Contents

1 Probability (Recap)

Suggested textbooks

- S.N. Wood. Core Statistics. Cambridge University Press, 2015.
- B. Efron, T. Hastie. Computer Age Statistical Inference – Algorithms, Evidence, and Data Science. Cambridge University Press, 201.

1.1 Random variables

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

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

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

Distribution of a r.v.

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

1. Probability functions
2. Cumulative distribution functions
3. Quantile functions

1.1.1 Discrete distributions

Discrete r.v. take values in a discrete set.

The **probability (mass) function** (p.m.f.) 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.

1.1.1.1 Mean and variance of a discrete r.v.

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 special case $g(X) = (X\mu)^2$, with $\mu = E(X)$, is the **variance** of X

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

The **standard deviation** is just given by $\sqrt{var(X)}$.

1.1.1.2 Notable discrete random variables

Discrete r.v. often used in applications:

- Binomial distribution
- Poisson distribution
- Negative binomial distribution
- Geometric distribution
- Hypergeometric distribution

The fist two deserve some further attention.

1.1.1.3 The binomial 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 function

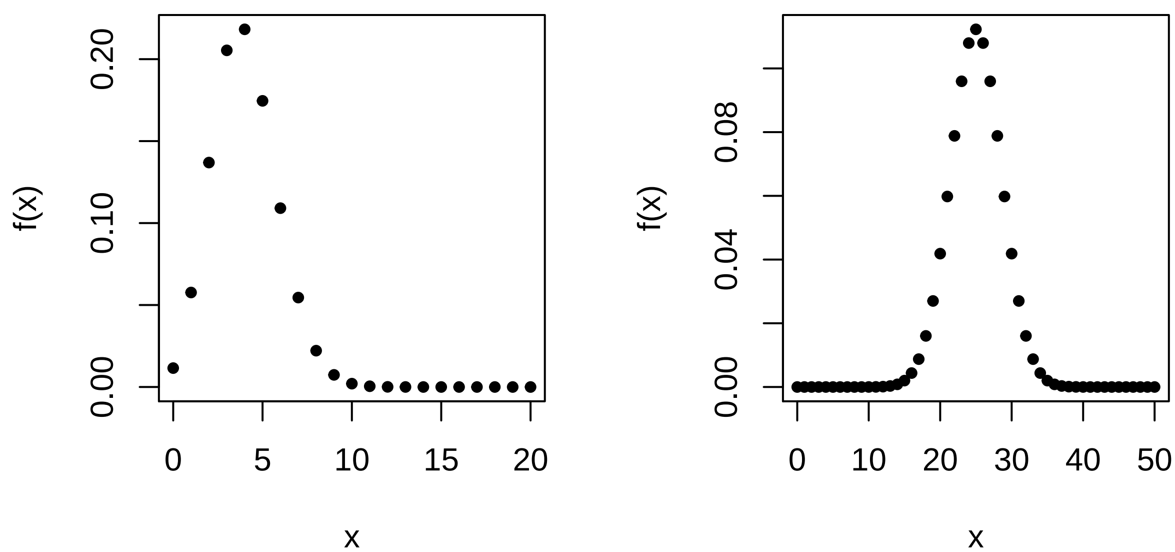
$$Pr(X = x) = \binom{n}{x} p^x (1p)^{nx}, \quad x = 0, \ldots, n.$$

The notation is $X \sim \mathcal{B}\rangle(n, p)$, and $E(X) = np$, $var(X) = np(1p)$.

The case when $n = 1$ is known as **Bernoulli distribution**.

R lab: the binomial distribution

```
par(mfrow=c(1,2), pty="s", pch = 20)
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)")
```



1.1.1.4 The Poisson distribution

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

The probability function is

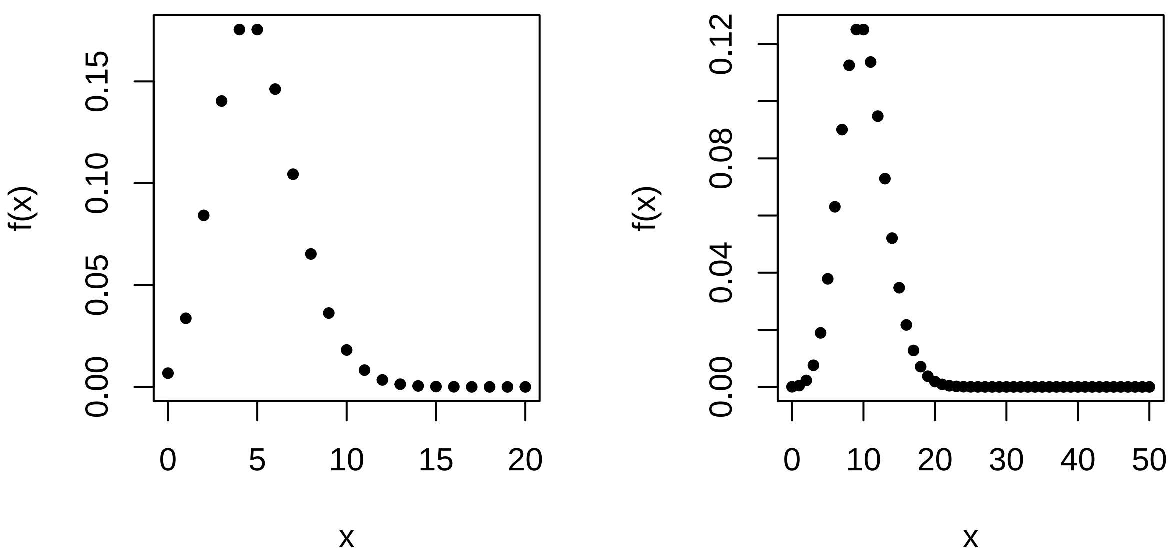
$$Pr(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \dots$$

with $\lambda > 0$.

The notation is $X \sim \mathcal{P}\wr(\lambda)$, and $E(X) = var(X) = \lambda$.

R lab: the Poisson distribution

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



1.1.2 Continuous distributions

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 $a \leq b$

$$Pr(a \leq X \leq b) = \int_a^b f(x)dx.$$

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

The probability density function defines the distribution of X.

