Probability and Distribution Refresher

Korbinian Strimmer

2 March 2024

Contents

Welcome

The Probability and Distribution Refresher notes were written by Korbinian Strimmer from 2018–2024. This version is from 2 March 2024.

If you have any questions, comments, or corrections then please email me at korbinian.strimmer@manchester.ac.uk.

Updates

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).

License

Preface

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.

About the notes

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

The notes are supporting information for a number of lecture notes 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.

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:

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

Scalar: plain type, typically lower case (x, θ), sometimes upper case (K).

Vector: bold type, lower case ($\boldsymbol{x}$, $\boldsymbol{\theta}$).

Matrix: bold type, upper case ($\boldsymbol{X}$, $\boldsymbol{\Sigma}$).

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 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

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

1.4 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.

For large n and n_k we can apply the De Moivre-Sterling approximation to the multinomial cofficient, yielding

$$\log \binom{n}{n_1, \ldots, n_K} = -n \sum_{k=1}^{K} \frac{n_k}{n} \log \left(\frac{n_k}{n} \right)$$

Note this is n times the Shannon entropy of a categorical distribution with n_k/n as class probabilities.

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 $x \in \Omega$.

For simplicity of notation we use the same symbol to denote the random variable and its elementary outcomes. [1] This convention greatly facilitates working with random vectors and matrices (multivariate statistics) and random parameters (Bayesian statistics).

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.

The set of all x for which $f(x)$ is positive is called the **support** of the PDF/PMF.

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

[1] For scalar random variables many texts use upper case to designate the random variable and lower case for its realisations. However, this notation quickly breaks down for random vectors and random matrices and for distributions describing the uncertainty of parameters. Hence, we use upper case primarily to indicate a matrix quantity (in bold type). Upper case (in plain type) may denote sets and some scalar quantities traditionally written in upper case (e.g. R^2, K). If a quantity is random we will always specify this explicitly in the context.

elementary events so that we can define the event $A = \{x : x \leq a\}$ and compute its probability as

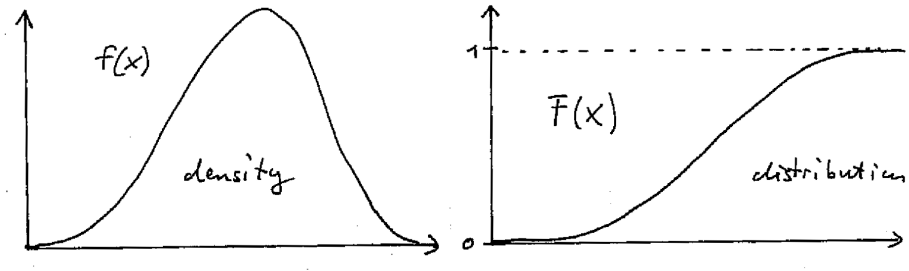
$$F(a) = \Pr(A) = \Pr(x \leq a) = \begin{cases} \sum_{x \in A} f(x) & \text{discrete case} \\ \int_{x \in A} f(x)dx & \text{continuous case} \end{cases}$$

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 Families of distributions

A distribution F_θ with a parameter θ constitutes a **distribution family** collecting all the distributions corresponding to particular instances of the parameter. The parameter θ therefore acts as an index of the distributions contained in the family.

The corresponding density (PDF) or probability mass function (PMF) is written either as $f_\theta(x)$, $f(x;\theta)$ or $f(x|\theta)$. The latter form is the most general is it suggests that the parameter θ may potentially also have its own distribution, with a joint density formed by $f(x,\theta) = f(x|\theta)f(\theta)$.

Note that any parameterisation is generally not unique, as a one-to-one transformation of θ will yield another equivalent index to the same distribution family. Typically, for most commonly used distribution families there are several standard parameterisations. Often we use those parameterisations where the parameters can be interpreted easily (e.g. in terms of moments).

If for any pair of different parameter values $\theta_1 \neq \theta_2$ we get distinct distributions with $F_{\theta_1} \neq F_{\theta_2}$ then the distribution family F_θ is said to be **identifiable** under the parameter θ.

2.5 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 $f(x)$:

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

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:

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

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 $\mathrm{E}_F(h(x))$.

2.6 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".

2.7 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. [2]

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

2.8 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).

2.9 Random vectors and their mean and variance

In addition to scalar random variables we often make use of random vectors and also random matrices. [3]

For a random vector $\boldsymbol{x} = (x_1, x_2, ..., x_d)^T \sim F$ the mean $\mathrm{E}(\boldsymbol{x}) = \boldsymbol{\mu}$ is given by the means of its components, i.e. $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^T$ with $\mu_i = \mathrm{E}(x_i)$. Thus, the mean of a random vector of dimension d is a vector 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

[3] In our notational conventions, a **vector** $\boldsymbol{x}$ is written in *lower case in bold type*, a **matrix** $\boldsymbol{M}$ in *upper case in bold type*. Hence random vectors and matrices as well as their realisations are indicated in bold type, with vectors given in lower case and matrices in upper case. Hence, as for scalar variables, upper vs. lower case does not indicate randomness vs. realisation.

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.

