Probability and Distribution Refresher

Korbinian Strimmer

10 October 2025

Table of contents

Welcome

The Probability and Distribution Refresher notes were written by Korbinian Strimmer from 2018–2025. This version is from 10 October 2025.

If you have any questions, comments, or corrections please get in touch! [1]

Updates

The lecture notes will be updated from time to time.

The most current version is found at the web page for the

- online version of the Probability and Distribution Refresher notes.

There you can also 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

[1] Email address: korbinian.strimmer@manchester.ac.uk

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 currently offered modules:

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

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

- MATH20802 Statistical Methods.

1 Combinatorics

1.1 Some basic mathematical notation

Scalar quantity: plain font, typically lower case (x, θ, n), sometimes upper case (K, R^2, distribution functions F, P, Q).

Sets: plain font, upper case ($\Omega, \mathcal{F}$)

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

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

Summation:

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

Product:

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

The multiplication sign $\times$ between the factors is usually omitted unless it is needed for clarity.

Indicator function (in Iverson bracket notation):

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

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 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

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

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 coefficient, 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-Gibbs entropy of a categorical distribution with n_k/n as class probabilities.

2 Probability

2.1 Random variables

A **random variable** describes a random experiment. The set of all possible outcomes is the **sample space** of the random variable and is denoted by Ω. If Ω is countable then the random variable is **discrete**, otherwise it is **continuous**. For a discrete random variable the sample space $\Omega = \{\omega_1, \omega_2, \ldots\}$ is composed of a finite or infinite number of **elementary outcomes** ω_i.

An event $A \subseteq \Omega$ is a subset of Ω. This includes as special cases the complete set Ω ("certain event") and the empty set $\emptyset$ ("impossible event"). The set of all possible events is denoted by $\mathcal{F}$. The complementary event $A^C = \Omega \setminus A$ is the complement of the set A in the sample space Ω. Two events A_1 and A_2 are mutually exclusive if the sets are disjoint with $A_1 \cap A_2 = \emptyset$.

For a discrete random variable, the elementary outcomes ω_i are referred to as **elementary events**, and they are all mutually exclusive. An event A consists of a number of elementary events $\omega_i \in A$ and the complementary event is given by $A^C = \{\omega_i \in \Omega : \omega_i \notin A\}$.

The **probability of an event** A is denoted by $\Pr(A)$. Broadly, $\Pr(A)$ provides a measure of the size of the set A relative to the set Ω. The probability measure $\Pr(A)$ satisfies the three axioms of probability:

1) $\Pr(A) \geq 0$, probabilities are non-negative,
2) $\Pr(\Omega) = 1$, the certain event has probability 1, and
3) $\Pr(A_1 \cup A_2 \cup \ldots) = \sum_i \Pr(A_i)$, the probability of countable mutually exclusive events A_i is additive.

This implies

- $\Pr(A) \leq 1$, probability values lie within the range $[0, 1]$,
- $\Pr(A^C) = 1 - \Pr(A)$, the probability of the complement, and
- $\Pr(\emptyset) = 0$, the impossible event has probability 0.

From the above it is evident that probability is closely linked to set theory, in particular to measure theory which serves as the theoretical foundations of probability and generalisations. For instance, if $\Pr(\emptyset) = 0$ is assumed instead of $\Pr(\Omega) = 1$, this leads to the axioms for a **positive measure** (of which probability is a special case).

2.2 Conditional probability

Consider two events A and B, which may not be be mutually exclusive. The probability of the event "A and B" is given by the probability of the set intersection $\Pr(A \cap B)$. The probability of the event "A or B" is given by the probability of the set union

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)\,.$$

This identity follows from the axioms.

The **conditional probability** of event A assuming event B has occurred is given by

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

Essentially, now B acts as the new sample space relative to which A is measured, restricting it from Ω. Note that $\Pr(A|B)$ is generally not the same as $\Pr(B|A)$, see Bayes' theorem below.

Importantly, it can be seen that any probability may be viewed as conditional, namely relative to Ω as $\Pr(A) = \Pr(A|\Omega)$.

From the definition of conditional probability we derive the **product rule**

$$\begin{aligned}\Pr(A \cap B) &= \Pr(A|B)\Pr(B) \\ &= \Pr(B|A)\Pr(A)\end{aligned}$$

which in turn yields **Bayes' theorem**

$$\Pr(A|B) = \Pr(B|A)\frac{\Pr(A)}{\Pr(B)}$$

This theorem is useful for changing the order of conditioning and it plays a key role in Bayesian statistics.

If $\Pr(A \cap B) = \Pr(A)\Pr(B)$ then the two events A and B are **independent** with $\Pr(A|B) = \Pr(A)$ and $\Pr(B|A) = \Pr(B)$.

2.3 Probability mass and density function

The **distribution** (or **law**) of a random variable x with sample space Ω gives the probability for each value or a range of values of x according to the underlying probability measure. This is done in practise by employing probability mass functions (for discrete random variables) or probability density functions (for continuous random variables).

Here x is a scalar random variable, denoted by lower case and plain font. We also write x for the outcomes of the random variable. Thus, we use the same symbol to denote a random variable and its realisations.[1]

For a discrete random variable we define the event $A = \{x : x = a\} = \{a\}$ (corresponding to a single elementary event) and get the probability

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

directly from the **probability mass function** (**pmf**). The pmf has the property that $\sum_{x\in\Omega} f(x) = 1$ and that $f(x) \in [0, 1]$.

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

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

Similarly, the probability of the event $A = \{x : a_1 < x \leq a_2\}$ is given by

$$\Pr(A) = \Pr(a_1 < x \leq a_2) = \int_{a_1}^{a_2} f(a)da\,.$$

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.

It is sometimes convenient to refer to a pdf or a pmf collectively **as probability density mass function** (**pdmf**) without specifying whether x is continuous or discrete.

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

[1] This notation is common in statistical machine learning and multivariate statistics, see for example Mardia, Kent, and Bibby (1979). In alternative notation, random variables are often written in upper case and the outcomes in lower case. However, that convention doesn't work well for multivariate random quantities (i.e. random vectors and random matrices) and is also ill-suited in Bayesian statistics where the uncertainty of parameters are modelled by random variables.

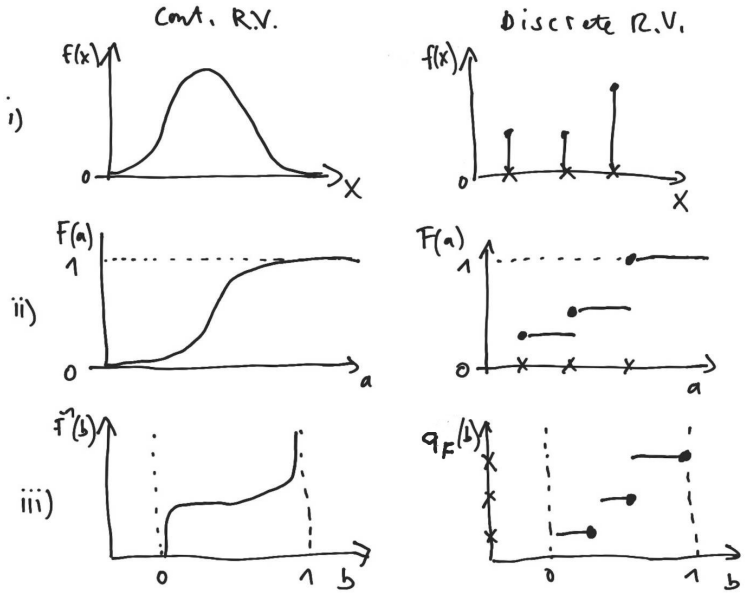

Figure 2.1: Illustration of i) pdmf, ii) distribution function and iii) quantile function for a continuous (first column) and a discrete random variable (second column).

Using the pdmf, the probability of general event $A \subseteq \Omega$ is given by

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

Figure 2.1 (first row) illustrates the pdmf for a continuous and discrete random variable.

In the above we denoted the pdmf by the lower case letter f though we also often use p or q.

2.4 Cumulative distribution function

As alternative to the pdmf we can describe the random variable using a **cumulative distribution function** (**cdf**). This requires an ordering so

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

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

Th cdf is denoted by the same letter as the pdmf but in upper case (usually F, P and Q). By construction the cumulative distribution function is monotonically non-decreasing and its value ranges from 0 to 1. For a discrete random variable $F(a)$ is a step function with jumps of size $f(\omega_i)$ at the elementary outcomes ω_i.

With the help of the cdf we can compute the probability of the event $A = \{x : a_1 < x \leq a_2\}$ simply as

$$\Pr(A) = F(a_2) - F(a_1)\,.$$

This works both for discrete and continuous random variables.

Figure 2.1 (second row) illustrates the distribution function for a continuous and discrete random variable.