1.1.2.1 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

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

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 apply to a **linear transformation** $a + bX$, with a, b constants:

$$E(a + bX) = a + bE(X)$$

$$var(a + bX) = b^2 var(X).$$

1.1.2.2 Notable continuous random variables

Important continuous distributions include:

- Normal distribution
- χ^2 distribution
- F distribution
- t and $Cauchy$ distributions
- Gamma, Weibull and exponential distributions

The normal distribution has a major role in statistics. Theχ^2, t and F distributions are relative of the normal distribution.

1.1.2.3 The normal distribution

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

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

The notation is $X \sim \mathcal{N}(\mu, \sigma^2)$, and $E(X) = \mu$ and $var(X) = \sigma^2$, $\sigma^2 > 0$, $\mu \in \Re$.

An important property is that for any constants a, b

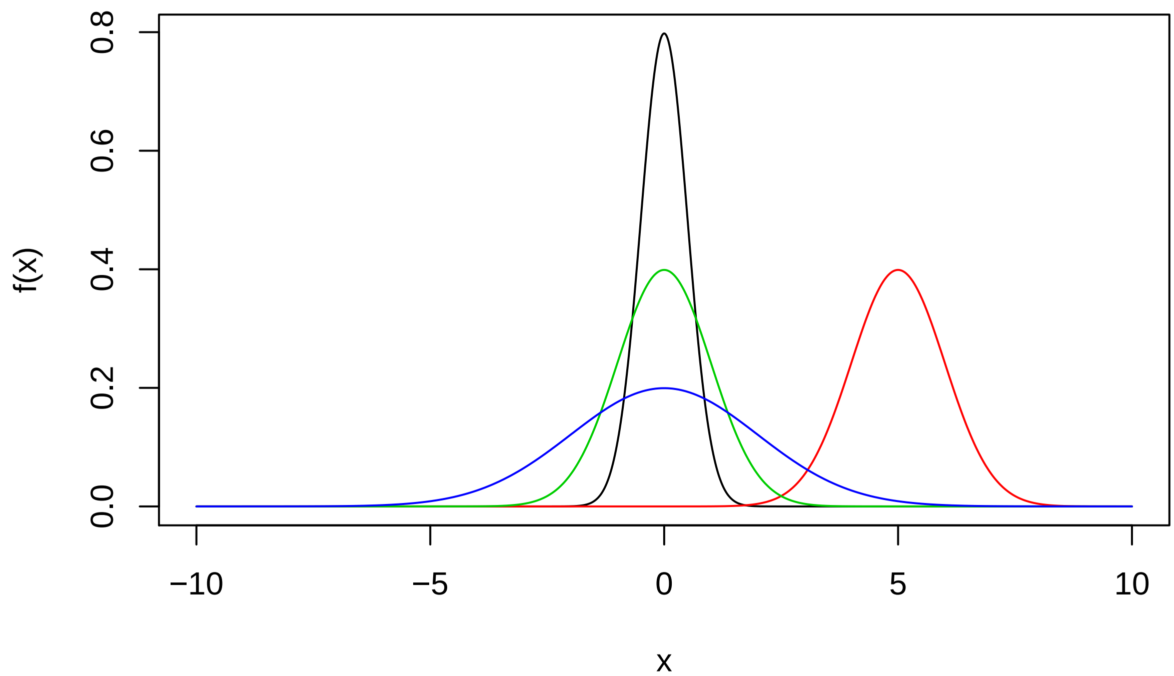
$$a + bX \sim \mathcal{N}(a + b\mu, b^2\sigma^2),$$

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

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

R lab: the normal distribution

```
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)
```



1.1.2.4 The χ^2 distribution

Let $Z_1, \ldots, Z_k$ be a set of independent $\mathcal{N}(0,1)$ 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**.

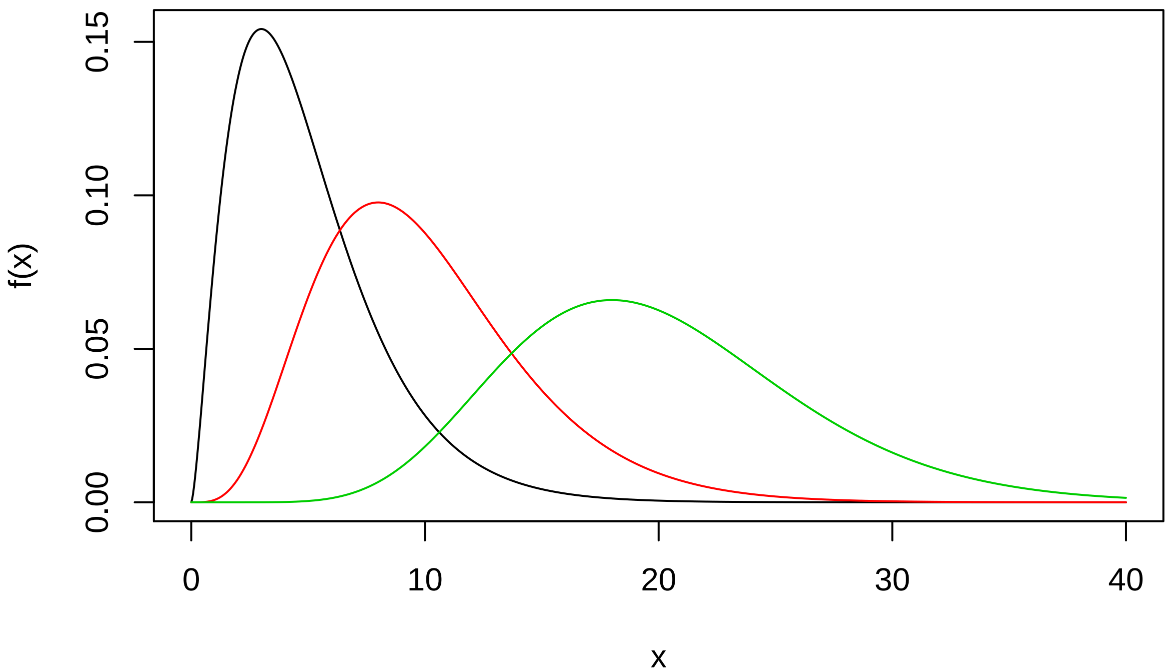
The notation is $X \sim \chi_k^2$, $E(X) = k$ and $var(X) = 2k$.