2.10 Correlation matrix

The **correlation matrix** $\boldsymbol{P}$ ("upper case rho", not "upper case p") is the variance standardised version of the covariance matrix $\boldsymbol{\Sigma}$.

Specifically, denote by $\boldsymbol{V}$ the diagonal matrix containing the variances

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

then the correlation matrix $\boldsymbol{P}$ is given by

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

Like the covariance matrix the correlation matrix is symmetric, and note that the diagonal of $\boldsymbol{P}$ contains only 1s.

Equivalently, in component notation the correlation between x_i and x_j is given by

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

Using the above, a covariance matrix can be factorised into the product of standard deviations $\boldsymbol{V}^{\frac{1}{2}}$ and the correlation matrix as follows:

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

Chapter 3

Transforming and combining random variables

3.1 Affine or location-scale transformation of random variables

Suppose a and b are constants and $x \sim F_x$ is a scalar random variable with mean $\mathrm{E}(x) = \mu_x$ and variance $\mathrm{Var}(x) = \sigma_x^2$. The random variable $y = a + bx$ is a **location-scale transformation** or **affine transformation** of x, where a plays the role of the **location parameter** and b is the **scale parameter**. For $a = 0$ with no translation this is a **linear transformation**. If $b \neq 0$ then the transformation is **invertible**, with $x = (y - a)/b$. Invertible transformations provide a one-to-one map between x and y.

The random variable $y \sim F_y$ has mean

$$\mathrm{E}(y) = a + b\mu_x$$

and variance

$$\mathrm{Var}(y) = b^2\sigma_x^2$$

If x is a continuous random variable with density $f_x(x)$ and assuming an invertible transformation the density for y is given by

$$f_y(y) = |b|^{-1} f_x\left(\frac{y-a}{b}\right)$$

For a random vector $\boldsymbol{x}$ of dimension d and location parameter $\boldsymbol{a}$ (a $m \times 1$ vector) and scale parameter $\boldsymbol{B}$ (a $m \times d$ matrix) the location-scale transformation is $\boldsymbol{y} = \boldsymbol{a} + \boldsymbol{B}\boldsymbol{x}$. For $m = d$ (square $\boldsymbol{B}$) and $\det(\boldsymbol{B}) \neq 0$ the affine transformation is

invertible, with forward transformation $\boldsymbol{y} = \boldsymbol{a} + \boldsymbol{B}\boldsymbol{x}$ and backtransformation $\boldsymbol{x} = \boldsymbol{B}^{-1}(\boldsymbol{y} - \boldsymbol{a})$.

Suppose the mean and variance of the original random vector $\boldsymbol{x}$ is $\mathrm{E}(\boldsymbol{x}) = \boldsymbol{\mu}_{\boldsymbol{x}}$ and $\mathrm{Var}(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$. Then the mean and variance of the transformed random vector $\boldsymbol{y}$ is

$$\mathrm{E}(\boldsymbol{y}) = \boldsymbol{a} + \boldsymbol{B}\,\boldsymbol{\mu}_{\boldsymbol{x}}$$

and

$$\mathrm{Var}(\boldsymbol{y}) = \boldsymbol{B}\,\boldsymbol{\Sigma}_{\boldsymbol{x}}\,\boldsymbol{B}^{T}$$

Assuming an invertible transformation for a continous random vector $\boldsymbol{x}$ with density $f_{\boldsymbol{x}}(\boldsymbol{x})$ the density for $\boldsymbol{y}$ is given by

$$f_{\boldsymbol{y}}(\boldsymbol{y}) = |\det(\boldsymbol{B})|^{-1} f_{\boldsymbol{x}}\left(\boldsymbol{B}^{-1}(\boldsymbol{y} - \boldsymbol{a})\right)$$

The constants $\boldsymbol{a}$ and $\boldsymbol{B}$ (or a and b in the univariate case) are the parameters of the **location-scale family** F_y created from F_x. Many important distributions are location-scale families (e.g. the normal distribution and the location-scale t distribution). Furthermore, key procedures in multivariate statistics such as orthogonal transformations (including PCA) or whitening transformations (e.g. the Mahalanobis transformation) are affine transformations.

3.2 General invertible transformation of random variables

As a generalisation of invertible affine transformations we now consider general invertible transformations. For a scalar random variable we assume the transfomation is specified by $y = y(x) = h(x)$ and the backtransformation by $x = x(y) = h^{-1}(y)$ For a random vector we assume $\boldsymbol{y}(\boldsymbol{x}) = \boldsymbol{h}(\boldsymbol{x})$ is invertible with $\boldsymbol{x}(\boldsymbol{y}) = \boldsymbol{h}^{-1}(\boldsymbol{y})$.

The mean and variance of the transformed random variable can typically only be approximated. Linearising the transformation and computing the correponding means and variance is known as the **delta method**.

In the univariate case the delta method yields

$$\mathrm{E}(y) \approx y(\mu_x)$$

and

$$\mathrm{Var}(y) \approx (Dy(\mu_x))^2\,\sigma_x^2$$

where $Dy(x) = y'(x)$ is the first derivative of the transformation $y(x)$ and $Dy(\mu_x)$ is the first derivative evaluated at the mean μ_x

The density of the transformed variable can be computed exactly and is given by

$$f_y(y) = |Dx(y)|\, f_x(x(y))$$

Note in the above the use of the inverse transformation $x(y)$ and its derivative $Dx(y)$.