It is common to use the same upper case letter as the cdf to name the distribution. Thus, if a random variable x has distribution F we write $x \sim F$, and this implies it has a pdmf $f(x)$ and cdf $F(x)$.

2.5 Quantile function and quantiles

The **quantile function** is defined as $q_F(b) = \min\{x : F(x) \geq b\}$. For a continuous random variable the quantile function simplifies to $q_F(b) = F^{-1}(b)$, i.e. it is the ordinary inverse $F^{-1}(b)$ of the distribution function.

Figure 2.1 (third row) illustrates the quantile function for a continuous and discrete random variable.

The quantile x of order b of the distribution F is often denoted by $x_b = q_F(b)$.

The 25% quantile $x_{1/4} = x_{25\%} = q_F(1/4)$ is called the **first quartile** or **lower quartile**.

The 50% quantile $x_{1/2} = x_{50\%} = q_F(1/2)$ is called the **second quartile** or **median**.

The 75% quantile $x_{3/4} = x_{75\%} = q_F(3/4)$ is called the **third quartile** or **upper quartile**.

The interquartile range is the difference between the upper and lower quartiles and equals $\text{IQR}(F) = q_F(3/4) - q_F(1/4)$.

The quantile function is also useful for generating general random variates from uniform random variates. If $y \sim \text{Unif}(0, 1)$ then $x = q_F(y) \sim F$.

2.6 Expectation or mean

The expected value of a random variable $x \sim F$ is defined as the weighted average over all possible outcomes, with the weight given by the pdmf $f(x)$:

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

We may can also write $\text{E}_F(x)$ as a reminder that the expectation is taken with regard to the distribution F. Usually, the subscript F is left out if there are no ambiguities. A further variant is to write the expectation as $\text{E}(F)$ to indicate that the mean is a functional of the distribution F.

Because the sum or integral may diverge, not all distributions have finite means so the mean does not always exist (in contrast to the median, or quantiles in general). For example, the location-scale t-distribution $t_\nu(\mu, \tau^2)$ does not have a mean for a degree of freedom in the range $0 < \nu \leq 1$ (see Section 4.6).

2.7 Variance

The variance of a random variable $x \sim F$ is the expected value of the squared deviation around the mean $\mu = \text{E}(x)$:

$$\begin{aligned} \text{Var}(F) &= \text{Var}(x) \\ &= \text{E}\left((x - \mu))^2\right) \\ &= \text{E}(x^2) - \mu^2 \end{aligned}$$

By construction, $\text{Var}(x) \geq 0$.

The notation $\mathrm{Var}(F)$ highlights that the variance is a functional of the distribution F. Occasionally, we write $\mathrm{Var}_F(x)$ indicate that the expectation is taken with regard to the distribution F.

Like the mean, the variance may diverge and hence not necessarily exists for all distribution. For example, the location-scale t-distribution $t_\nu(\mu, \tau^2)$ does not have a variance for the degree of freedom in the range $0 < \nu \leq 2$ (see Section 4.6).

2.8 Moments of a distribution

The n-th moment of a distribution F for a random variable x is defined as follows:

$$\mu_n(F) = \mathrm{E}(x^n)$$

Special important cases are the

- Zeroth moment: $\mu_0(F) = \mathrm{E}(x^0) = 1$ (since the pdmf integrates to one)
- First moment: $\mu_1(F) = \mathrm{E}(x^1) = \mathrm{E}(x) = \mu$ (=the mean)
- Second moment: $\mu_2(F) = \mathrm{E}(x^2)$

The n-th central moment centred around the mean $\mathrm{E}(x) = \mu$ is given by

$$m_n(F) = \mathrm{E}((x - \mu)^n)$$

The first few central moments are the

- Zeroth central moment: $m_0(F) = \mathrm{E}((x - \mu)^0) = 1$
- First central moment: $m_1(F) = \mathrm{E}((x - \mu)^1) = 0$
- Second central moment: $m_2(F) = \mathrm{E}\left((x - \mu)^2\right)$ (=the variance)

The moments of a distribution are not necessarily all finite, i.e. some moments may not exist. For example, the location-scale t-distribution $t_\nu(\mu, \tau^2)$ only has finite moments of degree smaller than the degree of freedom ν (see Section 4.6).

2.9 Expectation of a transformed random variable

Often, one needs to find the mean of a transformed random variable. If $x \sim F_x$ and $y = h(x)$ with $y \sim F_y$ then one can directly apply the above definition to obtain $\mathrm{E}(y) = \mathrm{E}(F_y)$. However, this requires knowledge of the transformed pdmf $f_y(y)$ (see Chapter 3 for more details about variable transformations).

As an alternative, the "law of the unconscious statistician"(LOTUS) provides a convenient shortcut to compute the mean of the transformed random variable $y = h(x)$ using only the pdmf of the original variable x:

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

Note this is not an approximation but equivalent to obtaining the mean using the transformed pdmf.

2.10 Probability as expectation

Probability itself can also be understood as an expectation.

For an event $A \subseteq \Omega$ we define a corresponding indicator function $[x \in A]$. From LOTUS it then follows immediately that

$$\begin{aligned} \mathrm{E}([x \in A]) &= \begin{cases} \sum_{x\in A} f(x) & \text{discrete case} \\ \int_{x\in A} f(x)\, dx & \text{continuous case} \end{cases} \\ &= \Pr(A) \end{aligned}$$

This relation is called the "fundamental bridge" between probability and expectation. Interestingly, one can develop the whole theory of probability from this perspective (e.g., Whittle 2000).

2.11 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".

An example of Jensen's inequality is $\mathrm{E}(x^2) \geq \mathrm{E}(x)^2$.

2.12 Random vectors and their mean and variance

In addition to scalar random variables we often make use of random vectors and random matrices.[2]

The mean of a random vector $\boldsymbol{x} = (x_1, x_2, ..., x_d)^T \sim F$ is given by

$$\begin{aligned}\mathrm{E}(F) &= \mathrm{E}(\boldsymbol{x}) \\ &= \underbrace{\boldsymbol{\mu}}_{d\times 1} = (\mu_1, \ldots, \mu_d)^T\end{aligned}$$

and thus is a vector of the same dimension as $\boldsymbol{x}$, where $\mu_i = \mathrm{E}(x_i)$ are the means of the individual components x_i.

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

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

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

[2] In our notational conventions, a scalar x is written in lower case plain font, a vector $\boldsymbol{x}$ is written in lower case bold font, a matrix $\boldsymbol{X}$ in upper case bold font.

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.

2.13 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

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

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

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

Like the covariance matrix the correlation matrix is symmetric. The elements of the diagonal of $\boldsymbol{P}$ are all set to 1.

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

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

Following from the definition above, a covariance matrix $\boldsymbol{\Sigma}$ can be factorised into the product of standard deviations $\boldsymbol{V}^{1/2}$ and the correlation matrix $\boldsymbol{P}$ as follows:

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

2.14 Parameters and families of distributions

A **distribution family** $F(\theta)$ is a collection of distributions obtained by varying a parameter θ. Each specific value of the parameter θ indexes one distribution in that family.

Common distribution families are usually denoted by familiar abbreviation such as $N(\mu, \sigma^2)$ for the normal family. We also call these simply "distributions" with parameters and omit the word "family".

If a random variable x has distribution $F(\theta)$ we write $x \sim F(\theta)$ or simply $x \sim N(\mu, \sigma^2)$ in case of a named normal distribution.

The associated pdmf is written $f(x; \theta)$ or $f(x|\theta)$. The conditional notation is more general because it implies the parameter θ may have its own distribution, yielding a joint density $f(x, \theta) = f(x|\theta)f(\theta)$. Similarly, the corresponding cumulative distribution function is written $F(x; \theta)$ or $F(x|\theta)$.

Note that parametrisations are generally not unique, as any one-to-one transformation of θ yields an equivalent index of the same distribution family. For most commonly used distribution families there exist several standard parametrisations. We usually prefer those whose parameters that can be interpreted easily (e.g. in terms of moments) or that help to simplify calculations.

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 $F(\theta)$ is said to be **identifiable** by the parameter θ.

3 Transformations

3.1 Affine or location-scale transformation

Transformation rule

Suppose $x \sim F_x$ is a scalar random variable. 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 $a = 0$ this is a **linear transformation**. If $b \neq 0$ then the transformation is **invertible**, with back-transformation

$$x = (y - a)/b$$

Invertible transformations provide a one-to-one map between x and y.

