 5)
  se_z <- sqrt(var(y) / n) * qnorm(0.975)
  mat.ci[i,] <- mean(y) + se_z * c(-1, 1)
}
mean(mat.ci[,1] < 5 & mat.ci[,2] > 5)
## [1] 0.91904
```

```
M <- 100000; n <- 100; mat.ci <- matrix(NA, nrow = M, ncol = 2)
for(i in 1:M) { y <- rnorm(n, 5)
se_z <- sqrt(var(y) / n) * qnorm(0.975)
mat.ci[i,] <- mean(y) + se_z * c(-1, 1)}
mean(mat.ci[,1] < 5 & mat.ci[,2] > 5)
## [1] 0.94676
```

▼ Confidence Interval for a Proportion

The method for approximate intervals can be readily used for confidence intervals on a proportion π, the success probability of a random sample of n binary variables:

$$Y_i \sim B(1, \pi), \qquad i = 1, ..., n$$

The pivot is given by:

$$Z(\pi) = \frac{\overline{Y} - \pi}{\sqrt{\frac{\overline{Y}(1-\overline{Y})}{n}}} \dot{\sim} \mathcal{N}(0,1), \qquad \forall \pi \in (0,1)$$

where $\hat{\pi} = \overline{Y}$, with the standard error $SE(\hat{\pi})$ approximated by $\sqrt{\frac{\pi(1-\pi)}{n}}$, which is estimated by plugging-in $\hat{\pi}$ in place of π.

▼ R Lab: Confidence Interval for a Proportion