Assuming $y(x) = a + bx$, with $x(y) = (y - a)/b$, $Dy(x) = b$ and $Dx(y) = b^{-1}$, recovers the location-scale transformation.

For the vector random variable the delta method yields as approximation for the mean and variance

$$\mathrm{E}(\boldsymbol{y}) \approx \boldsymbol{y}(\boldsymbol{\mu}_{\boldsymbol{x}})$$

and

$$\mathrm{Var}(\boldsymbol{y}) \approx D\boldsymbol{y}(\boldsymbol{\mu}_{\boldsymbol{x}})\, \boldsymbol{\Sigma}_{\boldsymbol{x}}\, D\boldsymbol{y}(\boldsymbol{\mu}_{\boldsymbol{x}})^T$$

where $D\boldsymbol{y}(\boldsymbol{x})$ is the Jacobian matrix (vector derivative) for the transformation $\boldsymbol{y}(\boldsymbol{x})$ and $D\boldsymbol{y}(\boldsymbol{\mu}_{\boldsymbol{x}})$ is the Jacobian matrix evaluated at the mean $\boldsymbol{\mu}_{\boldsymbol{x}}$.

The density for $\boldsymbol{y}$ is obtained by

$$f_{\boldsymbol{y}}(\boldsymbol{y}) = |\det(D\boldsymbol{x}(\boldsymbol{y}))|\;\; f_{\boldsymbol{x}}(\boldsymbol{x}(\boldsymbol{y}))$$

where $D\boldsymbol{x}(\boldsymbol{y})$ is the Jacobian matrix of the inverse transformation $\boldsymbol{x}(\boldsymbol{y})$.

Assuming $\boldsymbol{y}(\boldsymbol{x}) = \boldsymbol{a} + \boldsymbol{B}\boldsymbol{x}$, with $\boldsymbol{x}(\boldsymbol{y}) = \boldsymbol{B}^{-1}(\boldsymbol{y} - \boldsymbol{a})$, $D\boldsymbol{y}(\boldsymbol{x}) = \boldsymbol{B}$ and $D\boldsymbol{x}(\boldsymbol{y}) = \boldsymbol{B}^{-1}$, recovers the location-scale transformation.

3.3 Exponential tilting and exponential families

Another way to change the distribution of a random variable is by **exponential tilting**.

Suppose there is a vector valued function $\boldsymbol{u}(x)$ where each component is a transformation of x, usually a simple function such the identity x, the square x^2, the logarithm $\log(x)$ etc. These are called the **canonical statistics**. Typically, the dimension of $\boldsymbol{u}(x)$ is small.

The exponential tilt of a **base distribution** B with density or probability mass function $b(x)$ towards the linear combination $\boldsymbol{\eta}^T\boldsymbol{u}(x)$ of the canonical statistics $\boldsymbol{u}(x)$ yields the distribution family $P_{\boldsymbol{\eta}}$ with density or probability mass function

$$p(x|\boldsymbol{\eta}) = \underbrace{e^{\boldsymbol{\eta}^T\boldsymbol{u}(x)}}_{\text{exponential tilt}}\; b(x) \,/\, e^{\psi(\boldsymbol{\eta})}$$

where $\boldsymbol{\eta}$ are the canonical parameters. The normalising factor $e^{\psi(\boldsymbol{\eta})}$ ensures that $p(x|\boldsymbol{\eta})$ integrates to one following the exponential tilt.

The corresponding log-density / log probability mass function is

$$\log p(x|\boldsymbol{\eta}) = \boldsymbol{\eta}^T \boldsymbol{u}(x) + \log b(x) - \psi(\boldsymbol{\eta})$$

The **log-normaliser** or **log-partition function** $\psi(\boldsymbol{\eta})$ is obtained by computing

$$\psi(\boldsymbol{\eta}) = \log \int_x e^{\boldsymbol{\eta}^T \boldsymbol{u}(x)} \, b(x) \, dx$$

The set of values of $\boldsymbol{\eta}$ for which the integral is finite and hence $\psi(\boldsymbol{\eta}) < \infty$ defines the parameter space.

The distribution family $P_{\boldsymbol{\eta}}$ obtained by exponential tiling is called an **exponential family**.

Many commonly used distribution families are exponential families (most importantly the normal distribution). Exponential families are extremely important in probability and statistics. They provide highly effective models for statistical learning using entropy, likelihood and Bayesian approaches, allow for substantial data reduction via minimal sufficiency, and provide the basis of generalised linear models. Furthermore, exponential families often enable to generalise probabilistic results valid for the normal distribution to more general settings.