For a random vector $\boldsymbol{x} \sim F_{\boldsymbol{x}}$ of dimension d the location-scale transformation is

$$\boldsymbol{y} = \boldsymbol{a} + \boldsymbol{B}\boldsymbol{x}$$

where $\boldsymbol{a}$ (a $m \times 1$ vector) is the **location parameter** and $\boldsymbol{B}$ (a $m \times d$ matrix) the **scale parameter** For $m = d$ (square $\boldsymbol{B}$) and $\det(\boldsymbol{B}) \neq 0$ the affine transformation is **invertible** with back-transformation

$$\boldsymbol{x} = \boldsymbol{B}^{-1}(\boldsymbol{y} - \boldsymbol{a})$$

Density

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

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

where $|b|$ is the absolute value of b. Likewise, assuming an invertible transformation for a continuous random vector $\boldsymbol{x}$ with density $f_{\boldsymbol{x}}(\boldsymbol{x})$ the density for $\boldsymbol{y}$ is given by

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

where $|\det(\boldsymbol{B})|$ is the absolute value of the determinant $\det(\boldsymbol{B})$.

Moments

The transformed random variable $y \sim F_y$ has mean

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

and variance

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

where $\mathrm{E}(x) = \mu_x$ and $\mathrm{Var}(x) = \sigma_x^2$ are the mean and variance of the original variable x.

The mean and variance of the transformed random vector $\boldsymbol{y} \sim F_{\boldsymbol{y}}$ is

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

and

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

where $\mathrm{E}(\boldsymbol{x}) = \boldsymbol{\mu}_{\boldsymbol{x}}$ and $\mathrm{Var}(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$ are the mean and variance of the original random vector $\boldsymbol{x}$.

Importance of affine transformations

The constants $\boldsymbol{a}$ and $\boldsymbol{B}$ (or a and b in the univariate case) are the parameters of the **location-scale family** $F_{\boldsymbol{y}}$ created from $F_{\boldsymbol{x}}$. Many important distributions are location-scale families such as the normal distribution (cf. Section 4.3 and Section 5.3) and the location-scale t-distribution (Section 4.6 and Section 5.6). 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

Transformation rule

As above we assume $x \sim F_x$ is a scalar random variable and $\boldsymbol{x} \sim F_{\boldsymbol{x}}$ is a random vector.

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

Density

If x is a continuous random variable with density $f_x(x)$ the density of the transformed variable y can be computed exactly and is given by

$$f_y(y) = \left|Dx(y)\right| f_x(x(y))$$

where $Dx(y)$ is the derivative of the inverse transformation $x(y)$.

Likewise, for a continuous random vector $\boldsymbol{x}$ with density $f_{\boldsymbol{x}}(\boldsymbol{x})$ the density for $\boldsymbol{y}$ is obtained by

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

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

Moments

The mean and variance of the transformed random variable can typically only be approximated. Assume that $\mathrm{E}(x) = \mu_x$ and $\mathrm{Var}(x) = \sigma_x^2$ are the mean and variance of the original random variable x and $\mathrm{E}(\boldsymbol{x}) = \boldsymbol{\mu}_{\boldsymbol{x}}$ and $\mathrm{Var}(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$ are the mean and variance of the original random vector $\boldsymbol{x}$. In the **delta method** the transformation $y(x)$ resp. $\boldsymbol{y}(\boldsymbol{x})$ is linearised around the mean μ_x respectively $\boldsymbol{\mu}_{\boldsymbol{x}}$ and the mean and variance resulting from the linear transformation is reported.

Specifically, the linear approximation for the scalar-valued function is

$$y(x) \approx y\left(\mu_x\right) + Dy\left(\mu_x\right)\ \left(x - \mu_x\right)$$

where $Dy(x) = y'(x)$ is the first derivative of the transformation $y(x)$ and $Dy\left(\mu_x\right)$ is the first derivative evaluated at the mean μ_x, and for the vector-valued function

$$\boldsymbol{y}(\boldsymbol{x}) \approx \boldsymbol{y}\left(\boldsymbol{\mu}_{\boldsymbol{x}}\right) + D\boldsymbol{y}\left(\boldsymbol{\mu}_{\boldsymbol{x}}\right)\ \left(\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{x}}\right)$$

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

In the univariate case the delta method yields as approximation for the mean and variance of the transformed random variable y

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

and

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

For the vector random variable $\boldsymbol{y}$ the delta method yields

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

and

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

Assuming $y(x) = a + bx$, with $x(y) = (y - a)/b$, $Dy(x) = b$ and $Dx(y) = b^{-1}$, recovers the univariate location-scale transformation. Likewise, 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 multivariate 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 $\boldsymbol{t}(x)$ 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 $\boldsymbol{t}(x)$ is small.

The exponential tilt of a **base distribution** B with base function $h(x)$ (possibly unnormalised) toward the linear combination $\boldsymbol{\eta}^T\boldsymbol{t}(x)$ of the canonical statistics $\boldsymbol{t}(x)$ and the **canonical parameters** $\boldsymbol{\eta}$ yields the distribution family $P(\boldsymbol{\eta})$ with pdmf

$$p(x|\boldsymbol{\eta}) = \underbrace{e^{\boldsymbol{\eta}^T\boldsymbol{t}(x)}}_{\text{exponential tilt}}\ h(x)\,/\,z(\boldsymbol{\eta})$$

The **normaliser** or **partition function** $z(\boldsymbol{\eta})$ ensures that $p(x|\boldsymbol{\eta})$ integrates to one, with

$$z(\boldsymbol{\eta}) = \int_x e^{\boldsymbol{\eta}^T\boldsymbol{t}(x)}\, h(x)\, dx$$

In particular, $z(\boldsymbol{0}) = \int_x h(x)\, dx$ ensures that

$$p(x|\boldsymbol{0}) = h(x)/z(\boldsymbol{0}) = b(x)$$

is a valid base pdmf. If $h(x)$ is a pdmf then $z(\boldsymbol{0}) = 1$ and $b(x) = h(x)$.

A distribution family $P(\boldsymbol{\eta})$ obtained by exponential tiling is called an **exponential family**. The set of values of $\boldsymbol{\eta}$ for which $z(\boldsymbol{\eta}) < \infty$, and hence for which $p(x|\boldsymbol{\eta})$ is well defined, comprises the parameter space of the exponential family. Some choices of $h(x)$ and $\boldsymbol{t}(x)$ do not yield a finite normalising factor for any $\boldsymbol{\eta}$ and hence these cannot be used to form an exponential family.

The **log-normaliser** or **log-partition function** $a(\boldsymbol{\eta}) = \log z(\boldsymbol{\eta})$ is the cumulant generating function for the canonical statistics. Its gradient yields the mean

$$\mathrm{E}(\boldsymbol{t}(x)) = \boldsymbol{\mu}_{\boldsymbol{t}} = \nabla a(\boldsymbol{\eta})$$

and the Hessian matrix the variance

$$\mathrm{Var}(\boldsymbol{t}(x)) = \boldsymbol{\Sigma}_{\boldsymbol{t}} = \nabla\nabla^T a(\boldsymbol{\eta})$$

Many common distributions are exponential families, such as the normal distribution and the Bernoulli distribution. Exponential families are central in probability and statistics. They support effective statistical learning using likelihood and Bayesian approaches, enable data reduction via minimal sufficiency and provide the basis for generalised linear models. Furthermore, exponential families often allow to generalise results established for specific cases, such as the normal distribution, to a broader domain.

See also (Wikipedia): exponential family — table of distributions.

3.4 Sums of random variables and convolution

Moments

Suppose we have a sum of n *independent* random variables.

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

where each $x_i \sim F_{x_i}$ has its own distribution and corresponding probability density mass function $f_{x_i}(x)$.

With $\boldsymbol{x} = (x_1, \ldots, x_n)^T$ and $\mathbf{1}_n = (1, 1, \ldots, 1)^T$ the relationship between y and $\boldsymbol{x}$ can be written as affine transformation $y = \mathbf{1}_n^T \boldsymbol{x}$. Assuming $\mathrm{E}(x_i) = \mu_i$, $\mathrm{Var}(x_i) = \sigma_i^2$ and $\mathrm{Cov}(x_i, x_j) = 0$ for $i \neq j$ the mean and variance of the random variable y equals (cf. Section 3.1)

$$\mathrm{E}(y) = \mathbf{1}_n^T \boldsymbol{\mu} = \sum_{i=1}^{n} \mu_i$$

and

$$\mathrm{Var}(y) = \mathbf{1}_n^T \mathrm{Var}(\boldsymbol{x}) \mathbf{1}_n = \sum_{i=1}^{n} \sigma_i^2$$

Thus both the means and variance are additive (but note that for the variance this is only true because of the independence assumption).

Convolution

The pdmf for y is obtained by repeatedly convolving (denoted by the asterisk $*$ operator) the pdmfs of the x_i:

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

The **convolution** of two functions is defined as (continuous case)

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

and (discrete case)

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

Convolution is commutative and associative so so you may convolve multiple pdmfs in any order or grouping. Furthermore, the convolution of pdmfs yields another pdmf, i.e. the resulting function integrates to one.

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 (Wikipedia): list of convolutions of probability distributions.