```
M <- 100000; n <- 50; mat.ci <- matrix(NA, nrow = M, ncol = 2)
for(i in 1:M) {
  y <- rbinom(n, size = 1, prob = 0.25)
  p.hat <- mean(y)
  se_z <- sqrt(p.hat * (1 - p.hat) / n)
  mat.ci[i,] <- mean(y) + se_z * qnorm(0.975) * c(-1, 1)
}
mean(mat.ci[,1] < 0.25 & mat.ci[,2] > 0.25)
## [1] 0.94063
```

▼ Confidence Interval for the Difference of Means

A significant application of confidence intervals is comparing two means. For two independent random samples, the confidence interval for their difference $\delta = \mu_X - \mu_Y$ uses the approximate pivot:

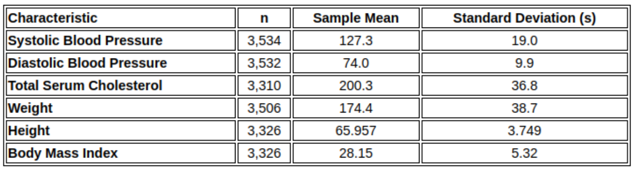
$$Z(\delta) = \frac{\hat{\delta} - \delta}{SE(\hat{\delta})}$$

where $\hat{\delta} = \overline{X} - \overline{Y}$ and $SE(\hat{\delta}) = \sqrt{SE(\overline{X})^2 + SE(\overline{Y})^2}$.

Again, for normal samples, exact solutions exist, both for the case of equal variances and for the case of unequal variances.

▼ Example

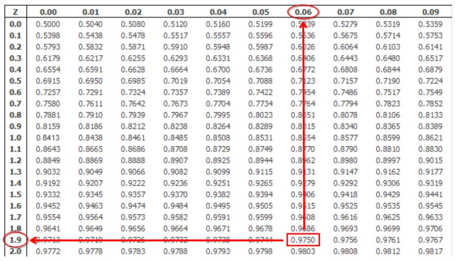
Descriptive statistics on variables measured in a sample of a $n = 3{,}539$ participants attending the 7^{th} examination of the offspring in the Framingham Heart Study are shown below.



Characteristic	n	Sample Mean	Standard Deviation (s)
Systolic Blood Pressure	3,534	127.3	19.0
Diastolic Blood Pressure	3,532	74.0	9.9
Total Serum Cholesterol	3,310	200.3	36.8
Weight	3,506	174.4	38.7
Height	3,326	65.957	3.749
Body Mass Index	3,326	28.15	5.32

We can generate a 95% confidence interval for systolic blood pressure using the following formula:

$$\overline{X} = \pm t \frac{s}{\sqrt{n}}$$



Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817

Since the confidence level is 95%, we want the middle of 95% of the distribution. This leaves 2.5% in each tail, so we're looking for the z value corresponding to 0.975.

The Z value for 95% confidence is $z_{95\%} = 1.96$.

Hypothesis testing

The idea of hypothesis testing

The fundamental goal of hypothesis testing within a parametric statistical model $f_\theta(y)$ is to determine whether the observed data could be reasonably generated by the model $f_{\theta_0}(y)$, where θ_0 is a specific value of the parameter θ. This concept is often simplified using the notation:

$$H_0 : \theta = \theta_0$$

where H_0 represents the **null hypothesis**. Additionally, an alternative hypothesis, H_1, is required. This hypothesis specifies values of θ that become plausible if H_0 is not true.

Example: Testing the Mean of a Normal Sample

Assume a model for independent observations $y_1, y_2, \ldots, y_n$ generated by a normal distribution with unknown mean μ and variance 1:

In this case, we might want to test:

$$H_0 : \mu = 0 \quad \text{against} \quad H_1 : \mu > 0$$

This test examines the null hypothesis that the data originates from a standard normal distribution, with the alternative hypothesis suggesting a positive mean value. This formulation of H_1 (a **one-sided alternative**) is appropriate if we can exclude negative values for μ. If not, a **two-sided alternative** might be better:

$$H_1 : \mu \neq 0$$

General Formulation

In a broader sense, hypotheses regarding a parameter θ can be framed as:

$$H_0 : \theta \in \Theta_0 \quad \text{against} \quad H_1 : \theta \in \Theta_1$$

where Θ_0 and Θ_1 form a partition of the **parameter space** Θ, which includes all possible values for θ.

Methods for handling this generality are often covered in sections on **likelihood methods**.

Steps of Hypothesis Testing

The process of hypothesis testing consists of several key steps:

Test Statistic: A function of the observed sample used to conduct the test.

A **test statistic** is a function of the random sample, chosen based on the context of the hypothesis being tested.

For the example above (testing $\mu = 0$ in a normal distribution), a natural test statistic is the standardized sample mean:

$$Z = \frac{\overline{Y}}{\sqrt{\frac{1}{n}}} = \sqrt{n}\,\overline{Y}$$

Two-sided hypothesis testing for the mean: $H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0$.

Case	Test Statistic	Acceptance Region
$X_i \sim N(\mu, \sigma^2)$, σ known	$W = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$	$\lvert W \rvert \leq z_{\frac{\alpha}{2}}$
n large, X_i non-normal	$W = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$	$\lvert W \rvert \leq z_{\frac{\alpha}{2}}$
$X_i \sim N(\mu, \sigma^2)$, σ unknown	$W = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$	$\lvert W \rvert \leq t_{\frac{\alpha}{2}, n-1}$

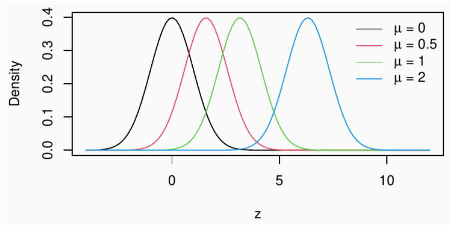
Null and Alternative Distributions: The distribution of the test statistic under each hypothesis.

The distribution of a test statistic generally depends on the actual value of θ. For example, if H_0 is true, then $Z \sim \mathcal{N}(0, 1)$, termed the **null distribution**. If H_1 holds, then:

$$Z \sim \mathcal{N}(\Delta, 1)$$

where $\Delta = \sqrt{n}\mu > 0$.

Distributions under H_1 are known as **alternative distributions**.



▼ **The p-value**: The probability of observing a test statistic as extreme as, or more extreme than, the observed value, assuming H_0 is true.

suppose we end up rejecting H_0 at at significance level $\alpha = 0.05$. Then we could ask: "How about if we require significance level $\alpha = 0.01$?" Can we still reject H_0?

More specifically, we can ask the following question: What is the lowest significance level α that results in rejecting the null hypothesis? The answer to the above question is called the *P*-value.

> **P-value** is the lowest significance level α that results in rejecting the null hypothesis.

Intuitively, if the *P*-value is small, it means that the observed data is very unlikely to have occurred under H_0, so we are more confident in rejecting the null hypothesis.

The **p-value** quantifies the distance between the data and H_0. A small p-value suggests that the data are inconsistent with H_0. Specifically, the p-value is the probability (under H_0) of observing a value of the test statistic as extreme as or more extreme than the observed value, z_{obs}:

$$p = P_{H_0}(Z \geq z_{obs}) = 1 - \Phi(z_{obs})$$

where Φ denotes the standard normal cumulative distribution function.

▼ **R Code Example: Computing p-value for a Sample**

If the null distribution isn't known, we can compute the p-value via simulation. For example, in R:

```
set.seed(13)
n <- 10
y_obs <- rnorm(n)
z_obs <- mean(y_obs) * sqrt(n)
print(z_obs)

M <- 100000
z_sim <- numeric(M)
for(i in 1:M) {
  y <- rnorm(n)
  z_sim[i] <- mean(y) * sqrt(n)
}
c(mean(z_sim >= z_obs), 1 - pnorm(z_obs))
```

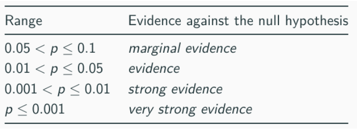
Comments on the p-value

1. The p-value **does not represent the probability** that H_0 is true, since the latter is not even an event.
2. Statistical tests should be **interpreted within context**; small p-values are not always meaningful (e.g. if the alternative hypothesis is logically implausible).
3. Hypothesis testing has limitations. For large samples, minor deviations can yield small p-values. Alternatives like **model selection** may be more appropriate.

▼ **Significance Level**: Define the decision thresholds.

A result is **significant at the 5% level** if the p-value is less than or equal to 0.05. Common thresholds are:

We commonly say that a the result of a test is significant at the 5% level whenever the p-value is smaller or equal to 0.05. Other levels of some practical interest are 1% or 0.1%.



Range	Evidence against the null hypothesis
$0.05 < p \leq 0.1$	*marginal evidence*
$0.01 < p \leq 0.05$	*evidence*
$0.001 < p \leq 0.01$	*strong evidence*
$p \leq 0.001$	*very strong evidence*

A test with fixed significance level arises when the significance level is fixed in advance, and then it is just reported whether the p-value is smaller than the fixed level. If this happens, it may be reported that H_0 **is rejected**, otherwise we may say that H_0 **is not rejected** (or **accepted**).

▼ **Rejection and Acceptance Regions**: Assessment of Type I and Type II errors, and the power of the test.

If we define the **sample space** as the set of the values that our available sample may take, the **rejection region** of a test with fixed significance level is the subset of the sample space corresponding to the samples that would lead to a rejection of H_0.

The remaining part of the sample space forms instead the **acceptance region**.

Both these two regions are determined by means of a test statistic.

For the example previously introduced: for $H_1 : \mu > 0$, the rejection region at level α is:

where $z_{1-\alpha}$ is the standard normal $(1-\alpha)$-quantile, i.e. 1.645 for $\alpha = 0.05$.

The acceptance region is just given by:

$$A_\alpha = y : Z < z_{1-\alpha}$$

Note

The computation of p-value, R_α and A_α would be exactly the same if the null hypothesis were of the form $H_0 : \mu \leq 0$, maintaining the same alternative hypothesis.)

▼ **Errors and power in Fixed-Significance Tests**: Assessment of Type I and Type II errors, and the power of the test.

With a fixed significance level, two types of errors are possible:

Type I error

We define **type I error** as the event that we reject H_0 when H_0 is true. Note that the probability of type I error in general depends on the real value of θ. More specifically:

$$\begin{aligned} P(\text{type I error} \mid \theta) &= P(\text{Reject } H_0 \mid \theta) \\ &= P(W \in R \mid \theta), \quad \text{ for } \theta \in S_0. \end{aligned}$$

If the probability of type I error satisfies

$$P(\text{type I error}) \leq \alpha, \quad \forall \theta \in S_0, \tag{3}$$

then we say the test has **significance level α** or simply the test is a level α test.

Type II error

The second possible error that we can make is to accept H_0 when H_0 is false. This is called the **type II error**. Since the alternative hypothesis, H_1, is usually a composite hypothesis (so it includes more than one value of θ), the probability of type II error is usually a function of θ and is usually shown by β:

$$\beta(\theta) = P(\text{Accept } H_0 \mid \theta), \quad \text{ for } \theta \in S_1. \tag{4}$$

In the example, $\Pr_{H_0}(Y \in R_\alpha) = \alpha$, and in fact the fixed significance level equals the probability of making a Type I error.

Power

For a test with fixed significance level, the power is the probability of (correctly) detecting that H_0 is false.

$$\Pr_{H_1}(\boldsymbol{Y} \in \mathcal{R}_\alpha)$$

The power of a test can be used for comparing alternative tests for the same problem, with tests with higher power being preferable. The power is often used for designing studies, in particular for choosing the sample size in medical or industrial studies. Indeed, for fixed significance level, the power increases with the sample size.

▼ **Example: Power of two tests**

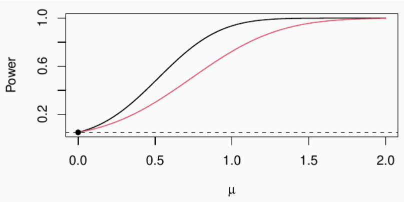
For the simple example (with $H_1 : \mu > 0$), an alternative (but silly) test statistic may be given by taking the same Z as above computed by using only half of the sample (for n even).

Fixing a significance level of 5%, the two tests have exactly the same probability of a Type I error, so for comparing them we must use their power.

The power is a function of the μ assumed under H_1 , and for a certain $\mu \geq 0$ we obtain (since $z_{0.95} = 1.645$)

$$\Pr_{\mu}(Z \geq 1.645) = 1 - \Phi(1.645 - \sqrt{n}\,\mu)$$