3.4 Sums of iid random variables and convolution

Suppose we have a sum of n independent and identically distributed (iid) random variables.

$$y = x_1 + x_2 + \ldots + x_n$$

where each $x_i \sim F_x$ with density or probability mass function $f_x(x)$. The density or probability mass function for y is obtained by repeated application of **convolution** (symbolised by the $*$ operator):

$$f_y(y) = (f_{x_1} * f_{x_2} * \ldots f_{x_n})(y)$$

The convolution of two functions is defined as (continous case)

$$(f_{x_1} * f_{x_2})(y) = \int_x f_{x_1}(x) \, f_{x_2}(y - x) dx$$

and (discrete case)

$$(f_{x_1} * f_{x_2})(y) = \sum_x f_{x_1}(x) \, f_{x_2}(y - x)$$

Convolution is commutative and associative so it can be applied in any order to compute the convolution of multiple functions. Furthermore, the convolution of probability densities / mass function yields another probability density / mass function.

Many commonly used random variables can be viewed as the outcome of convolutions. For example, the sum of Bernoulli variables yields a binomial random variable and the sum of normal variables yields another normal random variable. See also: list of convolutions of probability distributions.

The **central limit theorem**, first postulated by Abraham de Moivre (1667–1754), asserts that, under appropriate conditions, the distribution of the sum of independent and identically distributed random variables converges in the limit of large n to a normal distribution, even if the individual random variables are not normal. In other words, it asserts that for large n the convolution of n identical distributions typically converges to the normal distribution.

Chapter 4

Univariate distributions

4.1 Bernoulli distribution

The **Bernoulli distribution** $\text{Ber}(\theta)$ is the simplest of all distribution families. 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 $\text{Ber}(\theta)$ is

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

and

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

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 $\text{E}(x) = \theta$ and the variance is $\text{Var}(x) = \theta(1 - \theta)$.

4.2 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 $\text{Bin}(n, \theta)$.

If a random variable y follows the binomial distribution we write

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

The corresponding probability mass function is:

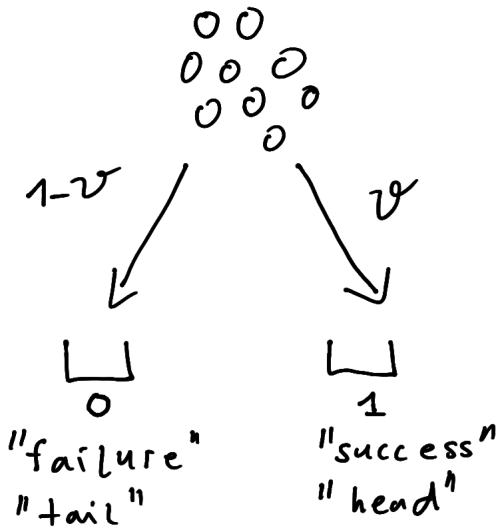
$$p(y|n, \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

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

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

If we standardise the support of the binomial variable to the unit interval with $\frac{y}{n} \in \left\{0, \frac{1}{n}, \ldots, 1\right\}$ then the mean is $\text{E}\left(\frac{y}{n}\right) = \theta$ and the variance is $\text{Var}\left(\frac{y}{n}\right) = \frac{\theta(1-\theta)}{n}$.

The binomial distribution may be illustrated by an urn model distributing n balls into 2 bins:



For $n = 1$ the binomial distribution reduces to the Bernoulli distribution.

In R the PMF of the binomial distribution is given by `dbinom()`, the cumulative distribution function is `pbinom()` and the quantile function is `qbinom()`. The binomial coefficient itself is computed by `choose()`.

As a result of the central limit theorem, the binomial distribution, obtained as the convolution of n Bernoulli distributions, can for large n be well approximated by a normal distribution (this is known as the De Moivre–Laplace theorem).

4.3 Beta distribution

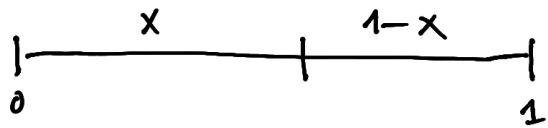
4.3.1 Standard parameterisation

A beta-distributed random variable is denoted by

$$x \sim \text{Beta}(\alpha, \beta)$$

where the support is $x \in [0, 1]$ and $\alpha > 0$ and $\beta > 0$ are two shape parameters.

The beta random variable can be visualised as breaking a unit stick of length one into two pieces of length x and $1 - x$:



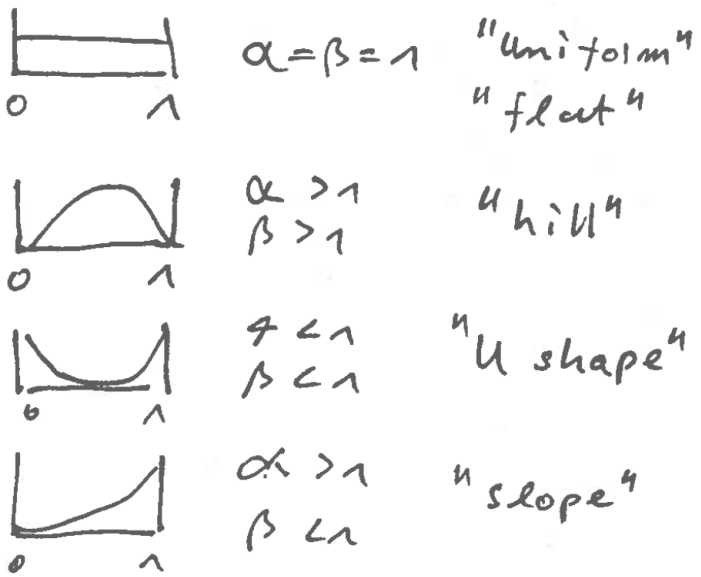
The density of the beta distribution $\text{Beta}(\alpha, \beta)$ is

