Probability and Distribution Refresher

Korbinian Strimmer

3 January 2024

Contents

Welcome **5**
About the author 5
License 6

1 Combinatorics **7**
1.1 Some basic mathematical notation 7
1.2 Number of permutations 7
1.3 Multinomial and binomial coefficient 8
1.4 De Moivre-Sterling approximation of the factorial 8

2 Probability **9**
2.1 Random variables 9
2.2 Probability mass and density function 10
2.3 Distribution function and quantile function 10
2.4 Expectation of a random variable 11
2.5 Jensen's inequality for the expectation 11
2.6 Probability as expectation 12
2.7 Moments and variance of a random variable 12
2.8 Distribution of sums of random variables 12
2.9 Transformation of random variables 13
2.10 Random vectors and their mean and variance 13
2.11 Correlation matrix 14

3 Univariate distributions **15**
3.1 Bernoulli distribution 15
3.2 Binomial distribution 16
3.3 Beta distribution 16
3.4 Normal distribution 18
3.5 Gamma distribution and special cases 20
3.6 Inverse gamma distribution 23
3.7 Location-scale t-distribution and special cases 24

4 Multivariate distributions **27**

4.1 Categorical distribution 27
4.2 Multinomial distribution 28
4.3 Multivariate normal distribution 29

Welcome

These supplementary notes aim to provide a quick refresher of some essentials in combinatorics and probability as well as offer an overview over selected univariate and multivariate distributions.

The notes are supporting information for the lecture notes of a number of statistical courses I am or have been teaching at the Department of Mathematics of the University of Manchester.

This includes the current modules:

- MATH27720 Statistics 2: Likelihood and Bayes and
- MATH38161 Multivariate Statistics,

as well as the retired module (not offered any more):

- MATH20802 Statistical Methods.

The Probability and Distribution Refresher notes were written by Korbinian Strimmer from 2018–2024. This version is from 3 January 2024.

The notes will be updated from time to time. To view the current version visit the online version of the Probability and Distribution Refresher notes.

You may also wish to download the Probability and Distribution Refresher notes as PDF in A4 format for printing (double page layout) or as 6x9 inch PDF for use on tablets (single page layout).

About the author

Hello! My name is Korbinian Strimmer and I am a Professor in Statistics. I am a member of the Statistics group at the Department of Mathematics of the University of Manchester. You can find more information about me on my home page.

License

Chapter 1

Combinatorics

1.1 Some basic mathematical notation

Summation:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n$$

Multiplication:

$$\prod_{i=1}^{n} x_i = x_1 \times x_2 \times \ldots \times x_n$$

Indicator function:

$$1_A = \begin{cases} 1 & \text{if } A \text{ is true} \\ 0 & \text{if } A \text{ is not true} \end{cases}$$

1.2 Number of permutations

The number of possible orderings, or permutations, of n distinct items is the number of ways to put n items in n bins with exactly one item in each bin. It is given by the factorial

$$n! = \prod_{i=1}^{n} i = 1 \times 2 \times \ldots \times n$$

where n is a positive integer. For $n = 0$ the factorial is defined as

$$0! = 1$$

as there is exactly one permutation of zero objects.

The factorial can also be obtained using the gamma function

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$$

which can be viewed as continuous version of the factorial with $\Gamma(x) = (x-1)!$ for any positive integer x.

1.3 Multinomial and binomial coefficient

The number of possible permutation of n items of K distinct types, with n_1 of type 1, n_2 of type 2 and so on, equals the number of ways to put n items into K bins with n_1 items in the first bin, n_2 in the second and so on. It is given by the **multinomial** coefficient

$$\binom{n}{n_1, \ldots, n_K} = \frac{n!}{n_1! \times n_2! \times \ldots \times n_K!}$$

with $\sum_{k=1}^K n_k = n$ and $K \leq n$. Note that it equals the number of permutation of all items divided by the number of permutations of the items in each bin (or of each type).

If all $n_k = 1$ and hence $K = n$ the multinomial coefficient reduces to the factorial.

If there are only two bins / types ($K = 2$) the multinomial coefficients becomes the **binomial coefficient**

$$\binom{n}{n_1} = \binom{n}{n_1, n-n_1} = \frac{n!}{n_1!(n-n_1)!}$$

which counts the number of ways to choose n_1 elements from a set of n elements.

1.4 De Moivre-Sterling approximation of the factorial

The factorial is frequently approximated by the following formula derived by Abraham de Moivre (1667–1754) and James Stirling (1692-1770)

$$n! \approx \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n}$$

or equivalently on logarithmic scale

$$\log n! \approx \left(n + \frac{1}{2}\right) \log n - n + \frac{1}{2} \log(2\pi)$$

The approximation is good for small n (but fails for $n = 0$) and becomes more and more accurate with increasing n. For large n the approximation can be simplified to

$$\log n! \approx n \log n - n$$

Chapter 2

Probability

2.1 Random variables

A **random variable** describes a random experiment. The set of all possible outcomes is the **sample space** or **state space** of the random variable and is denoted by $\Omega = \{\omega_1, \omega_2, \ldots\}$. The outcomes ω_i are the **elementary events**. The sample space Ω can be finite or infinite. Depending on type of outcomes the random variable is **discrete** or **continuous**.

An event $A \subseteq \Omega$ is a subset of Ω and thus itself a set composed of elementary events: $A = \{a_1, a_2, \ldots\}$. This includes as special cases the full set $A = \Omega$, the empty set $A = \emptyset$, and the elementary events $A = \omega_i$. The complementary event A^C is the complement of the set A in the set Ω so that $A^C = \Omega \backslash A = \{\omega_i \in \Omega : \omega_i \notin A\}$.