It is a special case of the Gamma distribution.

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

R lab: the χ^2 distribution

```
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)
```



1.1.2.5 The F distribution

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

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

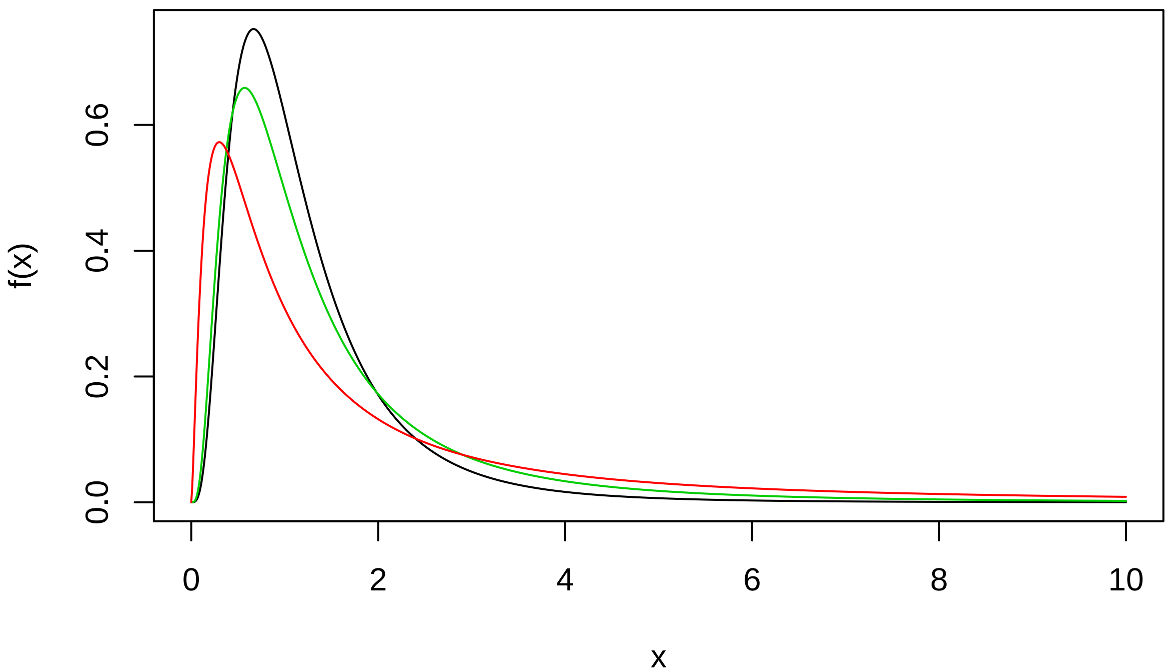
has an F **distribution with** n **and** m **degrees of freedom**.

The notation is $F \sim \mathcal{F}_{n,m}$, and $E(F) = m/(m-2)$ provided that $m > 2$.

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: the F distribution

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



1.1.2.6 The t and Cauchy distributions

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

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

has a t **distribution with** n **degrees of freedom**.

The notation is $T \sim \sqcup_n$, and $E(T) = 0$ provided that $n > 1$, whereas $var(T) = n/(n-2)$ provided that $n > 2$.

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

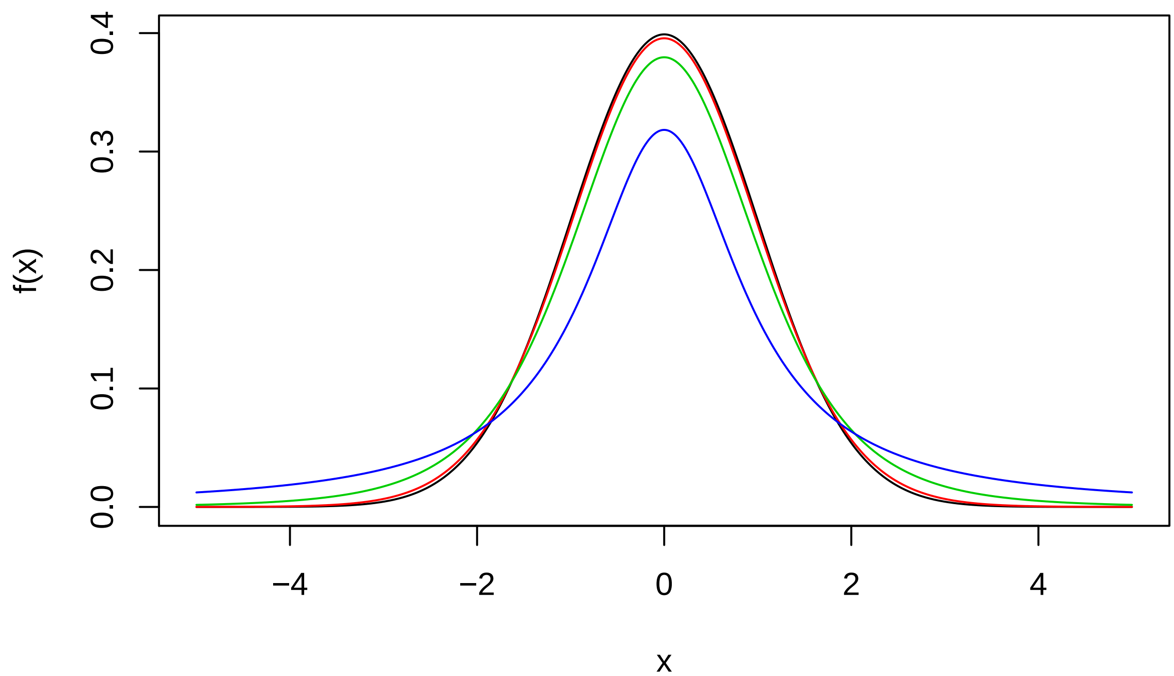
The case t_1 is the **Cauchy distribution**.