```
mu <- seq(0, 2 , l = 1000); n <- 10; n1 <- 5
plot(mu, 1 - pnorm(1.645 - sqrt(n) * mu), type = "l", ylab="Power", xlab = expression(mu))
lines(mu, 1 - pnorm(1.645 - sqrt(n1) * mu), col = 2)
abline(h=0.05, lty = 2); points(0, 0.05, pch = 16)
```



▼ **Some commonly used tests**

Several statistical tests are frequently used to evaluate hypotheses. Among these, the **t-test** is one of the most fundamental, commonly applied to test hypotheses about the means of normal distributions.

▼ **One-sample t test**

Given a normal random sample $y_1, \ldots, y_n$ with each $Y_i \sim \mathcal{N}(\mu, \sigma^2)$, the classical hypothesis test for μ in a two-sided form is:

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

The test statistic is:

$$T = \frac{\overline{Y} - \mu_0}{\sqrt{\frac{S^2}{n}}} \sim t_{n-1} \qquad \text{when } H_0 \text{ is true}$$

where S^2 is the sample variance.

The p-value is calculated as:

$$p = P_{H_0}(|T| \geq |t_{obs}|)$$

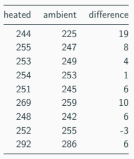
or, equivalently,

$$p = 2\,P_{H_0}(T \geq |t_{obs}|) = 2\,\{1 - F_{t,n-1}(|t_{obs}|)\}$$

since the t distribution is symmetric around zero.

- Example

The **pair65 dataset** from the DAAG package contains an experiment on the effect of heat on the elasticity of bands, showing the differences between heated and ambient conditions for nine bands:



heated	ambient	difference
244	225	19
255	247	8
253	249	4
254	253	1
251	245	6
269	259	10
248	242	6
252	255	-3
292	286	6

Focusing on the nine differences in stretch, we can test:

$$\begin{cases} H_0 : \mu = 0 \\ H_1 : \mu \neq 0 \end{cases}$$

by means of the t.test function, resulting in significance at 5% level