The probability of an event A is denoted by $\Pr(A)$. Essentially, to obtain this probability we need to count the elementary elements corresponding to A. To do this we assume as axioms of probability that

- $\Pr(A) \geq 0$, probabilities are positive,
- $\Pr(\Omega) = 1$, the certain event has probability 1, and
- $\Pr(A) = \sum_{a_i \in A} \Pr(a_i)$, the probability of an event equals the sum of its constituting elementary events a_i. This sum is taken over a finite or countable infinite number of elements.

This implies

- $\Pr(A) \leq 1$, i.e. probabilities all lie in the interval $[0, 1]$
- $\Pr(A^C) = 1 - \Pr(A)$, and
- $\Pr(\emptyset) = 0$

Assume now that we have two events A and B. The probability of the event "A and B" is then given by the probability of the set intersection $\Pr(A \cap B)$. Likewise

the probability of the event "A or B" is given by the probability of the set union $\Pr(A \cup B)$.

From the above it is clear that the definition and theory of probability is closely linked to set theory, and in particular to measure theory. Indeed, viewing probability as a special type of measure allows for an elegant treatment of both discrete and continuous random variables.

2.2 Probability mass and density function

To describe a random variable x with state space Ω we need a way to effectively store the probabilities of the corresponding elementary outcomes $x \in \Omega$. Note that for convenience we use the same symbol to denote the random variable and its elementary outcomes.

For a discrete random variable we define the event $A = \{x : x = a\} = \{a\}$ and get the probability

$$\Pr(A) = \Pr(x = a) = f(a)$$

directly from the **probability mass function** (PMF), here denoted by lower case f (but we frequently also use p or q). The PMF has the property that $\sum_{x\in\Omega} f(x) = 1$ and that $f(x) \in [0, 1]$.

For continuous random variables we need to use a **probability density function** (PDF) instead. We define the event $A = \{x : a < x \leq a + da\}$ as an infinitesimal interval and then assign the probability

$$\Pr(A) = \Pr(a < x \leq a + da) = f(a)da\,.$$

The PDF has the property that $\int_{x\in\Omega} f(x)dx = 1$ but in contrast to a PMF the density $f(x) \geq 0$ may take on values larger than 1.

2.3 Distribution function and quantile function

As alternative to using PMF/PDFs we may also use a **distribution function** to describe the random variable. This assumes an ordering exist among the elementary events so that we can define the event $A = \{x : x \leq a\}$ and compute its probability as

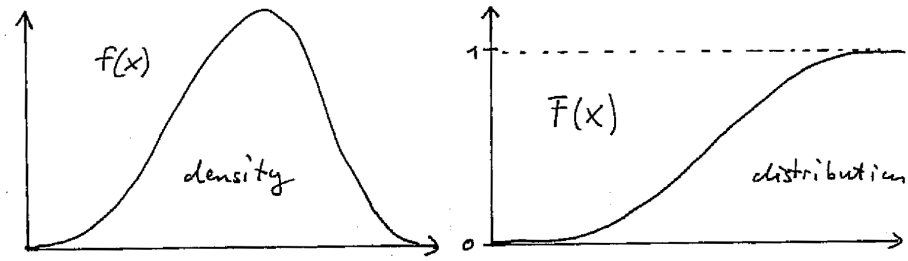
$$F(a) = \Pr(A) = \Pr(x \leq a) = \begin{cases} \sum_{x\in A} f(x) & \text{discrete case} \\ \int_{x\in A} f(x)dx & \text{continuous case} \end{cases}$$

This is also known **cumulative distribution function** (CDF) and is denoted by upper case F (or P and Q). By construction the distribution function is monotonically non-decreasing and its value ranges from 0 to 1. With its help we can compute the probability of an interval set such as

$$\Pr(a < x \leq b) = F(b) - F(a)\,.$$

The inverse of the distribution function $y = F(x)$ is the **quantile function** $x = F^{-1}(y)$. The 50% quantile $F^{-1}\left(\frac{1}{2}\right)$ is called the **median**.

If the random variable x has distribution function F we write $x \sim F$.



2.4 Expectation of a random variable

The expected value $\mathrm{E}(x)$ of a random variable is defined as the weighted average over all possible outcomes, with the weight given by the PMF / PDF $f(x)$:

$$\mathrm{E}_F(x) = \begin{cases} \sum_{x\in\Omega} f(x)x & \text{discrete case} \\ \int_{x\in\Omega} f(x)x\,dx & \text{continuous case} \end{cases}$$

Note the notation to emphasise that the expectation is taken with regard to the distribution F. The subscript F is usually left out if there are no ambiguities. Furthermore, because the sum or integral may diverge the expectation is not necessarily always defined (in contrast to quantiles).

The expected value of a function of a random variable $h(x)$ is obtained similarly:

$$\mathrm{E}_F(h(x)) = \begin{cases} \sum_{x\in\Omega} f(x)h(x) & \text{discrete case} \\ \int_{x\in\Omega} f(x)h(x)dx & \text{continuous case} \end{cases}$$

This is called the "law of the unconscious statistician", or short LOTUS. Again, to highlight that the random variable x has distribution F we write $\mathrm{E}_F(h(x))$.

2.5 Jensen's inequality for the expectation

If $h(\boldsymbol{x})$ is a *convex* function then the following inequality holds:

$$\mathrm{E}(h(\boldsymbol{x})) \geq h(\mathrm{E}(\boldsymbol{x}))$$

Recall: a con**ve**x function (such as x^2) has the shape of a "**va**lley".

2.6 Probability as expectation

Probability itself can also be understood as an expectation. For an event A we can define a corresponding indicator function $1_{x \in A}$ for an elementary element x to be part of A. From the above it then follows

$$\mathrm{E}(1_{x \in A}) = \Pr(A)\,,$$

Interestingly, one can develop the whole theory of probability from this perspective. [1]

2.7 Moments and variance of a random variable

The moments of a random variable are defined as follows:

- Zeroth moment: $\mathrm{E}(x^0) = 1$ by construction of PDF and PMF,
- First moment: $\mathrm{E}(x^1) = \mathrm{E}(x) = \mu$, the mean,
- Second moment: $\mathrm{E}(x^2)$
- The variance is the second moment centred about the mean μ:

$$\mathrm{Var}(x) = \mathrm{E}((x - \mu)^2) = \sigma^2$$

- The variance can also be computed by $\mathrm{Var}(x) = \mathrm{E}(x^2) - \mathrm{E}(x)^2$. This provides an example of Jensen's inequality, with $\mathrm{E}(x^2) = \mathrm{E}(x)^2 + \mathrm{Var}(x) \geq \mathrm{E}(x)^2$.