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

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

R lab: the t and Cauchy distributions

```
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)
```



1.1.3 C.d.f. and quantile functions

1.1.3.1 Cumulative distribution functions

The **cumulative distribution function** (c.d.f.) of a r.v. X is the function $F(x)$ such that

$$F(x) = Pr(X \leq x),$$

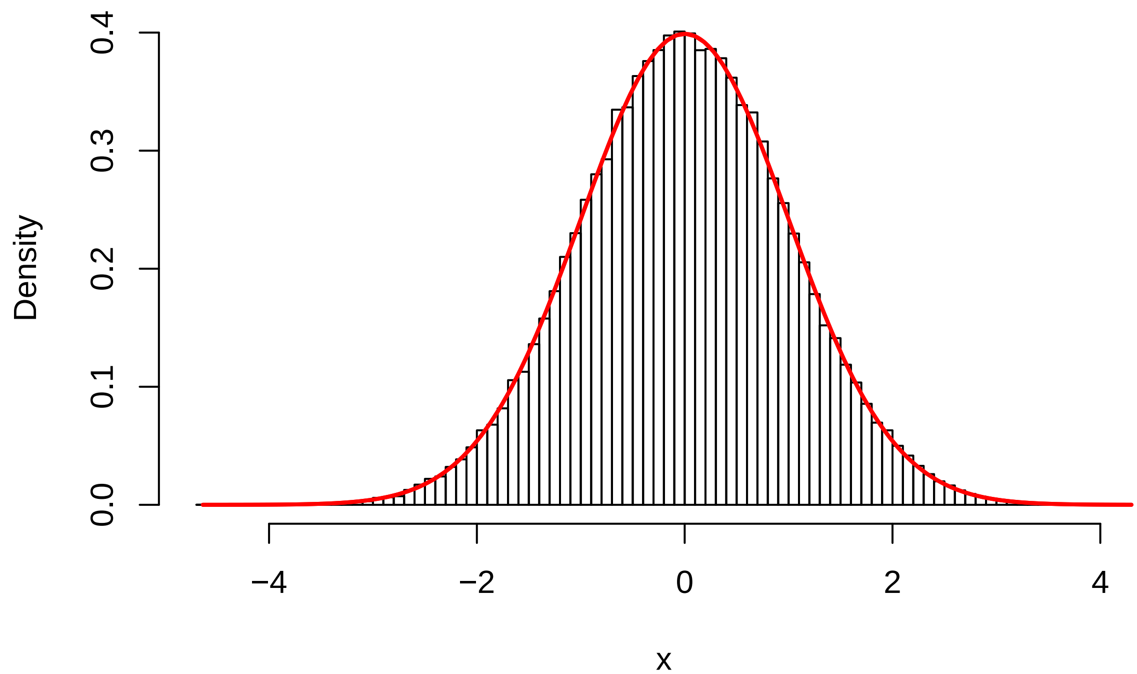
and it can be obtained from the probability function or the density function: the c.d.f. identifies the distribution.

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

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

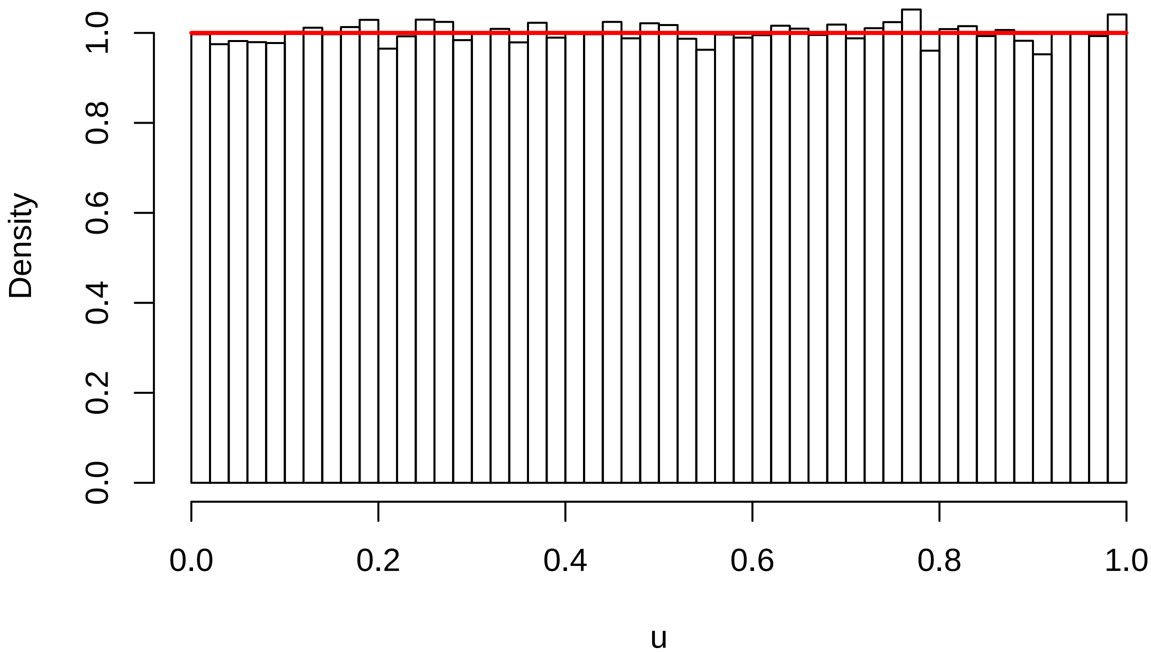
R lab: uniform transformation

```r
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 = 2, lwd = 2)
```



R lab: uniform transformation (cont'd)

```r
u <- pnorm(x) ### that's the uniform transformation
hist.scott(u, main="")
segments(0, 1, 1, 1, col = 2, lwd = 2)
curve(dunif(x), from=0, to=1, col = 2, lwd = 2, add=T)
```