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

This depends on the beta function defined as

$$B(z_1, z_1) = \frac{\Gamma(z_1)\Gamma(z_2)}{\Gamma(z_1 + z_2)}$$

The beta distribution is very flexible and can assume a number of different shapes, depending on the value of α and β. For example, for $\alpha = \beta = 1$ it becomes the uniform distribution over the unit interval:



In R the PDF of the beta distribution is given by `dbeta()`, the cumulative distribution function is `pbeta()` and the quantile function is `qbeta()`.

4.3.2 Mean parameterisation

Instead of employing α and β as parameters another useful reparameterisation $\text{Beta}(\mu, k)$ 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

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

and

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

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

The uniform distribution (with $\alpha = \beta = 1$) corresponds to $\mu = 1/2$ and $k = 2$.

Finally, note that the mean and variance of the continuous beta distribution closely match those of the unit-standardised discrete binomial distribution above.

4.4 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) and support $x \in]-\infty, \infty[$.

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

with mean

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

and variance

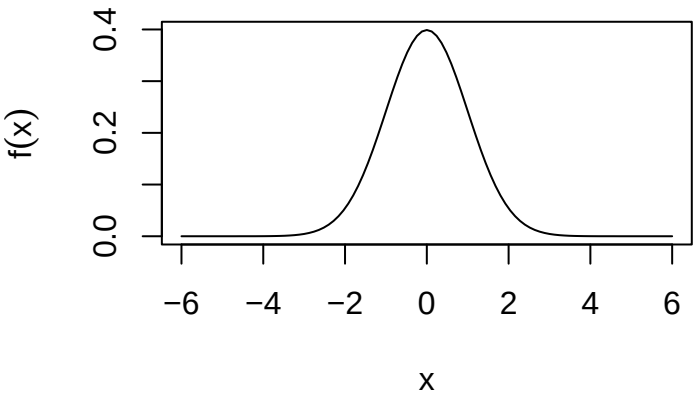
$$\mathrm{Var}(x) = \sigma^2$$

Probability density function (PDF):

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

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

Plot of the PDF of the standard normal:

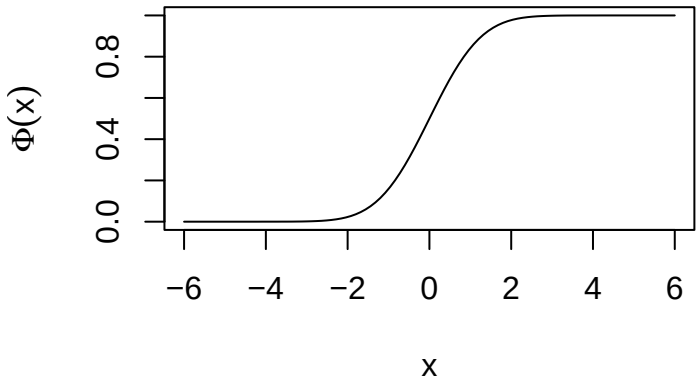


The cumulative distribution function (CDF) of the standard normal $N(0, 1)$ is

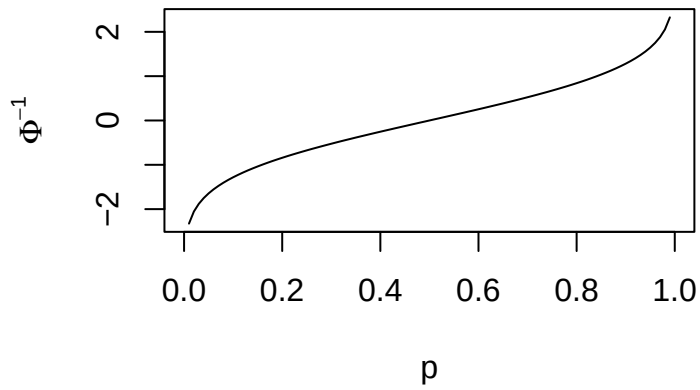
$$\Phi(x) = \int_{-\infty}^{x} p(x'|\mu = 0, \sigma^2 = 1)dx'$$

There is no analytic expression for $\Phi(x)$.

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 normal PDF is given by `dnorm()`, the cumulative distribution function is `pnorm()` and the quantile function is `qnorm()`.

4.5 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

4.5.1 Standard parameterisation

The gamma distribution $\text{Gam}(\alpha, \theta)$ is a continuous distribution with two parameters $\alpha > 0$ (shape) and $\theta > 0$ (scale):

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

and support $x \in [0, \infty[$ 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).

The probability density function (PDF) is:

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

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()`.