A distribution does not necessarily need to have any finite first or higher moments. An example is the Cauchy distribution that does not have a mean or variance (or any other higher moment).

2.8 Distribution of sums of random variables

The sum of two normal random variables is also normal (with the appropriate mean and variance).

The **central limit theorem**, first postulated by Abraham de Moivre (1667–1754), asserts that in many cases the distribution of the sum of identically distributed random variables converges to a normal distribution, even if the individual random variables are not normal.

For example, as a result the binomial distribution (as sum of Bernoulli random variables) can be approximated by a normal distribution.

[1] Whittle, P. 2000. Probability via Expectation (3rd ed.). Springer. https://doi.org/10.1007/978-1-4612-0509-8

2.9 Transformation of random variables

Linear transformation of random variables: if a and b are constants and x is a random variable, then the random variable $y = a + bx$ has mean $\mathrm{E}(y) = a + b\mathrm{E}(x)$ and variance $\mathrm{Var}(y) = b^2\mathrm{Var}(x)$.

For a general invertible coordinate transformation $y = h(x) = y(x)$ the backtransformation is $x = h^{-1}(y) = x(y)$.

The transformation of the infinitesimal volume element is $dy = \left|\frac{dy}{dx}\right| dx$.

The transformation of the density is $f_y(y) = \left|\frac{dx}{dy}\right| f_x(x(y))$.

Note that $\left|\frac{dx}{dy}\right| = \left|\frac{dy}{dx}\right|^{-1}$.

2.10 Random vectors and their mean and variance

Instead of scalar random variables one often also considers random vectors and also random matrices.

For a random vector $\boldsymbol{x} = (x_1, x_2, ..., x_d)^T$ the mean $\mathrm{E}(\boldsymbol{x}) = \boldsymbol{\mu}$ is simply comprised of the means of its components, i.e. $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^T$. Thus, the mean of a random vector of dimension is a vector of of the same length.

The variance of a random vector of length d, however, is not a vector but a matrix of size $d \times d$. This matrix is called the **covariance matrix**:

$$\begin{aligned}\mathrm{Var}(\boldsymbol{x}) &= \underbrace{\boldsymbol{\Sigma}}_{d\times d} = (\sigma_{ij}) = \begin{pmatrix} \sigma_{11} & \ldots & \sigma_{1d} \\ \vdots & \ddots & \vdots \\ \sigma_{d1} & \ldots & \sigma_{dd} \end{pmatrix} \\ &= \mathrm{E}\left(\underbrace{(\boldsymbol{x} - \boldsymbol{\mu})}_{d\times 1} \underbrace{(\boldsymbol{x} - \boldsymbol{\mu})^T}_{1\times d} \right) \\ &= \mathrm{E}(\boldsymbol{x}\boldsymbol{x}^T) - \boldsymbol{\mu}\boldsymbol{\mu}^T\end{aligned}$$

The entries of the covariance matrix $\sigma_{ij} = \mathrm{Cov}(x_i, x_j)$ describe the covariance between the random variables x_i and x_j. The covariance matrix is symmetric, hence $\sigma_{ij} = \sigma_{ji}$. The diagonal entries $\sigma_{ii} = \mathrm{Cov}(x_i, x_i) = \mathrm{Var}(x_i) = \sigma_i^2$ correspond to the variances of the components of $\boldsymbol{x}$. The covariance matrix is by construction **positive semi-definite**, i.e. the eigenvalues of $\boldsymbol{\Sigma}$ are all positive or equal to zero.

However, wherever possible one will aim to use models with non-singular covariance matrices, with all eigenvalues positive, so that the covariance matrix is invertible.

For univariate x and scalar constant a the variance of ax equals $\text{Var}(ax) = a^2\text{Var}(x)$. For a random vector $\boldsymbol{x}$ of dimension d and constant matrix $\boldsymbol{A}$ of dimension $m \times d$ this generalises to $\text{Var}(\boldsymbol{A}x) = \boldsymbol{A}\text{Var}(\boldsymbol{x})\boldsymbol{A}^T$.

2.11 Correlation matrix

A covariance matrix can be factorised into the product

$$\boldsymbol{\Sigma} = \boldsymbol{V}^{\frac{1}{2}}\boldsymbol{P}\boldsymbol{V}^{\frac{1}{2}}$$

where $\boldsymbol{V}$ is a diagonal matrix containing the variances

$$\boldsymbol{V} = \begin{pmatrix} \sigma_{11} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_{dd} \end{pmatrix}$$

and the matrix $\boldsymbol{P}$ ("upper case rho") is the symmetric **correlation matrix**

$$\boldsymbol{P} = (\rho_{ij}) = \begin{pmatrix} 1 & \dots & \rho_{1d} \\ \vdots & \ddots & \vdots \\ \rho_{d1} & \dots & 1 \end{pmatrix} = \boldsymbol{V}^{-\frac{1}{2}}\boldsymbol{\Sigma}\boldsymbol{V}^{-\frac{1}{2}}$$

Thus, the correlation between x_i and x_j is defined as

$$\rho_{ij} = \text{Cor}(x_i, x_j) = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$$

Chapter 3

Univariate distributions

3.1 Bernoulli distribution

The **Bernoulli distribution** Ber(θ) is simplest of all distributions. It is named after Jacob Bernoulli (1655-1705) who also discovered the law of large numbers.