1.1.3.2 The quantile function

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

$$F^1(p) = min(x|F(x) \geq p), \quad 0 \leq p \leq 1.$$

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

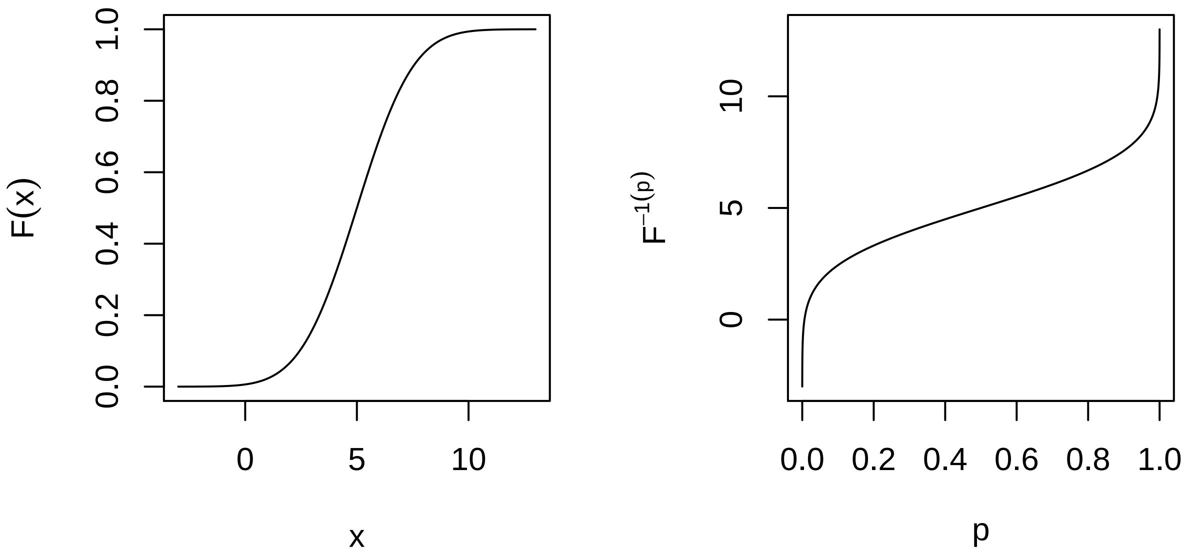
Another useful property is that if $U \sim \mathcal{U}(0,1)$, namely it has a **uniform distribution** in $[0,1]$, then the r.v. $X = F^1(U)$ has c.d.f. F.

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: normal cdf and quantile functions

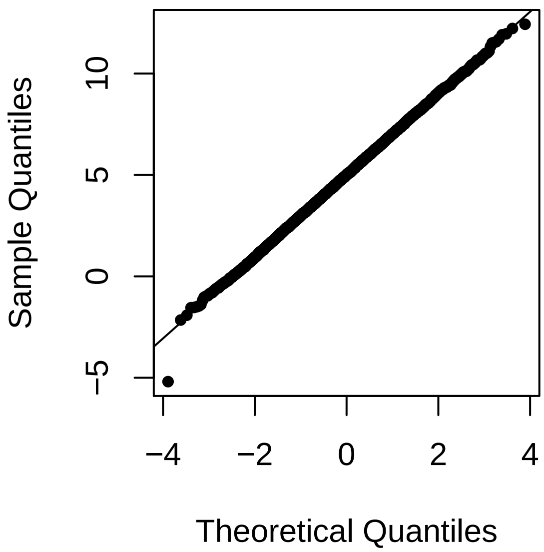
Let us consider the case of $X \sim N(5, 2^2)$, with c.d.f. and quantile functions given by pnorm and qnorm.

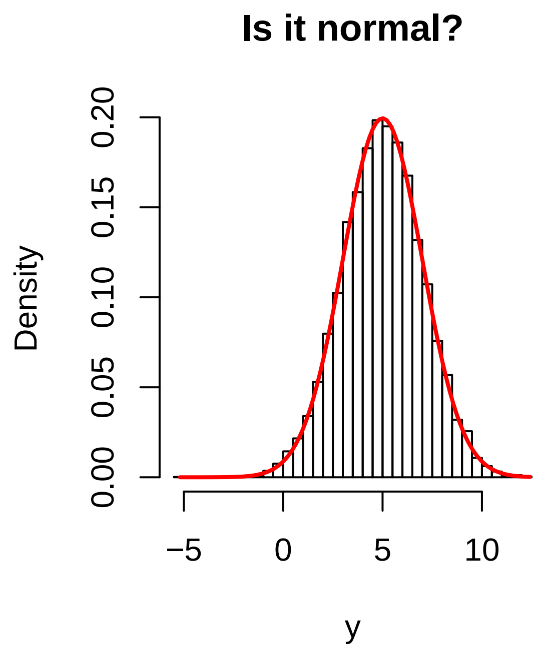
Make by exercise!



R lab: inversion sampling

```
u <- runif(10^4); y <- qnorm(u, m = 5, s = 2)
par(mfrow=c(1,2), pty = "s", pch=20)
qqnorm(y, main = "")
qqline(y)
## Now, trace the density function of y to check ...
```





Side note: quantile-quantile plot

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 $y = (y_1, \ldots, y_n)$.

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**

$$\hat{F}_n(t) = \frac{\text{number of elements of } y \leq t}{n}.$$

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

1.2 Random vectors

1.2.1 Joint distribution

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.

For continuous r.v., the **joint (probability) density function** extends the one-dimensional case: it is the $f(x, y)$ function such that, for any $A \subset \Re^2$

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

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

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

1.2.2 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).

1.2.3 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

$$f(y|X = x_0) = \frac{f(x_0, y)}{f(x_0)},$$

provided $f(x_0) > 0$.

The simplified notation $f(y|x_0)$ is often employed.

The conditional p.d.f. is properly defined, since $f(y|X = x_0) \geq 0$ and $\int_{-\infty}^{\infty} f(y|x_0) dy = 1$.

A symmetric definition applies to X given $Y = y_0$.