```
# One Sample t-test

data:
    difference

t = 3.1131,
df = 8,
p-value = 0.01438

alternative hypothesis:
    true mean is not equal to 0

95 percent confidence interval:
    1.641939 11.024728

sample estimates:
    mean of x -> 6.333333
```

- Approximate Tests

For large random samples, the **Central Limit Theorem (CLT)** ensures that:

$$\overline{Y} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

where $\mu = E(Y_i)$ and $\sigma^2 = \text{Var}(Y_i)$.

A test statistic for $H_0 : \mu = \mu_0$ is:

$$Z = \frac{\overline{Y} - \mu_0}{\sqrt{\frac{S^2}{n}}} \sim \mathcal{N}(0,1) \qquad \text{when } H_0 \text{ holds}$$

The estimator of the variance S^2 can be replaced by a more suitable one.

For binary data, if $Y_i \sim \mathcal{B}_i(1, \pi)$, common test statistics include:

$$Z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}} \quad \text{or} \quad Z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

Tests based on the CLT are known as **approximate tests,** for which the property concerning the Type I error level holds only approximately.

- Two-Sample t-Test

For two **independent normal samples** $X_i \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y_i \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$, the test statistic for testing the equality of the means is:

$$T = \frac{\overline{X} - \overline{Y}}{\text{SE}(\overline{X} - \overline{Y})}$$

where $\text{SE}(\overline{X} - \overline{Y})$ is estimated by $\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}$

A different formula is instead adopted is if it is possible to assume that $\sigma_X^2 = \sigma_Y^2$.

The distribution of T when H_0 is true is given by a suitable t distribution.

Like for the one-sample case, there are general formulas for large samples, employing the normal distribution.

- Paired t-Test

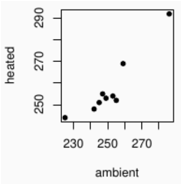
When we have paired observations, each unit in a sample is measured twice under different conditions, making it a case of dependent data: we end up again with two set of variables $X_i \sim N(\mu_X, \sigma_X^2), i = 1, ..., n$ and $Y_i \sim N(\mu_Y, \sigma_Y^2), i = 1, ..., n$.

Now, the pair (X_i, Y_i) refers to the same unit, so that the two samples $X_1, ..., X_n$ and $Y_1, ..., Y_n$ are **no longer independent**.

For such cases, we focus on the n differences $D_i = X_i - Y_i$ and check if the mean difference $\mu_D = E(D_i) = \mu_X - \mu_Y$ significantly deviates from zero. This approach simplifies to a one-sample t-test on the differences, where the null hypothesis $H_0 : \mu_D = 0$ represents no effect of the treatment.

▼ Example (pair65)

In the example of the **pair65** dataset, the paired t-test returned a p-value of about 0.014, suggesting that the treatment (e.g., heat) significantly impacted the outcome (e.g., stretchiness). Even though the pair65 data is very small, the fact that the two groups of observations are not independent is readily suggested by a scatterplot



The scatter plot clearly shows a relationship between paired observations, confirming that these data points are not independent. If we wrongly apply a test assuming independence, as illustrated, the results may vary significantly, here giving a misleading p-value of about 0.40.

▼ Relation between tests and confidence intervals

As displayed for the pair65 data testing, the `t.test` R function returns also the confidence interval for the parameter under testing, in that case the true mean of the differences in stretchiness.

This is not by chance, since there is a close connection between hypothesis testing on the value of a certain parameter and confidence intervals for that parameter.

For the case of a mean, for example, the basic idea is that

> *If the confidence interval for μ does not contain zero, this is equivalent to rejection of the hypothesis that the true mean is zero.*

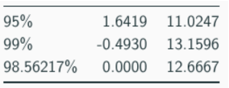