It describes a discrete binary random variable with two states $x = 0$ ("failure") and $x = 1$ ("success"), where the parameter $\theta \in [0, 1]$ is the probability of "success". Often the Bernoulli distribution is also referred to as "coin tossing" model with the two outcomes "heads" and "tails".

Correspondingly, the probability mass function of Ber(θ) is

$$p(x = 0) = \Pr(\text{"failure"}) = 1 - \theta$$

and

$$p(x = 1) = \Pr(\text{"success"}) = \theta$$

A compact way to write the PMF of the Bernoulli distribution is

$$p(x|\theta) = \theta^x (1 - \theta)^{1-x}$$

The log PMF is

$$\log p(x|\theta) = x \log \theta + (1 - x) \log(1 - \theta)$$

If a random variable x follows the Bernoulli distribution we write

$$x \sim \text{Ber}(\theta)\,.$$

The expected value is $\mathrm{E}(x) = \theta$ and the variance is $\mathrm{Var}(x) = \theta(1 - \theta)$.

3.2 Binomial distribution

Closely related to the Bernoulli distribution is the **binomial distribution** $\text{Bin}(n, \theta)$ which results from repeating a Bernoulli experiment n times and counting the number of successes among the n trials (without keeping track of the ordering of the experiments). Thus, if $x_1, \ldots, x_n$ are n independent $\text{Ber}(\theta)$ random variables then $y = \sum_{i=1}^{n} x_i$ is distributed as $\text{Bin}(n, \theta)$.

Its probability mass function is:

$$p(y|n, \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

for $y \in \{0, 1, 2, \ldots, n\}$. The binomial coefficient $\binom{n}{y}$ is needed to account for the multiplicity of ways (orderings of samples) in which we can observe y sucesses.

The expected value is $\text{E}(y) = n\theta$ and the variance is $\text{Var}(y) = n\theta(1-\theta)$.

If a random variable y follows the binomial distribution we write

$$y \sim \text{Bin}(n, \theta)$$

For $n = 1$ it reduces to the Bernoulli distribution $\text{Ber}(\theta)$.

In R the PMF of the binomial distribution is called `dbinom()`. The binomial coefficient itself is computed by `choose()`.

3.3 Beta distribution

3.3.1 Standard parameterisation

The density of the beta distribution $\text{Beta}(\alpha, \beta)$ is

$$p(x|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

with $x \in [0, 1]$ and $\alpha > 0$ and $\beta > 0$. The density depends on the beta function $B(z_1, z_1) = \frac{\Gamma(z_1)\Gamma(z_2)}{\Gamma(z_1+z_2)}$ which in turn is defined via Euler's gamma function $\Gamma(x)$. Note that $\Gamma(x) = (x-1)!$ for any positive integer x.

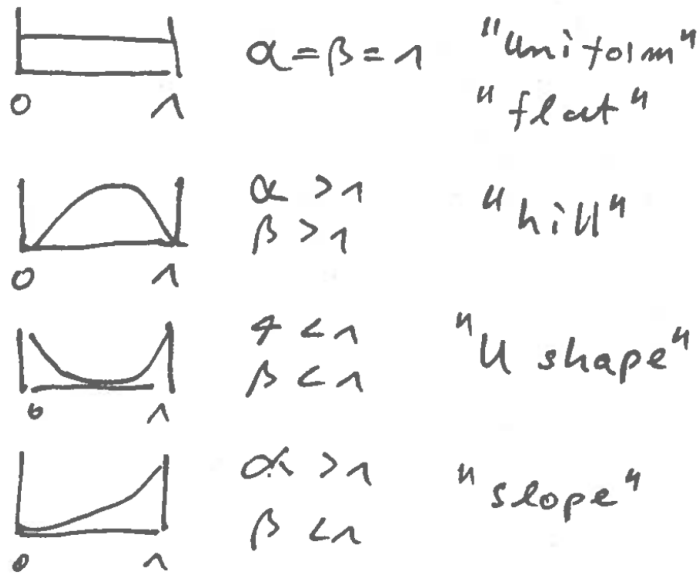
The mean of the beta distribution is

$$\text{E}(x) = \frac{\alpha}{\alpha + \beta}$$

and its variance is

$$\text{Var}(x) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

The beta distribution is very flexible and can assume a number of different shapes, depending on the value of α and β:



3.3.2 Mean parameterisation

A useful reparameterisation Beta(μ, k) of the beta distribution is in terms of a mean parameter $\mu \in [0, 1]$ and a concentration parameter $k > 0$. These are given by

$$k = \alpha + \beta$$

and

$$\mu = \frac{\alpha}{\alpha + \beta}$$

The original parameters can be recovered by

$$\alpha = \mu k$$

and

$$\beta = (1 - \mu)k$$

The mean and variance of the beta distribution expressed in terms of μ and k are

$$\mathrm{E}(x) = \mu$$

and

$$\mathrm{Var}(x) = \frac{\mu(1 - \mu)}{k + 1}$$

With increasing concentration parameter k the variance decreases and thus the probability mass becomes more concentrated around the mean.

3.4 Normal distribution

The **normal distribution** is the most important continuous probability distribution. It is also called **Gaussian distribution** named after Carl Friedrich Gauss (1777–1855).

The univariate normal distribution $N(\mu, \sigma^2)$ has two parameters μ (location) and σ^2 (scale):

$$x \sim N(\mu, \sigma^2)$$

with mean

$$\mathrm{E}(x) = \mu$$

and variance

$$\mathrm{Var}(x) = \sigma^2$$

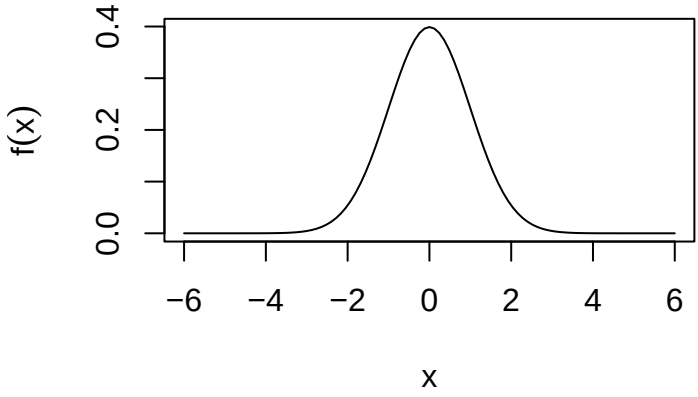
Probability density function (PDF):

$$p(x|\mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

In R the density function is called `dnorm()`.