Central limit theorem

The **central limit theorem**, first postulated by Abraham de Moivre (1667–1754) and later proved by Pierre-Simon Laplace (1749–1827) asserts that the distribution of the sum of n independent and identically distributed random variables with finite mean and finite variance converges in the limit of large n to a normal distribution (Section 4.3), even if the individual random variables are not themselves normal. In other words, it asserts that for large n the convolution of n identical distributions with finite first two moments converges to the normal distribution.

3.5 Loss functions and scoring rules

Loss function

A **loss or cost function** $L(x, a)$ evaluates a prediction a (for example a parameter or a probability distribution) on the basis of an observed outcome x, and returns a numerical score.

A loss function measures, informally, the error between x and a. During optimisation the prediction a is varied and the aim is minimisation of the error (hence a loss function has *negative orientation*, smaller is better).

Adding a constant or a positive scaling factor to the loss function will not change the location of its minimum, so such loss functions are considered equivalent.

A **utility or reward function** is a loss function with a reversed sign (hence it has *positive orientation*, larger is better).

Risk function

The **risk** of a under the distribution Q for x is defined as the expected loss

$$R_Q(a) = \mathrm{E}_Q(L(x, a))$$

If there is no ambiguity we drop the reference to Q and write

$$R(a) = \mathrm{E}(L(x, a))$$

The risk of a under the empirical distribution $\hat{Q}_n$ obtained from observations $x_1, \ldots, x_n$ is the **empirical risk**

$$\hat{R}(a) = R_{\hat{Q}_n}(a) = \frac{1}{n} \sum_{i=1}^{n} L(x_i, a)$$

where the expectation is replaced by the sample average.

Minimising $R(a)$ finds optimal predictions

$$a^* = \arg\min_a R(a)$$

Depending on the choice of underlying loss $L(x, a)$ minimising the risk provides a very general optimisation-based way to identify distributional features of the distribution Q and to obtain parameter estimates.

Scoring rules

A **scoring rule** $S(x, P)$ is special type of loss function[1] that assesses the probabilistic forecast P by assigning a numerical score based on P and the observed outcome x.

The associated **risk** of P under Q is

$$R_Q(P) = \mathrm{E}_Q\left(S(x, P)\right)$$

For a **proper** scoring rule the risk $R_Q(P)$ is minimised at $P = Q$, hence

$$R_Q(P) \geq R_Q(Q)$$

For a **strictly proper** scoring rule the minimum is achieved only at the true distribution Q, so equality holds exclusively for $P = Q$.

[1] As a loss function, scoring rules are negatively oriented. However, some authors consider them as utility functions with positive orientation.

A proper scoring rule induces a **divergence** between the distributions Q and P, as the difference between the risk and the minimum risk :

$$D(Q, P) = R_Q(P) - R_Q(Q) \geq 0$$

By construction, the divergence $D(Q, P)$ is always **non-negative** and equals zero if $P = Q$. For a stricly proper scoring rule the divergence vanishes exclusively for $P = Q$.

Proper scoring rules are very useful as they allow to identify the underlying distribution and their parameters by risk minimisation or minimisation of the associated divergences.

Proper scoring rules also have a number of further useful properties. For example, various **decompositions** exist for their risk, and the divergence satisfies a **generalised Pythagorean theorem**. Furthermore, there is a **correspondence** of proper scoring rules and their associated divergences with **Bregman divergences**.

Common loss functions

The **squared loss** or **squared error** is one of the most commonly used loss functions:

$$L(x, a) = (x - a)^2$$

The corresponding risk is the **mean squared loss** or **mean squared error** (MSE)

$$R(a) = \mathrm{E}((x - a)^2)$$

From $R(a) = \mathrm{E}((x - a)^2) = \mathrm{E}(x^2) - 2a\mathrm{E}(x) + a^2$ it follows $dR(a)/da = -2\mathrm{E}(x) + 2a$ and thus that the MSE is **minimised at the mean** $a^* = \mathrm{E}(x)$. The **achieved minimum risk** $R(a^*) = \mathrm{Var}(x)$ is the **variance**.

The **0-1 loss** function can be written as

$$L(x, a) = \begin{cases} -[x = a] & \text{discrete case} \\ -\delta(x - a) & \text{continuous case} \end{cases}$$

employing the indicator function and Dirac delta function, respectively. The corresponding risk assuming $x \sim Q$ and pdmf $q(x)$ is

$$R_Q(a) = -q(a)$$

which is **minimised at the mode** of the pdmf.

The **asymmetric loss** can be defined as

$$L(x, a; \tau) = \begin{cases} 2\tau(x-a) & \text{for } x \geq a \\ 2(1-\tau)(a-x) & \text{for } x < a \end{cases}$$

and the corresponding risk is **minimised at the quantile** x_τ.

For $\tau = 1/2$ it reduces to the **absolute loss**

$$L(x, a) = |x - a|$$

whose corresponding risk is **minimised at the median** $x_{1/2}$.

Logarithmic scoring rule

The most important scoring rule is the **logarithmic scoring rule** or **log-loss**

$$S(x, P) = -\log p(x)$$

The risk of P under Q based on the log-loss is the **mean log-loss** or **cross-entropy**

$$R_Q(P) = -\mathrm{E}_Q \log p(x) = H(Q, P)$$

which is uniquely minimised for $P = Q$. Thus, the log-loss is **strictly proper**. Furthermore, the log-loss is notably the only **local** strictly proper scoring rule, as it solely depends on the value of the pdmf at the observed outcome x, and not on any other features of the distribution P. The minimum risk is the Shannon-Gibbs **entropy** of Q:

$$R_Q(Q) = -\mathrm{E}_Q \log q(x) = H(Q)$$

The relationship $H(Q, P) \geq H(Q)$, with equality exclusively for $P = Q$, is known as **Gibbs' inequality**.

The divergence induced by the log-loss is the Kullback-Leibler (KL) divergence

$$D_{\mathrm{KL}}(Q, P) = H(Q, P) - H(Q) = \mathrm{E}_Q \log\left(\frac{q(x)}{p(x)}\right)$$

The KL divergence obeys the **data processing inequality**, i.e. applying a transformation to the underlying random variables cannot increase the KL divergence $D_{\mathrm{KL}}(Q, P)$ between Q and P. This property also holds for all f-divergences (of which the KL divergence is a principal example),

but is notably *not* satisfied by divergences of other proper scoring rules (and thus other Bregman divergences).

Furthermore, the KL divergence is the only divergence induced by proper scoring rules (and thus the only Bregman divergence), as well as the only f-divergence, that is **invariant against general coordinate transformations**. Coordinate transformations can be viewed as a special case of data processing, and for $D_{\mathrm{KL}}(Q, P)$ the data-processing inequality under general invertible transformations becomes an identity.

The empirical risk of a distribution family $P(\theta)$ based on the log-loss is proportional to the log-likelihood function

$$\hat{R}(\theta) = H(\hat{Q}_n, P(\theta)) = -\frac{1}{n}\sum_{i=1}^{n} \log p(x_i|\theta) = -\frac{1}{n}\ell_n(\theta)$$

Minimising the empirical risk $\hat{R}(\theta)$ is equivalent to maximising the log-likelihood function $\ell_n(\theta)$.

Similary, minimising the KL divergence $D_{\mathrm{KL}}(\hat{Q}_n, P(\theta))$ with regard to θ is equivalent to minimising the empirical risk $\hat{R}(\theta)$ and hence to maximum likelihood.

Brier or quadratic scoring rule

The **Brier scoring rule**, also known as **quadratic scoring rule**, evaluates a probabilistic categorical forecast P with corresponding class probabilities $p_1, \ldots, p_K$ given a realisation $\boldsymbol{x}$ from the categorical distribution Q with class probabilities $q_1, \ldots, q_K$. It can be written as

$$\begin{aligned} S(\boldsymbol{x}, P) &= \sum_{y=1}^{K} \left(x_y - p_y\right)^2 \\ &= 1 - 2\sum_{y=1}^{K} x_y p_y + \sum_{y=1}^{K} p_y^2 \\ &= 1 - 2p_k + \sum_{y=1}^{K} p_y^2 \end{aligned}$$

The indicator vector $\boldsymbol{x} = (x_1, \ldots, x_K)^T = (0, 0, \ldots, 1, \ldots, 0)^T$ contains zeros everywhere except for a single element $x_k = 1$. Unlike the log-loss,

the Brier score is *not local* as the pmf for P is evaluated across all K classes, not just at the realised class k.