4.5.2 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**

$$\text{W}_1\left(s^2, k\right)$$

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

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

and the variance

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

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

$$s^2\chi^2_k$$

Finally, we also often employ the Wishart distribution in **mean parameterisation**

$$\mathrm{W}_1\left(s^2 = \frac{\mu}{k}, k\right)$$

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

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

4.5.3 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

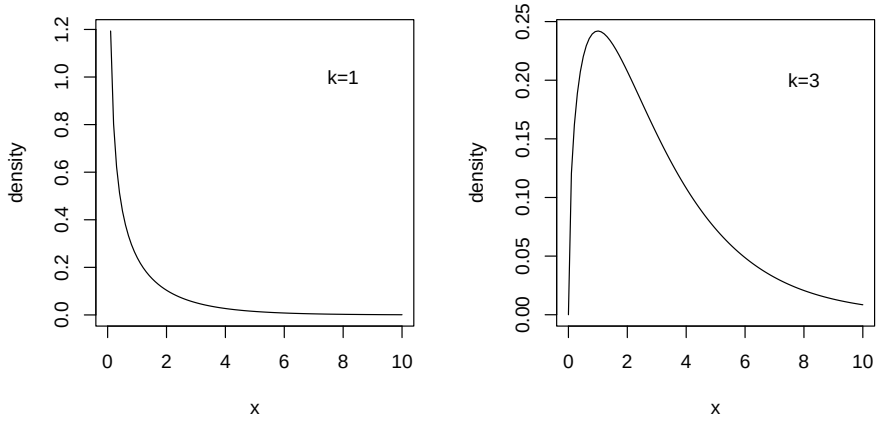
$$\begin{aligned} x \sim\ & s^2\chi^2_k \\ &= \mathrm{W}_1\left(s^2, k\right) \\ &= \mathrm{Gam}\left(\alpha = \frac{k}{2}, \theta = 2s^2\right) \end{aligned}$$

4.5.4 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 $\mathrm{Gam}(\alpha = k/2, \theta = 2) = \mathrm{W}_1(1, k)$.

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()`.

4.5.5 Exponential distribution

The **exponential distribution** $\text{Exp}(\theta)$ 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 $\text{Gam}(\alpha = 1, \theta) = \text{W}_1(s^2 = \theta/2, k = 2)$. 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.

4.6 Inverse gamma distribution

Also know as inverse univariate Wishart distribution.

4.6.1 Standard parameterisation

A random variable x following an **inverse gamma distribution** is denoted by

$$x \sim \text{Inv-Gam}(\alpha, \beta)$$

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

The inverse of x is then gamma distributed

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

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

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

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

The mean of the inverse gamma distribution is

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

and the variance

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

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

In the R `extraDistr` package the density of the inverse gamma distribution is given by `extraDistr::dinvgamma()`, the function `extraDistr::pinvgamma()` returns the corresponding cumulative density and `extraDistr::qinvgamma()` is the quantile function.

4.6.2 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**

$$\text{W}_1^{-1}(\psi, \nu)$$

with mean given by

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

for $\nu > 2$ and variance

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

for $\nu > 4$.

The inverse univariate Wishart and univariate Wishart distributions are linked. If a random variable x is inverse Wishart distributed

$$x \sim \mathrm{W}_1^{-1}(\psi, \nu)$$

then the inverse of x is Wishart distributed with inverted scale parameter:

$$\frac{1}{x} \sim \mathrm{W}_1(s^2 = \psi^{-1}, k = \nu)$$

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

Instead of ψ and ν we may also equivalently use $\kappa = \nu - 2$ and $\mu = \psi/(\nu - 2)$ as parameters for the inverse Wishart distribution, so that

$$\mathrm{W}_1^{-1}(\psi = \kappa\mu, \nu = \kappa + 2)$$

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 inverse gamma distribution as prior and posterior.

Finally, with $\mathrm{W}_1^{-1}(\psi = \nu\tau^2, \nu)$, 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 $\nu > 4$.

4.7 Location-scale t-distribution and special cases

4.7.1 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 $\text{lst}(\mu, \tau^2, \nu)$ is

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

with support $x \in]-\infty, \infty[$. The mean is (for $\nu > 1$)

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

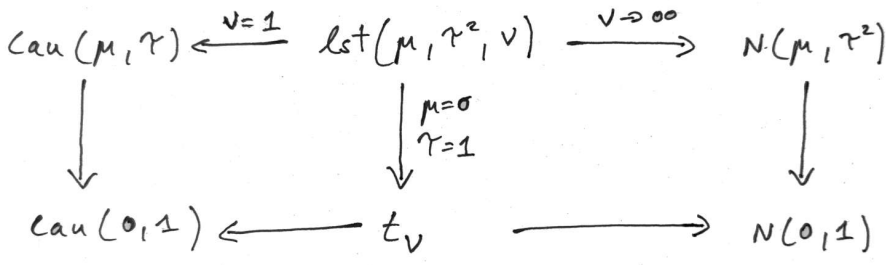
and the variance (for $\nu > 2$)

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

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

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

The following figure illustrates the relationship of the location-scale t distribution $\text{lst}(\mu, \tau^2, \nu)$ with related distributions such as the normal distribution $N(\mu, \tau^2)$, Student's t-distribution t_ν and the Cauchy distribution $\text{Cau}(\mu, \tau)$ discussed further below.