The standard normal distribution is $N(0, 1)$ with mean 0 and variance 1.

Plot of the PDF of the standard normal:

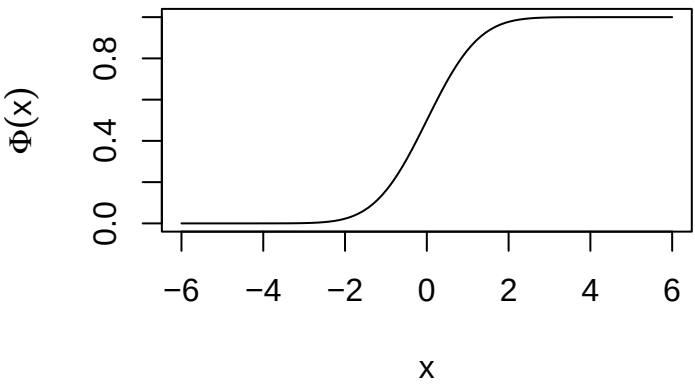


The cumulative distribution function (CDF) of the standard normal $N(0, 1)$ is

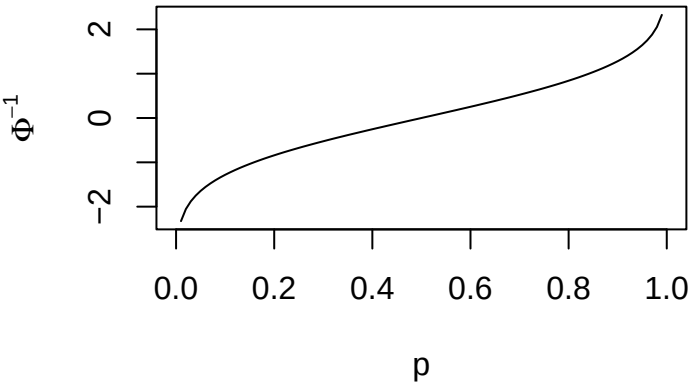
$$\Phi(x) = \int_{-\infty}^{x} p(x'|\mu = 0, \sigma^2 = 1)dx'$$

There is no analytic expression for $\Phi(x)$. In R the function is called `pnorm()`.

Plot of the CDF of the standard normal:



The inverse $\Phi^{-1}(p)$ is called the quantile function of the standard normal. In R the function is called `qnorm()`.



3.5 Gamma distribution and special cases

The gamma distribution is widely used in statistics, and also appears in various parameterisations and under some other names, such as univariate Wishart and scaled chi-squared distribution

3.5.1 Standard parameterisation

Another important continuous distribution is the gamma distribution $\text{Gam}(\alpha, \theta)$. It has two parameters $\alpha > 0$ (shape) and $\theta > 0$ (scale):

$$x \sim \text{Gam}(\alpha, \theta)$$

with mean

$$\text{E}(x) = \alpha\theta$$

and variance

$$\text{Var}(x) = \alpha\theta^2$$

The gamma distribution is also often used with a rate parameter $\beta = 1/\theta$ (so one needs to pay attention which parameterisation is used).

Probability density function (PDF):

$$p(x|\alpha, \theta) = \frac{1}{\Gamma(\alpha)\theta^\alpha} x^{\alpha-1} e^{-x/\theta}$$

The density of the gamma distribution is available in the R function `dgamma()`. The cumulative density function is `pgamma()` and the quantile function is `qgamma()`.

3.5.2 Wishart parameterisation and scaled chi-squared distribution

The gamma distribution is often used with a different set of parameters $k = 2\alpha$ and $s^2 = \theta/2$ (hence conversely $\alpha = k/2$ and $\theta = 2s^2$). In this form it is known as **univariate or one-dimensional Wishart distribution**

$$W_1\left(s^2, k\right)$$

named after John Wishart (1898–1954). In the Wishart parameterisation the mean is

$$\mathrm{E}(x) = ks^2$$

and the variance

$$\mathrm{Var}(x) = 2ks^4$$

Another name for the one-dimensional Wishart distribution with exactly the same parameterisation is **scaled chi-squared distribution** denoted as

$$s^2\chi_k^2$$

Finally, we also often employ the Wishart distribution in **mean parameterisation** $W_1\left(s^2 = \mu/k, k\right)$ with parameters $\mu = ks^2$ and k (and thus $\theta = 2\mu/k$). In this parameterisation the mean is

$$\mathrm{E}(x) = \mu$$

and the variance

$$\mathrm{Var}(x) = \frac{2\mu^2}{k}$$

3.5.3 Construction as sum of squared normals

A gamma distributed variable can be constructed as follows. Assume k independent normal random variables with mean 0 and variance s^2:

$$z_1, z_2, \ldots, z_k \sim N(0, s^2)$$

Then the sum of the squares

$$x = \sum_{i=1}^{k} z_i^2$$

follows

$$x \sim \sigma^2\chi_k^2 = W_1\left(s^2, k\right)$$

or equivalently

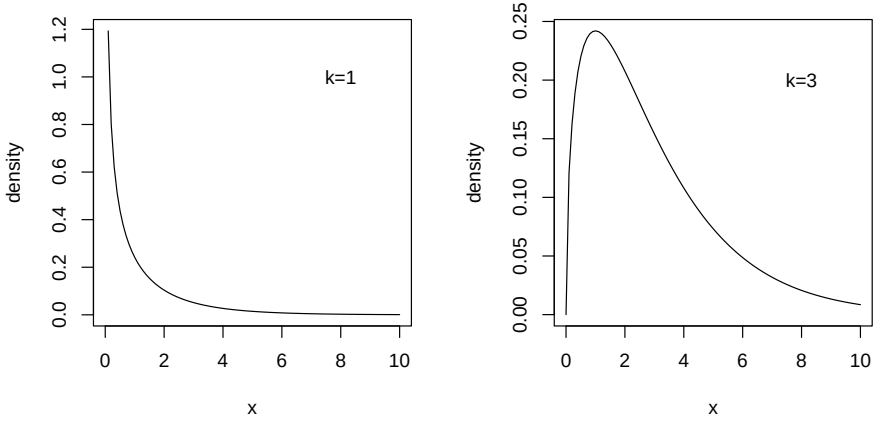
$$x \sim \mathrm{Gam}\left(\alpha = \frac{k}{2}, \theta = 2s^2\right)$$

3.5.4 Chi-squared distribution

The **chi-squared distribution** χ^2_k is a special one-parameter restriction of the gamma resp. Wishart distribution obtained when setting $s^2 = 1$ or, equivalently, $\theta = 2$ or $\mu = k$.

It has mean $\mathrm{E}(x) = k$ and variance $\mathrm{Var}(x) = 2k$. The chi-squared distribution χ^2_k equals $\mathrm{Gam}(\alpha = k/2, \theta = 2) = W_1(1, k)$.

Here is a plot of the density of the chi-squared distribution for degrees of freedom $k = 1$ and $k = 3$:



In R the density of the chi-squared distribution is given by `dchisq()`. The cumulative density function is `pchisq()` and the quantile function is `qchisq()`.

3.5.5 Exponential distribution

The **exponential distribution** $\mathrm{Exp}(\theta)$ with scale parameter θ is another special one-parameter restriction of the gamma distribution with shape parameter set to $\alpha = 1$ (or equivalently $k = 2$).

It thus equals $\mathrm{Gam}(\alpha = 1, \theta) = W_1(s^2 = \theta/2, k = 2)$. It has mean θ and variance θ^2.

Just like the gamma distribution the exponential distribution is also often specified using a rate parameter $\beta = 1/\theta$ instead of a scale parameter θ.

In R the command `dexp()` returns the density of the exponential distribution, `pexp()` is the corresponding cumulative density function and `qexp()` is the quantile function.

3.6 Inverse gamma distribution

Also know as inverse univariate Wishart distribution.

3.6.1 Standard parameterisation

The inverse gamma (IG) distribution Inv-Gam(α, β) has density

$$\frac{\beta^\alpha}{\Gamma(\alpha)}(1/x)^{\alpha+1}e^{-\beta/x}$$

with two parameters $\alpha > 0$ (shape parameter) and $\beta > 0$ (scale parameter) and support $x > 0$.

The mean of the inverse gamma distribution is

$$\mathrm{E}(x) = \frac{\beta}{\alpha - 1}$$

and the variance

$$\mathrm{Var}(x) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$$

Thus, for the mean to exist we have the restriction $\alpha > 1$ and for the variance to exist $\alpha > 2$.

The IG distribution is closely linked with the gamma distribution. If $x \sim$ Inv-Gam(α, β) is IG-distributed then the inverse of x is gamma distributed:

$$\frac{1}{x} \sim \mathrm{Gam}(\alpha, \theta = \beta^{-1})$$

where α is the shared shape parameter and θ the scale parameter of the gamma distribution.

3.6.2 Wishart parameterisation

The inverse gamma distribution is frequently used with a different set of parameters $\psi = 2\beta$ (scale parameter) and $\nu = 2\alpha$ (shape parameter), or conversely $\alpha = \nu/2$ and $\beta = \psi/2$. In this form it is called **one-dimensional inverse Wishart distribution** $W_1^{-1}(\psi, \nu)$ with mean and variance given by

$$\mathrm{E}(x) = \frac{\psi}{\nu - 2} = \mu$$

for $\nu > 2$ and

$$\mathrm{Var}(x) = \frac{2\psi^2}{(\nu-4)(\nu-2)^2} = \frac{2\mu^2}{\nu-4}$$

for $\nu > 4$.

Instead of ψ and ν we may also equivalently use μ and $\kappa = \nu - 2$ as parameters for the inverse Wishart distribution, so that $W_1^{-1}(\psi = \kappa\mu, \nu = \kappa + 2)$ has mean

$$\mathrm{E}(x) = \mu$$

with $\kappa > 0$ and the variance is

$$\mathrm{Var}(x) = \frac{2\mu^2}{\kappa - 2}$$

with $\kappa > 2$. This **mean parameterisation** is useful when employing the IG distribution as prior and posterior.

Finally, with $W_1^{-1}(\psi = \nu\tau^2, \nu)$, where $\tau^2 = \mu\frac{\kappa}{\kappa+2} = \frac{\psi}{\nu}$ is a biased mean parameter, we get the **scaled inverse chi-squared distribution** $\tau^2\text{Inv-}\chi^2_\nu$ with

$$\mathrm{E}(x) = \tau^2\frac{\nu}{\nu - 2}$$

for $\nu > 2$ and

$$\mathrm{Var}(x) = \frac{2\tau^4}{\nu - 4}\frac{\nu^2}{(\nu - 2)^2}$$

for $\nu > 4$.

The inverse univariate Wishart and univariate Wishart distributions are linked. If $x \sim W_1^{-1}(\psi, \nu)$ is inverse-Wishart distributed then the inverse of x is Wishart distributed with inverted scale parameter:

$$\frac{1}{x} \sim W_1(s^2 = \psi^{-1}, k = \nu)$$

where k is the shape parameter and s^2 the scale parameter of the Wishart distribution.