The corresponding risk is

$$R_Q(P) = \mathrm{E}_Q(S(\boldsymbol{x}, P)) = 1 - 2\sum_{y=1}^{K} q_y p_y + \sum_{y=1}^{K} p_y^2$$

which is uniquely minimised for $P = Q$. Thus, the Brier score is **strictly proper**. The minimum risk is

$$R_Q(Q) = 1 - \sum_{y=1}^{K} q_y^2$$

The divergence induced by the Brier score is the **squared Euclidean distance** between the two pmfs:

$$D(Q, P) = R_Q(P) - R_Q(Q) = \sum_{y=1}^{K} \left(q_y - p_y\right)^2$$

Other proper scoring rules

An example of a proper, but not strictly proper, scoring rule is the squared error relative to the mean of the quoted model P:

$$S(x, P) = (x - \mathrm{E}(P))^2$$

The corresponding risk is

$$R_Q(P) = \mathrm{E}_Q\left((x - \mathrm{E}(P))^2\right)$$

which is minimised for $P = Q$ but also for any other distribution P with the same mean as Q. The minimum risk is the variance of Q:

$$R_Q(Q) = \mathrm{E}_Q\left((x - \mathrm{E}(Q))^2\right) = \mathrm{Var}(Q)$$

The associated divergence is the squared distance between the two means

$$D(Q, P) = R_Q(P) - \mathrm{Var}(Q) = (\mathrm{E}(Q) - \mathrm{E}(P))^2$$

which vanishes for $P = Q$ but also for any other P with $\mathrm{E}(P) = \mathrm{E}(Q)$.

Other useful strictly proper scoring rules include

- the continuous ranked probability score (CRPS),
- the energy score, and
- the Hyvärinen scoring rule.

See also (Wikipedia): scoring rule.

4 Univariate distributions

4.1 Binomial distribution

The **binomial distribution** $\text{Bin}(n, \theta)$ is a discrete distribution counting binary outcomes.

The **Bernoulli distribution** $\text{Ber}(\theta)$ is a special case of the binomial distribution.

Standard parametrisation

A binomial random variable x describes the number of successful outcomes in n identical and independent trials. We write

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

where $\theta \in [0, 1]$ is the probability of a positive outcome ("success") in a single trial. Conversely, $1 - \theta \in [0, 1]$ is the complementary probability ("failure"). The support is $x \in \{0, 1, 2, \ldots, n\}$ which notably depends on n.

The binomial distribution is often motivated by a coin tossing experiment where θ is the probability of "head" when flipping the coin and x is the number of observed "heads" among n throws. Another common interpretation is that of an urn model where n items are distributed into two bins (Figure 4.1). Here θ is the probability to put an item into one urn (representing "success", "head") and $1 - \theta$ the probability to put it in the other urn (representing "failure", "tail").

The expected value is

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

and the variance is

$$\text{Var}(x) = n\theta(1 - \theta)$$

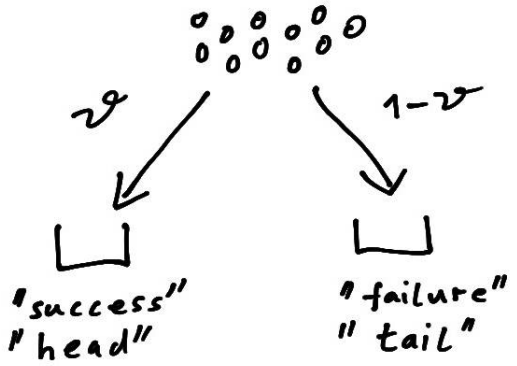

Figure 4.1: Binomial urn model.

The corresponding pmf is

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

The binomial coefficient $\binom{n}{x}$ in the pdf accounts for the multiplicity of ways in which we can observe x successes in n trials.

> R code
>
> The pmf of the binomial distribution is given by `dbinom()`, the distribution function is `pbinom()` and the quantile function is `qbinom()`. The corresponding random number generator is `rbinom()`.

Mean parametrisation

Instead of θ one may also use a mean parameter $\mu \in [0, n]$ so that

$$x \sim \text{Bin}\left(n, \theta = \frac{\mu}{n}\right)$$

The mean parameter μ can be obtained from θ and n by $\mu = n\theta$.

The mean and variance of the binomial distribution expressed in terms of μ and n are

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

and

$$\text{Var}(x) = \mu - \frac{\mu^2}{n}$$

Special case: Bernoulli distribution

For $n = 1$ the binomial distribution reduces to the **Bernoulli distribution** $\text{Ber}(\theta)$. This is the simplest of all distribution families and is named after Jacob Bernoulli (1655-1705) who also discovered the law of large numbers.

If a random variable x follows the Bernoulli distribution we write

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

with "success" probability $\theta \in [0, 1]$. Conversely, the complementary "failure" probability is $1 - \theta \in [0, 1]$. The support is $x \in \{0, 1\}$. The variable x acts as an indicator variable, with "success" indicated by $x = 1$ and "failure" indicated by $x = 0$.

Often the Bernoulli distribution is referred to as "coin flipping" model. Then θ is the probability of "head" and $1 - \theta$ the complementary probability of "tail" and $x = 1$ corresponds to the outcome "head" and $x = 0$ to the outcome "tail".

The expected value is

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

and the variance is

$$\text{Var}(x) = \theta(1 - \theta)$$

The pmf of $\text{Ber}(\theta)$ is

$$p(x|\theta) = \theta^x(1 - \theta)^{1-x} = \begin{cases} \theta & \text{if } x = 1 \\ 1 - \theta & \text{if } x = 0 \end{cases}$$

Convolution property and normal approximation

The convolution of n binomial distributions, each with identical success probability θ but possibly different number of trials n_i, yields another binomial distribution with the same parameter θ:

$$\sum_{i=1}^{n} \text{Bin}(n_i, \theta) \sim \text{Bin}\left(\sum_{i=1}^{n} n_i, \theta\right)$$

It follows that the binomial distribution with n trials is the result of the convolution of n Bernoulli distributions:

$$\sum_{i=1}^{n} \text{Ber}(\theta) \sim \text{Bin}(n, \theta)$$

Thus, repeating the same Bernoulli trial n times and counting the total number of successes yields a binomial random variable.

As a consequence, following the central limit theorem (Section 3.4), for large n the binomial distribution can be well approximated by a normal distribution (Section 4.3) with the same mean and variance. This is known as the De Moivre–Laplace theorem.

4.2 Beta distribution

The **beta distribution** $\text{Beta}(\alpha_1, \alpha_2)$ is a continuous distribution that is useful to model proportions or probabilities for $K = 2$ classes.

It includes the **uniform distribution** over the unit interval as a special case.

Standard parametrisation

A beta-distributed random variable is denoted by

$$x \sim \text{Beta}(\alpha_1, \alpha_2)$$

with shape parameters $\alpha_1 > 0$ and $\alpha_2 > 0$. Let $m = \alpha_1 + \alpha_2$. The support of x is the unit interval given by $x \in [0, 1]$. Thus, the beta distribution is defined over a one-dimensional space.

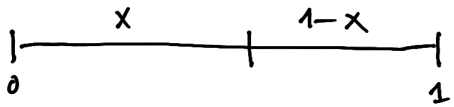

Figure 4.2: Stick breaking visualisation of a beta random variable.

A beta random variable can be visualised as breaking a unit stick of length one into two pieces of length $x_1 = x$ and $x_2 = 1 - x$ (Figure 4.2).

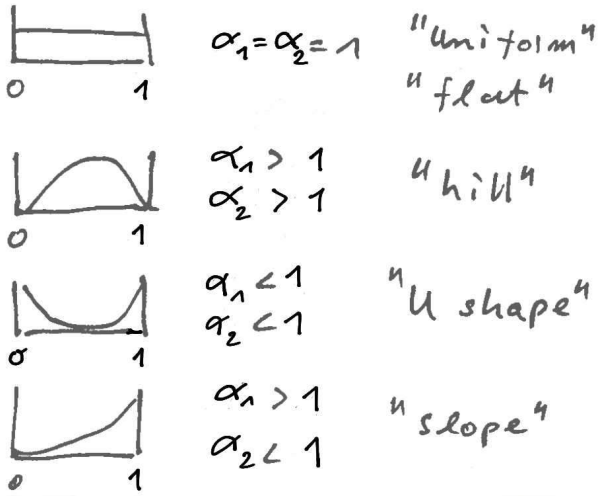

Figure 4.3: Shapes of the pdf of the beta distribution.

Thus, the x_i may be used as the exclusive proportions or probabilities for $K = 2$ classes.

The mean is

$$\mathrm{E}(x) = \mathrm{E}(x_1) = \frac{\alpha_1}{m}$$

and hence

$$\mathrm{E}(1 - x) = \mathrm{E}(x_2) = \frac{\alpha_2}{m}$$

The variance is

$$\mathrm{Var}(x) = \mathrm{Var}(x_1) = \mathrm{Var}(x_2) = \frac{\alpha_1 \alpha_2}{m^2(m+1)}$$

The pdf of the beta distribution $\mathrm{Beta}(\alpha_1, \alpha_2)$ is

$$p(x|\alpha_1, \alpha_2) = \frac{1}{B(\alpha_1, \alpha_2)} x^{\alpha_1 - 1}(1 - x)^{\alpha_2 - 1}$$

This depends on the beta function with arguments α_1 and α_2 defined as

$$B(\alpha_1, \alpha_1) = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(m)}$$

The beta distribution can assume a number of different shapes, depending on the values of α_1 and α_2 (see Figure 4.3).