Conditional distribution: useful properties

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

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

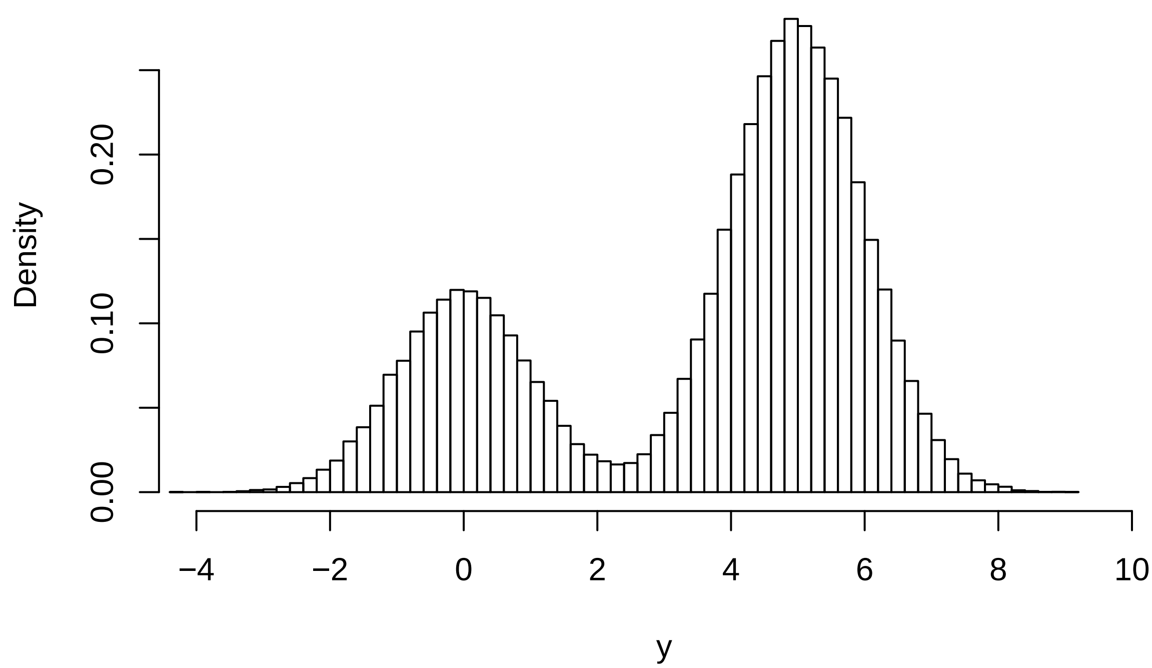
Extensions to higher dimensions require some care:

$$\begin{gathered} f(x, y, z) = f(x, y|z) f(z) \\ f(x, y|z) = f(x|y, z) f(y|z) = f(x|z) f(y|x, z) \\ f(x, y, z) = f(x|y, z) f(y, z) \\ f(x, y, z) = f(x|y, z) f(y|z) f(z) \end{gathered}$$

$$f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2|x_1)f(x_3|x_2, x_1)\ldots f(x_n|x_{n1}, \ldots, x_2, x_1)$$

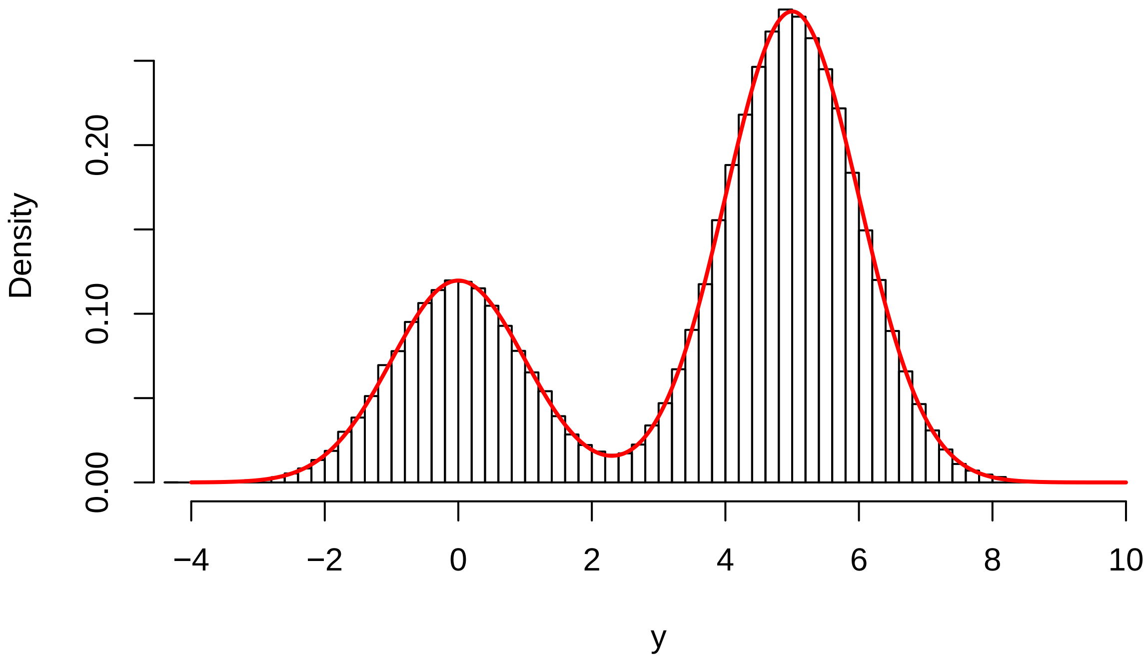
R lab: simulation from joint distributions

```
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))
```



R lab: simulation from joint distributions (cont'd)

```
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)
```



1.2.4 Bayes theorem

From 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**

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

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

1.2.5 Independence and conditional independence

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

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

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

$$f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\ldots f(x_n).$$

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.

1.2.5.1 Example of conditional independence: the Markov property

The general factorization defined above

$$f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2|x_1)f(x_3|x_2, x_1)\ldots f(x_n|x_{n1}, \ldots, x_2, x_1)$$

will simplify considerably when the **first order Markov property** holds:

$$f(x_i|x1, \ldots, x_{i1}) = f(x_i|x_{i1})$$

which means that X_i is independent of $X_1, \ldots, X_{i2}$ given X_{i1}.

We get

$$f(x_1, x_2, \ldots, x_n) = f(x_1)\sum_{i=2}^{n} f(x_i|x_{i1}).$$

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.

1.2.5.2 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 need first to introduce the **covariance** between X and Y

$$cov(X,Y) = E\{(X\mu_x)(Y\mu_y)\} = E(XY)\mu_x\mu_y,$$

where $\mu_x = E(X)$ and $\mu_y = E(Y)$.

Then

$$var(aX + bY) = a^2var(X) + b^2var(Y) + 2abcov(X,Y).$$

Note: for X, Y independent it follows that $cov(X,Y) = 0$. The reverse is not true, unless the joint distribution of X and Y is multivariate normal.