Important: the connection is between two-sided confidence intervals and two-sided alternative hypotheses. For one-sided alternative hypotheses, the connection is with one-sided confidence intervals.

Equivalence between the two methods

1. Given a method to find a confidence interval of level (1 − α)% for a certain scalar parameter θ, we can establish whether the p-value for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ is smaller than the significance level α by checking if θ_0 is included in the interval
2. Given a method to find a p-value for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, we can obtain a confidence interval of level $1 - \alpha$ by selecting all the θ_0 values that will lead to a p-value larger than α

▼ Example: pair65 Data with Confidence Intervals

The **pair65** dataset's confidence intervals for the mean difference at 95% and 99% confidence intervals for the mean of the differences are, respectively:



95%	1.6419	11.0247
99%	-0.4930	13.1596
98.56217%	0.0000	12.6667

The 95% confidence interval does not contain zero, while the wider 99% does, implying that the hypothesis $\mu = 0$ is rejected for $\alpha = 0.05$, but not for $\alpha = 0.01$.

Note

For a confidence interval of level $1 - p = 0.9856217$, we obtain a lower limit exactly equal to 0:

> the p-value, in fact, corresponds to a significance level which is borderline between rejection and non-rejection of H_0 .

▼ Non parametric tests (WIP)

▼ The likelihood function

The likelihood function for a certain statistical model $f_\theta(y)$ for the data y is given by the following function of the parameter θ:

$$\begin{aligned} L \quad : \quad & \Theta \rightarrow \mathbb{R}^+ \\ & \theta \rightarrow c(y) f_\theta(y) \end{aligned}$$

where $c(y) > 0$ is an arbitrary constant of proportionality.

We may write $L(\theta; y)$ to stress the fact that the data enter the function, though its argument is given by θ.

▼ Example

We extract n balls with replacement. We have:

$$Y_i \underset{iid}{\sim} Be(\pi), \qquad \begin{cases} 1 & \text{if ball is white} \\ 0 & \text{if ball is red} \end{cases}$$

We suppose to have the following results for $n = 5$:

`0, 1, 1, 0, 1` whose probability is $(1 - \pi) \cdot \pi \cdot \pi \cdot (1 - \pi) \cdot \pi$

more in general we have:

$$\pi^{n_1}(1-\pi)^{n-n_1}$$

let's try with some values:

- $\pi = 0.5 \Rightarrow P = 0.5^3 \cdot 0.5^2 = 0.5^5 = 0.03125$
- $\pi = 0.3 \Rightarrow P = 0.3^3 \cdot 0.7^2 = \qquad = 0.01323$

Roadmap

- Select a model
- Calculate the probability of obtaining a sample set according to the distribution itself
- Find the most likelifood function that generates that sample

Interpreting the likelihood function

The likelihood function assigns support (credibility) to possible values of θ meaning that if $L(\theta_1) > (\theta_2)$ then θ_1 is more supported by the observed data then θ_2.

So the likelihood ratio $L(\theta_1)/L(\theta_2)$ allows for the comparison between θ_1 and θ_2; note that the constant c(y) cancels out.

A mathematical justification for the above interpretation is given by the

Wald inequality: if θ_t is the true parameter value, then

$$E_{\theta_t}\{\log L(\theta_t; Y)\} > E_{\theta_t}\{\log L(\theta; Y)\} \qquad \theta \neq \theta_t$$

The above fact can be proven by straightforward application of the Jensen's inequality

The log likelihood function

Example: the Poisson model

For a random sample $y_1, \ldots, y_n$, with $Y_i \sim \mathcal{P}(\lambda)\ \ i.i.d,$

we readily get:

$$L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} = \frac{\lambda^{\sum y_i} e^{-n\lambda}}{\prod_{i=1}^{n} y_i!}$$

so that:

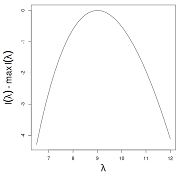
$$\ell(\lambda) = \log(\lambda) \sum_{i=1}^{n} y_i - n\lambda$$

neglecting the term which does not depend on λ.

Example : R lab

Assume that for a sample we have $n = 10, \sum y_i = 90$