R code

The pdf of the beta distribution is given by `dbeta()`, the distribution function is `pbeta()` and the quantile function is `qbeta()`. The corresponding random number generator is `rbeta()`.

Mean parametrisation

Instead of employing α_1 and α_2 as parameters another useful reparametrisation of the beta distribution is in terms of a mean parameter $\mu \in [0, 1]$ and a concentration parameter $m > 0$ so that

$$x \sim \mathrm{Beta}(\alpha_1 = m\mu, \alpha_2 = m(1 - \mu))$$

The concentration and mean parameters can be obtained from α_1 and α_2 by $m = \alpha_1 + \alpha_2$ and $\mu = \alpha_1/m$.

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

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

and

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

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

Special case: symmetric beta distribution

For $\alpha_1 = \alpha_2 = \alpha$ the beta distribution becomes the **symmetric beta distribution** with a single shape parameter $\alpha > 0$. In mean parametrisation the symmetric beta distribution corresponds to $\mu = 1/2$ and $m = 2\alpha$.

Special case: uniform distribution

For $\alpha_1 = \alpha_2 = 1$ the beta distribution becomes the **uniform distribution over the unit interval** with pdf $p(x) = 1$. In mean parametrisation the uniform distribution corresponds to $\mu = 1/2$ and $m = 2$.

4.3 Normal distribution

The **normal distribution** $N(\mu, \sigma^2)$ is the most important continuous probability distribution. It is also called **Gaussian distribution** named after Carl Friedrich Gauss (1777–1855).

Special cases are the **standard normal distribution** $N(0, 1)$ and the **delta distribution** δ.

Standard parametrisation

The univariate normal distribution $N(\mu, \sigma^2)$ has two parameters μ (location) and $\sigma^2 > 0$ (variance) and support $x \in \mathbb{R}$.

If a random variable x is normally distributed we write

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

with mean

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

and variance

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

The pdf is given by

$$\begin{aligned} p(x|\mu, \sigma^2) &= (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \\ &= (\sigma^2)^{-1/2}(2\pi)^{-1/2} e^{-\Delta^2/2} \end{aligned}$$

Here $\Delta^2 = (x-\mu)^2/\sigma^2$ is the squared distance between x and μ weighted by the variance σ^2, also known as **squared Mahalanobis distance**.

The normal distribution is sometimes also used by specifying the precision $1/\sigma^2$ instead of the variance σ^2.

> R code
>
> The normal pdf is given by `dnorm()`, the distribution function is `pnorm()` and the quantile function is `qnorm()`. The corresponding random number generator is `rnorm()`.

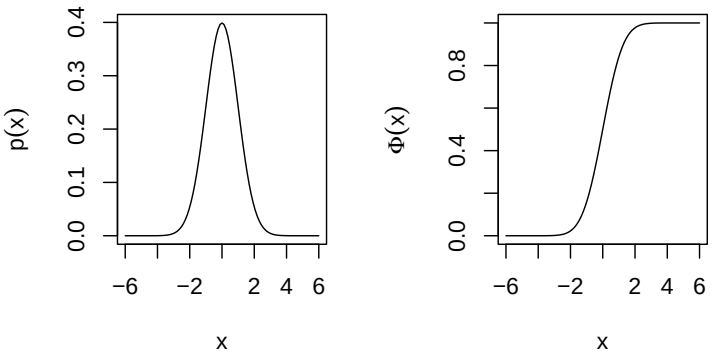

Figure 4.4: Probability density function (left) and cumulative density function (right) of the standard normal distribution.

Scale parametrisation

Instead of the variance parameter σ^2 it is often also convenient to use the standard deviation $\sigma = \sqrt{\sigma^2} > 0$ as scale parameter. Similarly, instead of the precision $1/\sigma^2$ one may wish to use the inverse standard deviation $w = 1/\sigma$.

The scale parametrisation is central for location-scale transformations (see below).

Special case: standard normal distribution

The **standard normal distribution** $N(0, 1)$ has mean $\mu = 0$ and variance $\sigma^2 = 1$. The corresponding pdf is

$$p(x) = (2\pi)^{-1/2} e^{-x^2/2}$$

with the squared Mahalanobis distance reduced to $\Delta^2 = x^2$.

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 $\Phi(x)$. The inverse $\Phi^{-1}(p)$ is called the quantile function of the standard normal distribution.

Figure 4.4 shows the pdf and cdf of the standard normal distribution.

Special case: delta distribution

The **delta distribution** δ is obtained as the limit of $N(0, \varepsilon\sigma^2)$ for $\varepsilon \to 0$ and where σ^2 is a positive number (e.g. $\sigma^2 = 1$). Thus δ is a distribution that behaves like an infinite spike at zero.

The corresponding pdf $\delta(x)$ is called the **Dirac delta function**, even though it is not an ordinary function. It satisfies $\delta(x) = 0$ for all $x \neq 0$ and integrates to one, thus representing a point mass at zero.

Location-scale transformation

Let $\sigma > 0$ be the positive square root of the variance σ^2 and $w = 1/\sigma$.

If $x \sim N(\mu, \sigma^2)$ then $y = w(x - \mu) \sim N(0, 1)$. This location-scale transformation corresponds to centring and standardisation of a normal random variable, reducing it to a standard normal random variable.

Conversely, if $y \sim N(0, 1)$ then $x = \mu + \sigma y \sim N(\mu, \sigma^2)$. This location-scale transformation generates the normal distribution from the standard normal distribution.

Convolution property

The convolution of n independent, but not necessarily identical, normal distributions results in another normal distribution with corresponding mean and variance:

$$\sum_{i=1}^{n} N(\mu_i, \sigma_i^2) \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

Hence, any normal random variable can be constructed as the sum of n suitable independent normal random variables.

Since n is an arbitrary positive integer the normal distribution is said to be **infinitely divisible**.

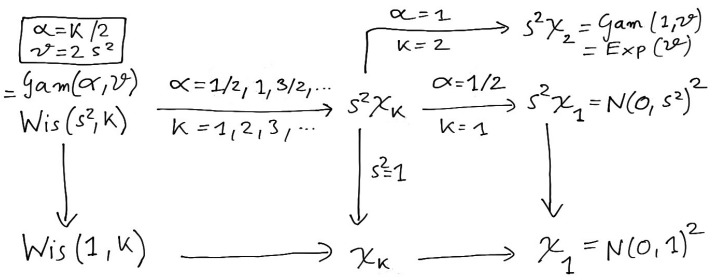

Figure 4.5: The gamma and the univariate Wishart distribution and its relatives.

4.4 Gamma distribution

The **gamma distribution** $\text{Gam}(\alpha, \theta)$ is another widely used continuous distribution and is also known as **univariate Wishart distribution** $\text{Wis}\left(s^2, k\right)$ using a different parametrisation.

It contains as special cases the **scaled chi-squared distribution** $s^2\chi_k^2$ (two parameter restrictions) as well as the **univariate standard Wishart distribution** $\text{Wis}\,(1, k)$, the **chi-squared distribution** χ_k^2 and the **exponential distribution** $\text{Exp}(\theta)$ (one parameter restrictions). Figure 4.5 illustrates the relationship of the gamma and the univariate Wishart distribution with these related distributions.

Standard parametrisation

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 $\beta = 1/\theta$. Therefore one needs to pay attention which parametrisation is used.

The pdf is

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

> R code
>
> The pdf of the gamma distribution is available in the function `dgamma()`, the distribution function is `pgamma()` and the quantile function is `qgamma()`. The corresponding random number generator is `rgamma()`.

Wishart parametrisation

The gamma distribution is often used with a different set of parameters $s^2 = \theta/2 > 0$ (scale) and $k = 2\alpha > 0$ (shape or concentration). In this form it is known as **univariate or one-dimensional Wishart distribution**

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

named after John Wishart (1898–1954).

In the above the scale parameter s^2 is scalar and hence the resulting Wishart distribution is univariate. If instead a matrix-valued scale parameter $\boldsymbol{S}$ is used this yields the multivariate or d-dimensional Wishart distribution, see Section 5.4.