1.2.5.3 Mean vector

For a random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$, the **mean vector** is just

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

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

$$E(A\mathbf{X} + \mathbf{b}) = \mathbf{A}E(\mathbf{X}) + \mathbf{b}.$$

1.2.5.4 Variance-covariance matrix

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

$$\Sigma = E\{(\mathbf{X} - \mu_x)(\mathbf{Y} - \mu_y)^T\} = \begin{pmatrix} var(X_1) & cov(X_1, X_2) & \ldots & cov(X_1, X_n) \\ cov(X_1, X_2) & var(X_2) & \ldots & cov(X_2, X_n) \\ \vdots & \vdots & \vdots & \vdots \\ cov(X_1, X_n) & cov(X_2, X_n) & \ldots & var(X_n) \end{pmatrix}$$

Useful properties:

$$\Sigma_{\mathbf{AX}+\mathbf{b}} = \mathbf{A}\Sigma\mathbf{A}^T$$
$$var(\mathbf{a}^T\mathbf{X}) = \mathbf{a}^T\Sigma\mathbf{a}\,(\geq 0)$$

1.2.5.5 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

$$f_{\mathbf{y}}(\mathbf{y}) = f_{\mathbf{x}}\{g^{-1}(\mathbf{y})\}\left|\mathbf{J}\right|.$$

with $J_{ij} = \partial x_i / \partial y_j$.

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

1.3 The multivariate normal distribution

Start from a set of n i.i.d. $Z_i \sim N(0, 1)$, so that $E(\mathbf{Z}) = \mathbf{0}$ and covariance matrix $\mathbf{I}_n$. If $\mathbf{B}$ is $m \times n$ matrix of coefficients and μ a m-vector of coefficients, then the m-dimensional random vector $\mathbf{X}$

$$\mathbf{X} = \mathbf{BZ} + \mu$$

has a **multivariate normal distribution** with covariance matrix $\Sigma = \mathbf{BB}^T$.

The notation is

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

Joint p.d.f.

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

$$f_{\mathbf{x}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^m |\Sigma|}} \exp\left\{(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)\right\}, \quad \mathbf{x} \in \Re^m$$

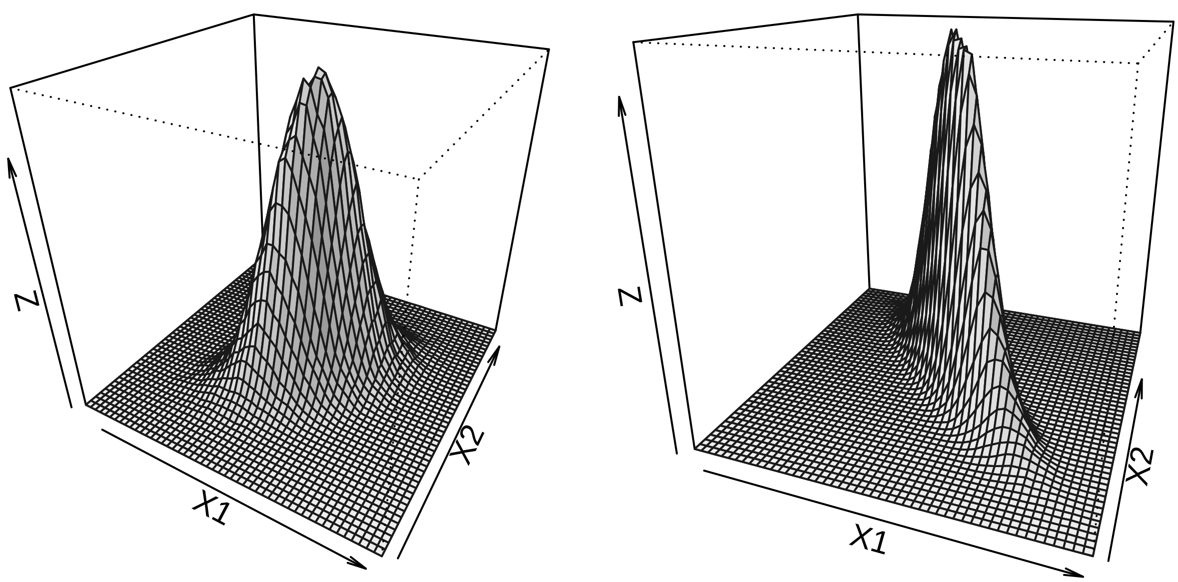
provided that Σ has full rank m.

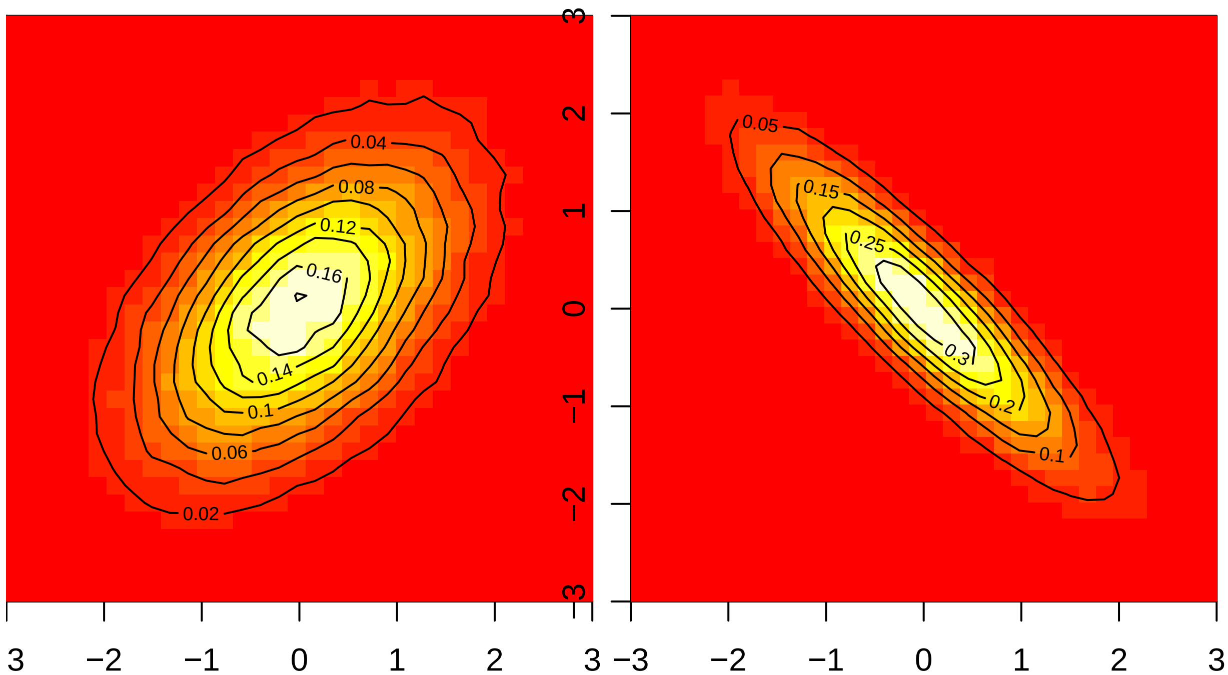
The result can be extended to singular Σ by recourse to the pseudo-inverse of Σ: this is used, for example, in the analysis of *compositional data*.