```
lik_pois <- function(lam, n, sumy) log(lam) * sumy - n * lam
xx <- seq(6.5, 12, l = 30)
ll <- sapply(xx, lik_pois, sumy = 90, n = 10) # evaluates the combination consider all the possible values of lambda -
par(pty = "s")
plot(xx, ll - max(ll), type = "l", xlab = expression(lambda), ylab = expression(l(lambda) - max(l(lambda))), cex.lab =
```



```
for (i in 1:100) {
    sumy <- sum(rpois(10, 9))
    ll <- lik_pois(xx, 10, sumy)
    lines(xx, ll, max(...
}
```

Example: the normal model

For a random sample $y_1, \ldots y_n$, with $Y_i \sim \mathcal{N}(\mu, \sigma^2)$, i.i.d.

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i-\mu)^2}{2\sigma^2}\right\}$$

and then with some simple algebra

$$\ell(\mu, \sigma^2) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2}\frac{(y_i-\mu)^2}{\sigma^2}$$

LThe log likelihood function is simply the logarithm of $L(\theta)$:

$$\ell(\theta) = \log L(\theta)$$

The log likelihood function carries the same information of the likelihood function, but is much more manageable. Indeed, for a random sample

$$L(\theta) = \prod_{I=1}^{n} f_\theta(y_i)$$

but

$$\ell(\theta) = \sum_{I=1}^{n} \log f_\theta(y_i)$$

Notice that $\ell(\theta)$ is defined up to an additive constant, depending only on the data y.

- **Sufficient statistic**

  The definition of sufficient statistic, given in the probability part, can be re-interpreted for the log likelihood function: $t(y)$ is sufficient for \theta if $L(\theta)$ can be written:

  $$L(\theta) = h(\theta) g_\theta\{t(y)\}$$

  The minimal sufficient statistic allows for the maximal reducrion of dimensionality, in teh sense that a minimal sufficient statistic is a function of every other sufficient statistic.

  For the Poisson model, the $\sum_i y_i$ (or, equivalently, the sample mean $\bar{y}$) is sufficient dor \lambda, whereas for the normal model the sufficient statistic os given by the pair $\left(\sum_i y_i, \sum_o i y_i^2\right)$ (or, equivalently, by the pair $(\bar{y}, s^2)$)

- **Maximum likelihood estimation**

  Given the interpretation of the (log) likelihood, the maximum of $\ell(\theta)$ is the ***maximum likelihood estimator*** (MLE), which is defined as the value of θ that maximizes $L(\theta)$ or, equivalently, $\ell(\theta)$.

  $$\hat{\theta} = \arg\max_{\theta \in \Theta} \ell(\theta)$$

  Notice that since $\ell(\theta)$ is also a function of y, the MLE is a statistic.

  In practice, we often find the MLE by solving the likelihood equation, which is obtained by setting the derivative of the log-likelihood function to zero:

  $$\frac{\partial \ell(\theta)}{\partial \theta} = 0$$

  This approach is particularly useful when dealing with complex likelihood functions, as it can simplify the optimization process.

  - **Examples (continue of poisson and normal model)**

    For the Poisson model, simple calculus gives:

    $$\frac{\partial \ell(\lambda)}{\partial \lambda} = \frac{\partial}{\partial \lambda} \log(\lambda) \sum_{i=1}^{n} y_i - n\lambda = \frac{\sum y_i}{\lambda} - n = 0$$

    $$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} y_i = \bar{y}$$

    For the normal moel, we need to maximize a function of two variables, and we get:

    $$\begin{cases} \hat{\mu} = \bar{y} \\ \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2 \end{cases}$$

...