In the Wishart parametrisation the mean is

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

and the variance

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

The pdf in terms of s^2 and k is

$$p(x|s^2,k) = \frac{1}{\Gamma(k/2)(2s^2)^{k/2}} x^{(k-2)/2} e^{-s^{-2}x/2}$$

Mean parametrisation

Finally, we also often employ the Wishart resp. gamma distribution in **mean parametrisation**

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

with parameters $\mu = ks^2 > 0$ and $k > 0$. In this parametrisation the mean is

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

and the variance

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

Special case: univariate standard Wishart distribution

For $s^2 = 1$ the univariate Wishart distribution reduces to the **univariate standard Wishart distribution**

$$x \sim \text{Wis}\left(1, k\right) = \text{Gam}\left(\alpha = \frac{k}{2}, \theta = 2\right)$$

with mean

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

and variance

$$\text{Var}(x) = 2k$$

The pdf is

$$p(x|k) = \frac{1}{\Gamma(k/2)2^{k/2}} x^{(k-2)/2} e^{-x/2}$$

Special case: scaled chi-squared distribution

If the shape parameter k of the Wishart distribution $\text{Wis}\left(s^2, k\right)$ is restricted to the positive integers $k \in \{1, 2, \ldots\}$ the Wishart distribution becomes the **scaled chi-squared distribution** $s^2\chi^2_k$ where k is called the degree of freedom.

This is equivalent to restricting the shape parameter α of the gamma distribution $\text{Gam}(\alpha = k/2, \theta = 2s^2)$ to $\alpha \in \{1/2, 1, 3/2, 2, \ldots\}$.

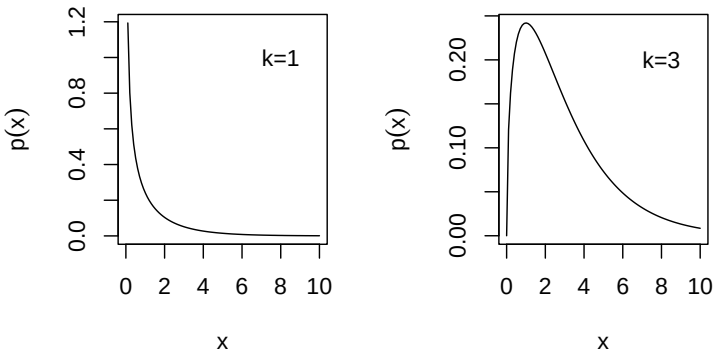

Figure 4.6: Probability density function of the chi-squared distribution.

The scaled chi-squared distribution with $k = 1$ is the distribution of a squared normal random variable with mean zero. Specifically, if $z \sim N(0, s^2)$ then $z^2 \sim s^2\chi^2_1 = \text{Wis}(s^2, 1) = N(0, s^2)^2$.

Special case: chi-squared distribution

If k is restricted to the positive integers $k \in \{1, 2, \ldots\}$ the univariate standard Wishart distribution reduces to the **chi-squared distribution**

$$x \sim \chi^2_k = \text{Wis}\left(s^2 = 1, k\right) = \text{Gam}\left(\alpha = \frac{k}{2}, \theta = 2\right)$$

where k is called the degree of freedom.

The chi-squared distribution has mean

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

and variance

$$\text{Var}(x) = 2k$$

Figure 4.6 shows the pdf of the chi-squared distribution for degrees of freedom $k = 1$ and $k = 3$.

The chi-squared distribution with $k = 1$ is the distribution of a squared standard normal random variable. Specifically, if $z \sim N(0, 1)$ then $z^2 \sim \chi^2_1 = \text{Wis}(1, 1) = N(0, 1)^2$.

R code

The pdf of the chi-squared distribution is given by `dchisq()`. The distribution function is `pchisq()` and the quantile function is `qchisq()`. The corresponding random number generator is `rchisq()`.

Special case: exponential distribution

If the shape parameter α of the gamma distribution $\text{Gam}(\alpha, \theta)$ is set to $\alpha = 1$, or if the shape parameter k of the Wishart distribution $\text{Wis}(s^2, k)$ is set to $k = 2$, we obtain the **exponential distribution**

$$x \sim \text{Exp}(\theta) = \text{Gam}(\alpha = 1, \theta) = \text{Wis}(s^2 = \theta/2, k = 2)$$

with scale parameter θ.

It has mean

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

and variance

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

and the pdf is

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

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.

R code

The command `dexp()` returns the pdf of the exponential distribution, `pexp()` is the distribution function and `qexp()` is the quantile function. The corresponding random number generator is `rexp()`.

Scale transformation

If $x \sim \text{Gam}(\alpha, \theta)$ then the scaled random variable bx with $b > 0$ is also gamma distributed with $bx \sim \text{Gam}(\alpha, b\theta)$.

Hence,

- $\theta \,\text{Gam}(\alpha, 1) = \text{Gam}(\alpha, \theta)$,
- $\theta \,\text{Exp}(1) = \text{Exp}(\theta)$,
- $(\mu/k)\,\text{Wis}(1, k) = \text{Wis}(s^2 = \mu/k, k)$ and
- $s^2\,\text{Wis}(1, k) = \text{Wis}(s^2, k)$.

As χ^2_k equals $\text{Wis}(1, k)$ the last example demonstrates that the **scaled chi-squared distribution** $s^2\chi^2_k$ equals the univariate Wishart distribution $\text{Wis}(s^2, k)$.

Convolution property

The convolution of n gamma distributions with the same scale parameter θ but possible different shape parameters α_i yields another gamma distribution:

$$\sum_{i=1}^{n} \text{Gam}(\alpha_i, \theta) \sim \text{Gam}\left(\sum_{i=1}^{n} \alpha_i, \theta\right)$$

Thus, any gamma random variable can be obtained as the sum of n suitable independent gamma random variables.

In Wishart parametrisation this becomes

$$\sum_{i=1}^{n} \text{Wis}(s^2, k_i) \sim \text{Wis}\left(s^2, \sum_{i=1}^{n} k_i\right)$$

As a result, since n is an arbitrary positive integer, the gamma resp. univariate Wishart distribution is **infinitely divisible**.

The above includes the following two specific constructions:

a) If $x_1, \ldots, x_n \sim \text{Exp}(\theta)$ are independent samples from $\text{Exp}(\theta)$ then the sum $y = \sum_{i=1}^{n} x_i \sim \text{Gam}(\alpha = n, \theta)$ is gamma distributed with the same scale parameter.

b) The sum of k independent scaled chi-squared random variables $s^2\chi^2_1$ with one degree of freedom and identical scale parameter s^2 yields a scaled chi-squared random variable $s^2\chi^2_k$ with degree of freedom k and the same scale parameter. Thus, if $z_1, z_2, \ldots, z_k \sim N(0, 1)$ are k independent samples from $N(0, 1)$ then $\sum_{i=1}^{k} z_i^2 \sim \chi^2_k$.

4.5 Inverse gamma distribution

The **inverse gamma distribution** $\text{IG}(\alpha, \beta)$ is a continuous distribution and is also known as **univariate inverse Wishart distribution** $\text{IW}(\psi, k)$ using a different parametrisation. It is linked to the gamma distribution $\text{Gam}(\alpha, \theta)$ aka univariate Wishart distribution $\text{Wis}\left(s^2, k\right)$ (Section 4.4).

Special cases include the **inverse chi-squared distribution** $\text{Inv-}\chi^2_k$ and the **scaled inverse chi-squared distribution** $s^2\text{Inv-}\chi^2_k$.

Standard parametrisation

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

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

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

The mean of the inverse gamma distribution is (for $\alpha > 1$)

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

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

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

The inverse gamma distribution $\text{IG}(\alpha, \beta)$ has pdf

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

The inverse gamma distribution and the gamma distribution are directly linked. If $x \sim \text{IG}(\alpha, \beta)$ then the inverse of x is gamma distributed with inverted scale parameter

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

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

R code

The `extraDistr` package implements the inverse gamma distribution. The function `extraDistr::dinvgamma()` provides the pdf, `extraDistr::pinvgamma()` the distribution function and `extraDistr::qinvgamma()` is the quantile function. The corresponding random number generator is `extraDistr::rinvgamma()`.

Wishart parametrisation

The inverse gamma distribution is frequently used with a different set of parameters $\psi = 2\beta$ (scale parameter) and $k = 2\alpha$ (shape parameter). In this form it is called **univariate inverse Wishart distribution**

$$x \sim \text{IW}(\psi, k) = \text{IG}\left(\alpha = \frac{k}{2}, \beta = \frac{\psi}{2}\right)$$

In the above the scale parameter ψ is scalar and hence the resulting inverse Wishart distribution is univariate. If instead a matrix-valued scale parameter $\mathbf{\Psi}$ is used this yields the multivariate or d-dimensional inverse Wishart distribution, see Section 5.5.

In the Wishart parametrisation the mean is (for $k > 2$)

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

and the variance is (for $k > 4$)

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

The pdf in terms of ψ and k is

$$p(x|\psi, k) = \frac{(\psi/2)^{(k/2)}}{\Gamma(k/2)}\, x^{-(k+2)/2}\, e^{-\psi x^{-1}/2}$$

The univariate inverse Wishart and the univariate Wishart distributions are linked. If $x \sim \text{IW}(\psi, k)$ then the inverse of x is Wishart distributed with inverted scale parameter:

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

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

Mean parametrisation

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

$$x \sim \text{IW}(\psi = \kappa\mu, k = \kappa + 2) = \text{IG}\left(\alpha = \frac{\kappa+2}{2}, \beta = \frac{\mu\kappa}{2}\right)$$

has mean (for $\kappa > 0$)

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

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

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

The **mean parametrisation** is useful in Bayesian analysis when employing the inverse gamma aka univariate inverse Wishart distribution as prior and posterior distribution.