4.7.2 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 $\text{E}(x|s^2) = \mu$ and variance $\text{Var}(x|s^2) = s^2$.

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

$$s^2 \sim \text{W}_1^{-1}(\psi = \kappa\sigma^2, \nu = \kappa + 2) = \text{W}_1^{-1}(\psi = \tau^2\nu, \nu)$$

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

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

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

The probability density of t_ν is

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

with support $x \in]-\infty, \infty[$.

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.

4.7.4 Cauchy and standard Cauchy distribution

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

Chapter 5

Multivariate distributions

5.1 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.

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

The form of the categorical covariance matrix follows directly from the definition of the variance $\mathrm{Var}(\boldsymbol{x}) = \mathrm{E}(\boldsymbol{x}\boldsymbol{x}^T) - \mathrm{E}(\boldsymbol{x})\mathrm{E}(\boldsymbol{x})^T$ 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 $\sum_{k=1}^{K} x_k = 1$.

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

5.2 Multinomial distribution

The **multinomial distribution** $\mathrm{Mult}(n, \boldsymbol{\pi})$ 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 $\mathrm{Cat}(\boldsymbol{\pi})$ random categorical variables then $\boldsymbol{y} = \sum_{i=1}^{n} \boldsymbol{x}_i$ is distributed as $\mathrm{Mult}(n, \boldsymbol{\pi})$.

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

$$\mathrm{E}(\boldsymbol{y}) = n\boldsymbol{\pi}$$

which in component notation is $\mathrm{E}(y_k) = n\pi_k$. The covariance matrix is

$$\mathrm{Var}(\boldsymbol{y}) = n(\mathrm{Diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T)$$

which in component notation is $\mathrm{Var}(x_i) = n\pi_i(1-\pi_i)$ and $\mathrm{Cov}(x_i, x_j) = -n\pi_i\pi_j$.

Standardised to unit interval we get:

$$\frac{y_i}{n} \in \left\{0, \frac{1}{n}, \frac{2}{n}, ..., 1\right\}$$

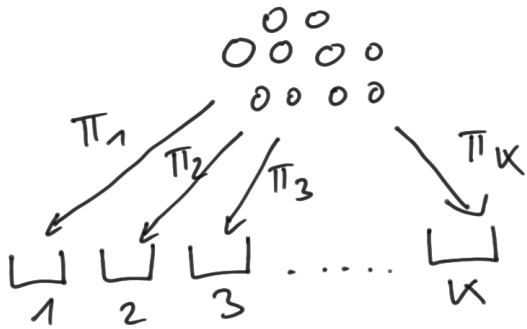
$$\mathrm{E}\left(\frac{\boldsymbol{y}}{n}\right) = \boldsymbol{\pi}$$

$$\mathrm{Var}\left(\frac{\boldsymbol{y}}{n}\right) = \frac{\mathrm{Diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T}{n}$$

$$\mathrm{Var}\left(\frac{y_i}{n}\right) = \frac{\pi_i(1-\pi_i)}{n}$$

$$\mathrm{Cov}\left(\frac{y_i}{n}, \frac{y_j}{n}\right) = -\frac{\pi_i\pi_j}{n}$$

The multinomial distribution may be illustrated by an urn model distributing n balls into K bins:



For $n = 1$ the multinomial distribution reduces to the categorical distribution.

For $K = 2$ the multinomial distribution reduces to the Binomial distribution.

5.3 Dirichlet distribution

5.3.1 Standard parameterisation

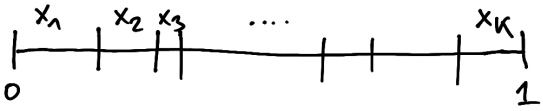
The Dirichlet distribution is the multivariate generalisation of the beta distribution.

A Dirichlet distributed random vector is denoted by

$$\boldsymbol{x} \sim \text{Dir}(\boldsymbol{\alpha})$$

with parameter $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_K)^T > 0$ and $K \geq 2$ and where the support of $\boldsymbol{x}$ is the $K-1$ dimensional simplex with $x_i \in [0, 1]$ and $\sum_{i=1}^{K} x_i = 1$.

The Dirichlet random variable can be visualised as breaking a unit stick of length one in K pieces of length x_1 to x_K:



The density of the Dirichlet distribution $\text{Dir}(\boldsymbol{\alpha})$ is

$$p(\boldsymbol{x}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} x_k^{\alpha_k - 1}$$

This depends on the beta function with multivariate argument defined as

$$B(\boldsymbol{z}) = \frac{\prod_{k=1}^{K} \Gamma(z_k)}{\Gamma\left(\sum_{k=1}^{K} z_k\right)}$$

For $K = 2$ the Dirichlet distribution reduces to the beta distribution.