3.7 Location-scale t-distribution and special cases

3.7.1 Location-scale t-distribution

The location-scale t-distribution $\mathrm{lst}(\mu, \tau^2, \nu)$ is a generalisation of the normal distribution. It has an additional parameter $\nu > 0$ (degrees of freedom) that controls the probability mass in the tails. For small values of ν the distribution is heavy-tailed — indeed so heavy that for $\nu \leq 1$ even the mean is not defined and for $\nu \leq 2$ the variance is undefined.

The probability density of $\mathrm{lst}(\mu, \tau^2, \nu)$ is

$$p(x|\mu, \tau^2, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu\tau^2}\,\Gamma(\frac{\nu}{2})}\left(1 + \frac{(x-\mu)^2}{\nu\tau^2}\right)^{-(\nu+1)/2}$$

The mean is (for $\nu > 1$)

$$\mathrm{E}(x) = \mu$$

and the variance (for $\nu > 2$)

$$\mathrm{Var}(x) = \tau^2 \frac{\nu}{\nu - 2}$$

For $\nu \to \infty$ the location-scale t-distribution $\mathrm{lst}(\mu, \tau^2, \nu)$ becomes the normal distribution $N(\mu, \tau^2)$.

In the R `extraDistr` package the command `dlst()` returns the density of the location-scale t-distribution, `plst()` is the corresponding cumulative density function and `qlst()` is the quantile function.

3.7.2 Location-scale t-distribution as compound distribution

Suppose that

$$x|s^2 \sim N(\mu, s^2)$$

with corresponding density $p(x|s^2)$ and mean $\mathrm{E}(x|s^2) = \mu$ and variance $\mathrm{Var}(x|s^2) = s^2$.

Now let the variance s^2 be distributed as inverse gamma / inverse Wishart

$$s^2 \sim W^{-1}(\psi = \kappa\sigma^2, \nu = \kappa + 2) = W^{-1}(\psi = \tau^2\nu, \nu)$$

with corresponding density $p(s^2)$ and mean $\mathrm{E}(s^2) = \sigma^2 = \tau^2\nu/(\nu - 2)$. Note we use here both the mean parameterisation (σ^2, κ) and the inverse chi-squared parameterisation (τ^2, ν).

The joint density for x and s^2 is $p(x, s^2) = p(x|s^2)p(s^2)$. We are interested in the marginal density for x:

$$p(x) = \int p(x, s^2) ds^2 = \int p(s^2) p(x|s^2) ds^2$$

This is a compound distribution of a normal with fixed mean μ and variance s^2 varying according the inverse gamma distribution. Calculating the integral results in the location-scale t-distribution with parameters

$$x \sim \mathrm{lst}\left(\mu, \sigma^2 \frac{\kappa}{\kappa + 2}, \kappa + 2\right) = \mathrm{lst}\left(\mu, \tau^2, \nu\right)$$

with mean

$$\mathrm{E}(x) = \mu$$

and variance

$$\mathrm{Var}(x) = \sigma^2 = \tau^2 \frac{\nu}{\nu - 2}$$

From the law of total expectation and variance we can also directly verify that

$$\mathrm{E}(x) = \mathrm{E}(\mathrm{E}(x|s^2)) = \mu$$

and

$$\mathrm{Var}(x) = \mathrm{E}(\mathrm{Var}(x|s^2)) + \mathrm{Var}(\mathrm{E}(x|s^2)) = \mathrm{E}(s^2) = \sigma^2 = \tau^2 \frac{\nu}{\nu - 2}$$

3.7.3 Student's t-distribution

For $\mu = 0$ and $\tau^2 = 1$ the location-scale t-distribution becomes the Student's t-distribution t_ν with mean 0 (for $\nu > 1$) and variance $\frac{\nu}{\nu-2}$ (for $\nu > 2$).

It can thus be viewed as a generalisation of the standard normal distribution $N(0, 1)$.

If $y \sim t_\nu$ then $x = \mu + \tau y$ is distributed as $x \sim \mathrm{lst}(\mu, \tau^2, \nu)$.

For $\nu \longrightarrow \infty$ the t-distribution becomes equal to $N(0, 1)$.

In R the command `dt()` returns the density of the t-distribution, `pt()` is the corresponding cumulative density function and `qt()` is the quantile function.

3.7.4 Cauchy and standard Cauchy distribution

For $\nu = 1$ the location-scale t-distribution becomes the Cauchy distribution $\mathrm{Cau}(\mu, \tau)$ with density $p(x|\mu, \tau) = \frac{\tau}{\pi(\tau^2 + (x-\mu)^2)}$.

For $\nu = 1$ the t-distribution becomes the standard Cauchy distribution $\mathrm{Cau}(0, 1)$ with density $p(x) = \frac{1}{\pi(1+x^2)}$.

Chapter 4

Multivariate distributions

4.1 Categorical distribution

The **categorical distribution** is a generalisation of the Bernoulli distribution from two classes to K classes.

The categorical distribution $\text{Cat}(\boldsymbol{\pi})$ describes a discrete random variable with K states ("categories", "classes", "bins") where the parameter vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)^T$ specifies the probability of each of class so that Pr("class k") $= \pi_k$. The parameters satisfy $\pi_k \in [0, 1]$ and $\sum_{k=1}^{K} \pi_k = 1$, hence there are $K - 1$ independent parameters in a categorical distribution (and not K).

There are two main ways to numerically represent "class k":

i) by "integer encoding", i.e. by the corresponding integer k.
ii) by "one hot encoding", i.e. by an indicator vector $\boldsymbol{x} = (x_1, \ldots, x_K)^T = (0, 0, \ldots, 1, \ldots, 0)^T$ containing zeros everywhere except for the element $x_k = 1$ at position k. Thus all $x_k \in \{0, 1\}$ and $\sum_{k=1}^{K} x_k = 1$.