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

1.3.0.1 Example: bivariate case

We take $\mu_1 = \mu_2 = 0$, $\sigma_1^2 = sigma_2^2 = 1$, $\sigma_{12} = .5$ (at left), $\sigma_{12} = -.9$ (at right).





1.3.1 Linear transformations

It is simple to verify that if $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$ and $\mathbf{A}$ is a $k \times m$ matrix of constants then

$$\mathbf{AX} \sim \mathcal{N}(\mathbf{A}\mu, \mathbf{A}\Sigma\mathbf{A}^T).$$

A special case is obtained when $k = 1$, in that for a m-dimensional vector $\mathbf{a}$

$$\mathbf{a}^T\mathbf{X} \sim \mathcal{N}(\mathbf{a}^T\mu, \mathbf{a}^T\Sigma\mathbf{a}).$$

Note that for suitable choices of $\mathbf{a}$ (when all the elements 0s or 1s) it follows that the marginal distribution of any subvector of $\mathbf{X}$ is multivariate normal.

Normality of the marginal distributions, instead, does not imply multivariate normality

---

1.3.2 Conditional distributions

Consider two random vectors $\mathbf{X}$ and $\mathbf{Y}$ with multivariate normal joint distribution, and partition their joint covariance matrix as

$$\Sigma = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix}$$

and similarly for the mean vector $\mu = (\mu_{\mathbf{X}}, \mu_{\mathbf{Y}})^T$.

Using results on partitioned matrices, it follows that the **conditional distributions are multivariate normal**.

For instance

$$\mathbf{Y}|\mathbf{x} \sim \mathcal{N}(\mu_{\mathbf{Y}} + \Sigma_{\mathbf{YX}}\Sigma_{\mathbf{XX}}^{-1}(\mathbf{x} - \mu_{\mathbf{x}}), \Sigma_{\mathbf{YY}} - \Sigma_{\mathbf{YX}}\Sigma_{\mathbf{XX}}^{-1}\Sigma_{\mathbf{XY}}).$$

1.4 Statistics

1.4.1 Random sample

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

- $X_1, X_2, \ldots, X_n$ 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.)

1.4.2 Statistics

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 $y_1, y_2, \ldots, y_n$:

$$\bar{y} = \frac{1}{n}\sum_i^n y_i, \qquad s^2 = \frac{1}{n-1}\sum_i^n (y_i \bar{y})^2.$$

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

If a statistic $t(\mathbf{y})$ is such that $f_\theta(\mathbf{y})$ can be written as

$$f_\theta(\mathbf{y}) = h(\mathbf{y})g_\theta t(\mathbf{y}),$$

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

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. $Y_i \sim \mathcal{N}(\mu, \sigma^2)$, it follows that $\theta = (\mu, \sigma^2)$ and

$$\begin{aligned} f_\theta(\mathbf{y}) &= \prod_i^n \tfrac{1}{\sigma\sqrt{2\pi}} \exp\left\{\tfrac{1}{2\sigma^2}(y_i - \mu)^2\right\} \\ &= \tfrac{1}{\sigma^n(\sqrt{2\pi})^n} \exp\left\{\tfrac{1}{2\sigma^2}(y_i - \mu)^2\right\}. \end{aligned}$$

By some simple algebra, it is possible to show that the two-dimensional statistic $t(\mathbf{y}) = (\bar{y}, s^2)$ is sufficient for (μ, σ^2).

1.5 Complements & large-sample results

1.5.1 Moment generating function

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

$$M_X(t) = E(e^{tX}), \quad \text{for } t \text{ real.}$$

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

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

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, $M_{X+Y}(t) = M_X(t)M_Y(t)$.

1.5.2 The central limit theorem

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

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

More formally, the theorem says that for any $x \in \Re$ the c.d.f. of $Z_n = (\bar{X}_n \mu)/\sqrt{\sigma^2/n}$ 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 the normal approximation to the distribution of a r.v. that can be viewed as the sum of other r.v.

1.5.3 The law of large numbers

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

The **strong law of large numbers** states that, for any positive ϵ

$$Pr\left(lim_{n\to\infty} |\bar{X}_n \mu| < \epsilon\right) = 1,$$

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

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

$$lim_{n\to\infty} Pr(|\bar{X}_n \mu| \geq \epsilon) = 0.$$

The proof of the weak law of large numbers uses the **Chebyshev's inequality**.

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

1.5.4 Jensen's inequality

This is another useful result, that 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

$$E(X)^2 \geq E(X^2) \;\Rightarrow\; E(X)^2 \leq E(X^2),$$

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

1.6 In-course exercise

1.6.1 The binomial distribution: approximation with CLT

Using the central limit theorem, we may approximate the binomial distribution with a normal distribution. In fact, by means of CLT, we already know that

$$\bar{X}_n \sim \mathcal{N}(p, pq/n)$$

Then, it is easy to show that

$$n\bar{X}_n \sim \mathcal{N}(np, npq)$$

For large n, the binomial distribution may be approximated by a normal distribution with mean $\mu = np$ and variance $\sigma^2 = npq$.

Exercise on CLT

1. Write a R code for checking the validity of the CLT when the distribustion of X is binomial (Hint: for different p incerease n). Use plots for visualizing the results.
2. Use the code above for checking that a Poisson distribution can be approximated by a Normal by increasing λ.
3. Check the validity of the CLT for the distribution of the mean of n uniform variables in $[0, 1]$.