5.3.2 Mean parameterisation

Instead of employing $\boldsymbol{\alpha}$ as parameter vector another useful reparameterisation $\text{Dir}(\boldsymbol{\pi}, k)$ of the Dirichlet distribution is in terms of a mean parameter $\boldsymbol{\pi}$, with $\pi_i \in [0, 1]$ and $\sum_{i=1}^{K} \pi_i = 1$, and a concentration parameter $k > 0$. These are given by

$$k = \sum_{i=1}^{K} \alpha_i$$

and

$$\boldsymbol{\pi} = \frac{\boldsymbol{\alpha}}{k}$$

The original parameters can be recovered by $\alpha = \boldsymbol{\pi} k$.

The mean and variance of the Dirichlet distribution expressed in terms of $\boldsymbol{\pi}$ and k are

$$\text{E}(\boldsymbol{x}) = \boldsymbol{\pi}$$

and

$$\text{Var}(\boldsymbol{x}) = \frac{\text{Diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T}{k + 1}$$

which in component notation is

$$\text{Var}(x_i) = \frac{\pi_i(1-\pi_i)}{k+1}$$

and

$$\text{Cov}(x_i, x_j) = -\frac{\pi_i \pi_j}{k+1}$$

Finally, note that the mean and variance of the continuous Dirichlet distribution closely match those of the unit-standardised discrete multinomial distribution above.

5.4 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 occurence of the dimension d.

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

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.

5.5 Wishart distribution

The Wishart distribution is a multivariate generalisation of the gamma distribution.

Recall that the gamma distribution can be motivated as the distribution of sums of squared normal random variables. Likewise, the Wishart distribution can be understood as the sum of squared multivariate normal variables:

$$z_1, z_2, \ldots, z_k \overset{\text{iid}}{\sim} N_d(0, \boldsymbol{S})$$

with $\boldsymbol{S} = (s_{ij})$ the specified covariance matrix. The random variable

$$\underbrace{\boldsymbol{X}}_{d\times d} = \sum_{i=1}^{k} \underbrace{\boldsymbol{z}_i \boldsymbol{z}_i^T}_{d\times d}$$

is a random *matrix* and is distributed as

$$\boldsymbol{X} \sim \mathrm{W}_d(\boldsymbol{S}, k)$$

with mean

$$\mathrm{E}(\boldsymbol{X}) = k\boldsymbol{S}$$

and variances

$$\mathrm{Var}(x_{ij}) = k\left(s_{ij}^2 + s_{ii}s_{jj}\right)$$

We often also employ the Wishart distribution in **mean parameterisation**

$$\mathrm{W}_d\left(\boldsymbol{S} = \frac{\boldsymbol{M}}{k}, k\right)$$

with parameters $\boldsymbol{M} = k\boldsymbol{S}$ and k. In this parameterisation the mean is

$$\mathrm{E}(\boldsymbol{X}) = \boldsymbol{M} = (\mu_{ij})$$

and variances are

$$\mathrm{Var}(x_{ij}) = \frac{\mu_{ij}^2 + \mu_{ii}\mu_{jj}}{k}$$

If $\boldsymbol{S}$ or $\boldsymbol{M}$ is a scalar rather than a matrix then the multivariate Wishart distribution reduces to the univariate Wishart aka gamma distribution.

5.6 Inverse Wishart distribution

The inverse Wishart distribution is a multivariate generalisation of the inverse gamma distribution and is linked to the Wishart distribution.

A random matrix $\boldsymbol{X}$ following an **inverse Wishart distribution** is denoted by

$$\boldsymbol{X} \sim \mathrm{W}_d^{-1}(\boldsymbol{\Psi}, \nu)$$

where $\boldsymbol{\Psi}$ is the scale parameter and ν the shape parameter. The corresponding mean is given by

$$\mathrm{E}(\boldsymbol{X}) = \frac{\boldsymbol{\Psi}}{\nu - d - 1}$$

and the variances are

$$\mathrm{Var}(x_{ij}) = \frac{(\nu - d + 1)\psi_{ij}^2 + (\nu - d - 1)\,\psi_{ii}\psi_{jj}}{(\nu - d)(\nu - d - 1)^2(\nu - d - 3)}$$

The inverse of $\boldsymbol{X}$ is then Wishart distributed:

$$\boldsymbol{X}^{-1} \sim \mathrm{W}_d\left(\boldsymbol{S} = \boldsymbol{\Psi}^{-1}, k = \nu\right)$$

Instead of $\boldsymbol{\Psi}$ and ν we may use the mean parameterisation with parameters $\kappa = \nu - d - 1$ and $\boldsymbol{M} = \boldsymbol{\Psi}/(\nu - d - 1)$:

$$\boldsymbol{X} \sim \mathrm{W}_d^{-1}(\boldsymbol{\Psi} = \kappa \boldsymbol{M}, \nu = \kappa + d + 1)$$

with mean

$$\mathrm{E}(\boldsymbol{X}) = \boldsymbol{M}$$

and variances

$$\mathrm{Var}(x_{ij}) = \frac{(\kappa + 2)\mu_{ij}^2 + \kappa\,\mu_{ii}\mu_{jj}}{(\kappa + 1)(\kappa - 2)}$$

If $\boldsymbol{\Psi}$ or $\boldsymbol{M}$ is a scalar rather than a matrix then the multivariate inverse Wishart distribution reduces to the univariate inverse Wishart aka inverse gamma distribution.