In the following we use "one hot encoding". Therefore sampling from a categorical distribution with parameters $\boldsymbol{\pi}$

$$\boldsymbol{x} \sim \text{Cat}(\boldsymbol{\pi})$$

yields a random index vector $\boldsymbol{x}$.

The corresponding probability mass function (PMF) can be written conveniently in terms of x_k as

$$p(\boldsymbol{x}|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{x_k} = \begin{cases} \pi_k & \text{if } x_k = 1 \end{cases}$$

and the log PMF as

$$\log p(\boldsymbol{x}|\boldsymbol{\pi}) = \sum_{k=1}^{K} x_k \log \pi_k = \begin{cases} \log \pi_k & \text{if } x_k = 1 \end{cases}$$

In order to be more explicit that the categorical distribution has $K-1$ and not K parameters we rewrite the log-density with $\pi_K = 1 - \sum_{k=1}^{K-1} \pi_k$ and $x_K = 1 - \sum_{k=1}^{K-1} x_k$ as

$$\begin{aligned} \log p(\boldsymbol{x}|\boldsymbol{\pi}) &= \sum_{k=1}^{K-1} x_k \log \pi_k + x_K \log \pi_K \\ &= \sum_{k=1}^{K-1} x_k \log \pi_k + \left(1 - \sum_{k=1}^{K-1} x_k\right) \log \left(1 - \sum_{k=1}^{K-1} \pi_k\right) \end{aligned}$$

Note that there is no particular reason to choose π_K as dependent of the probabilities of the other classes, in its place any other of the π_k may be selected.

For $K = 2$ the categorical distribution reduces to the Bernoulli $\text{Ber}(\theta)$ distribution, with $\pi_1 = \theta$ and $\pi_2 = 1 - \theta$.

The expected value is $\text{E}(\boldsymbol{x}) = \boldsymbol{\pi}$, in component notation $\text{E}(x_k) = \pi_k$. The covariance matrix is $\text{Var}(\boldsymbol{x}) = \text{Diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T$, which in component notation is $\text{Var}(x_i) = \pi_i(1 - \pi_i)$ and $\text{Cov}(x_i, x_j) = -\pi_i \pi_j$.

The form of the categorical covariance matrix follows directly from the definition of the variance $\text{Var}(\boldsymbol{x}) = \text{E}(\boldsymbol{x}\boldsymbol{x}^T) - \text{E}(\boldsymbol{x})\text{E}(\boldsymbol{x})^T$ and noting that $x_i^2 = x_i$ and $x_i x_j = 0$ if $i \neq j$. Furthermore, the categorical covariance matrix is singular by construction, as the K random variables $x_1, \ldots, x_K$ are dependent through the constraint $\sum_{k=1}^{K} x_k = 1$.

4.2 Multinomial distribution

The **multinomial distribution** $\text{Mult}(n, \boldsymbol{\pi})$ arises from repeated categorical sampling, in the same fashion as the binomial distribution arises from repeated Bernoulli sampling. Thus, if $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are n independent $\text{Cat}(\boldsymbol{\pi})$ random categorical variables then $\boldsymbol{y} = \sum_{i=1}^{n} \boldsymbol{x}_i$ is distributed as $\text{Mult}(n, \boldsymbol{\pi})$.

The corresponding PMF describes the probability of a pattern $y_1, \ldots, y_K$ of samples distributed across K classes (with $n = \sum_{k=1}^{K} y_k$):

$$p(\boldsymbol{y}|n, \theta) = \binom{n}{y_1, \ldots, y_n} \prod_{k=1}^{K} \pi_k^{y_k}$$

where $\binom{n}{y_1, \ldots, y_n}$ is the multinomial coefficient.

The expected value is $\mathrm{E}(\boldsymbol{y}) = n\boldsymbol{\pi}$, in component notation $\mathrm{E}(y_k) = n\pi_k$. The covariance matrix is $\mathrm{Var}(\boldsymbol{y}) = n\mathrm{Diag}(\boldsymbol{\pi}) - n\boldsymbol{\pi}\boldsymbol{\pi}^T$, which in component notation is $\mathrm{Var}(x_i) = n\pi_i(1-\pi_i)$ and $\mathrm{Cov}(x_i, x_j) = -n\pi_i\pi_j$.

4.3 Multivariate normal distribution

The univariate normal distribution for a random scalar x generalises to the **multivariate normal distribution** for a random vector $\boldsymbol{x} = (x_1, x_2, ..., x_d)^T$.

If $\boldsymbol{x}$ follows a multivariate normal distribution we write

$$\boldsymbol{x} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

where $\boldsymbol{\mu}$ is the mean (location) parameter and $\boldsymbol{\Sigma}$ the variance (scale) parameter.

The corresponding density is

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \det(2\pi\boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\left(\underbrace{-\frac{1}{2} \underbrace{(\boldsymbol{x}-\boldsymbol{\mu})^T}_{1\times d} \underbrace{\boldsymbol{\Sigma}^{-1}}_{d\times d} \underbrace{(\boldsymbol{x}-\boldsymbol{\mu})}_{d\times 1}}_{1\times 1=\text{scalar!}} \right)$$

As $\det(2\pi\boldsymbol{\Sigma})^{-\frac{1}{2}} = \det(2\pi\boldsymbol{I}_d)^{-\frac{1}{2}} \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} = (2\pi)^{-d/2} \det(\boldsymbol{\Sigma})^{-\frac{1}{2}}$ the density can also be written as

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{d}{2}} \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

with explicit occurance of the dimension d.

The expectation of $\boldsymbol{x}$ is $\mathrm{E}(\boldsymbol{x}) = \boldsymbol{\mu}$ and the variance is $\mathrm{Var}(\boldsymbol{x}) = \boldsymbol{\Sigma}$.

For $d = 1$ the random vector $\boldsymbol{x} = x$ is a scalar and $\boldsymbol{\mu} = \mu$ and $\boldsymbol{\Sigma} = \sigma^2$ and the multivariate normal density reduces to the univariate normal density.