Biased mean parametrisation

Using $\tau^2 = \frac{\psi}{k}$ as biased mean parameter together with $\nu = k$ we arrive at the **biased mean parametrisation**

$$x \sim \text{IW}(\psi = \nu\tau^2, k = \nu) = \text{IG}\left(\alpha = \frac{\nu}{2}, \beta = \frac{\nu\tau^2}{2}\right)$$

with mean (for $\nu > 2$)

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

and variance ($\nu > 4$)

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

As $\tau^2 = \mu(\nu - 2)/\nu$ for large ν the parameter τ^2 will become identical to the true mean μ.

This parametrisation is useful to derive the location-scale t-distribution with its matching parameters (see Section 4.6). It is also common in Bayesian analysis.

Special case: inverse chi-squared distribution

If the scale parameter in $\mathrm{IW}(\psi, k)$ is set to $\psi = 1$ and k is restricted to the positive integers $\{1, 2, \ldots\}$ then the univariate inverse Wishart distribution reduces to the **inverse chi-squared distribution**

$$x \sim \text{Inv-}\chi^2_k = \mathrm{IW}(\psi = 1, k) = \mathrm{IG}\left(\alpha = \frac{k}{2}, \beta = \frac{1}{2}\right)$$

where k is called the degree of freedom.

The inverse chi-squared distribution has mean (for $k > 2$)

$$\mathrm{E}(x) = \frac{1}{k - 2}$$

and the variance is (for $k > 4$)

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

The inverse chi-squared distribution and the chi-squared distribution are linked. If $x \sim \text{Inv-}\chi^2_k$ then the inverse of x is chi-squared distributed:

$$\frac{1}{x} \sim \chi^2_k$$

where k is the degree of freedom.

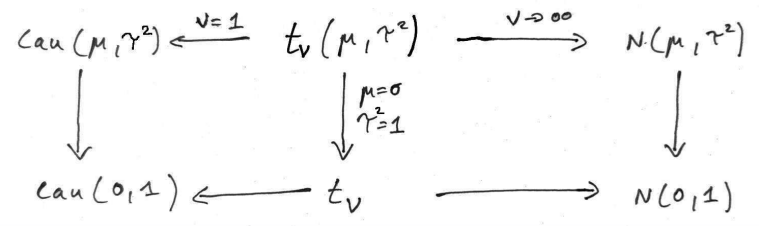

Figure 4.7: The location-scale t-distribution and its relatives.

Scale transformation

If $x \sim \mathrm{IG}(\alpha, \beta)$ then the scaled random variable bx with $b > 0$ is also inverse gamma distributed with $bx \sim \mathrm{IG}(\alpha, b\beta)$.

Hence,

- $\beta\,\mathrm{IG}(\alpha, 1) = \mathrm{IG}(\alpha, \beta)$,
- $\psi\,\mathrm{IW}(1, k) = \mathrm{IW}(\psi, k)$,
- $\kappa\mu\,\mathrm{IW}(1, k = \kappa + 2) = \mathrm{IW}(\psi = \kappa\mu, k = \kappa + 2)$ and
- $\nu\tau^2\,\mathrm{IW}(1, k = \nu) = \mathrm{IW}(\psi = \nu\tau^2, k = \nu)$

As $\text{Inv-}\chi^2_k$ equals $\mathrm{IW}(\psi = 1, k)$ the **scaled inverse chi-squared distribution** $\psi\text{Inv-}\chi^2_k$ is thus equivalent to the univariate inverse Wishart distribution $\mathrm{IW}(\psi, k)$. If a random variable follows the scaled inverse chi-squared distribution its inverse follows the corresponding scaled chi-squared distribution. Specifically, if $x \sim \psi\,\text{Inv-}\chi^2_k$ then $1/x \sim \psi^{-1}\,\chi^2_k$.

The scaled inverse chi-squared distribution is frequently used in the biased mean parametrisation with $\tau = \psi/\nu$ and $\nu = k$. Then $\psi\,\text{Inv-}\chi^2_k$ is equal to $\nu\tau^2\,\text{Inv-}\chi^2_\nu = \mathrm{IW}(\psi = \nu\tau^2, k = \nu)$ which is sometimes also written as $\text{Inv-}\chi^2(\nu, \tau^2)$. If $x \sim \nu\tau^2\,\text{Inv-}\chi^2_\nu$ then $1/x \sim 1/(\nu\tau^2)\,\chi^2_\nu$.

4.6 Location-scale t-distribution

The **location-scale t-distribution** $t_\nu(\mu, \tau^2)$ is a continuous distribution and is a generalisation of the normal distribution $N(\mu, \tau^2)$ (Section 4.3) with an additional parameter $\nu > 0$ (degrees of freedom) controlling the probability mass in the tails.

Special cases include the **Student's t-distribution** t_ν, the **normal distribution** $N(\mu, \tau^2)$ and the **Cauchy distribution** $\mathrm{Cau}(\mu, \tau^2)$. Figure 4.7

illustrates the relationship of the location-scale t-distribution $t_\nu(\mu, \tau^2)$ with these related distributions.

Standard parametrisation

If a random variable $x \in \mathbb{R}$ follows the **location-scale t-distribution** we write

$$x \sim t_\nu(\mu, \tau^2)$$

where μ is the location and τ^2 the dispersion parameter. The parameter $\nu > 0$ prescribes the degrees of freedom. For small values of ν the distribution is heavy-tailed and as a result only moments of order smaller than ν are finite and defined.

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

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

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

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

The pdf of $t_\nu(\mu, \tau^2)$ is

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

with $\Delta^2 = (x - \mu)^2/\tau^2$ the squared Mahalanobis distance between x and μ.

R code

The package `extraDistr` implements the location-scale t-distribution. The function `extraDistr::dlst()` returns the pdf, `extraDistr::plst()` the distribution function function and `extraDistr::qlst()` is the quantile function. The corresponding random number generator is `extraDistr::rlst()`.

Scale parametrisation

Instead of the dispersion parameter τ^2 it is often also convenient to use the scale parameter $\tau = \sqrt{\tau^2} > 0$. Similarly, instead of the inverse dispersion $1/\tau^2$ one may wish to use the inverse scale $w = 1/\tau$.

The scale parametrisation is central for location-scale transformations (see below).

Special case: Student's t-distribution

With $\mu = 0$ and $\tau^2 = 1$ the location-scale t-distribution reduces to the **standard t-distribution** $t_\nu = t_\nu(0, 1)$. It is commonly known **Student's t-distribution** named after "Student" which was the pseudonym of William Sealy Gosset (1876–1937). It is a generalisation of the standard normal distribution $N(0, 1)$ to allow for heavy tails.

The distribution has mean $\mathrm{E}(x) = 0$ (for $\nu > 1$) and variance $\mathrm{Var}(x) = \frac{\nu}{\nu-2}$ (for $\nu > 2$).

The pdf of t_ν is

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

with the squared Mahalanobis distance reducing to $\Delta^2 = x^2$.

R code

The command `dt()` returns the pdf of the t-distribution, `pt()` is distribution function and `qt()` the quantile function. The corresponding random number generator is `rt()`.

Special case: normal distribution

For $\nu \to \infty$ the location-scale t-distribution $t_\nu(\mu, \tau^2)$ reduces to the **normal distribution** $N(\mu, \tau^2)$ (Section 4.3). Correspondingly, for $\nu \to \infty$ the Student's t-distribution becomes equal to the **standard normal distribution** $N(0, 1)$.

See Section 5.6 for further details.

Special case: Cauchy distribution

For $\nu = 1$ the location-scale t-distribution becomes the **Cauchy distribution** $\mathrm{Cau}(\mu, \tau^2) = t_1(\mu, \tau^2)$ named after Augustin-Louis Cauchy (1789–1857).

Its mean, variance and other higher moments are all undefined.

It has pdf

$$\begin{aligned} p(x|\mu,\tau^2) &= (\tau^2)^{-1/2}(\pi(1+\Delta^2))^{-1} \\ &= \frac{\tau}{\pi(\tau^2+(x-\mu)^2)} \end{aligned}$$

with $\tau = \sqrt{\tau^2} > 0$.

Note that in the above we employ τ^2 as dispersion parameter as this parallels the location-scale t-distribution and the normal distribution but very often the Cauchy distribution is used with $\tau > 0$ as scale parameter.

> R code
>
> The command `dcauchy()` returns the pdf of the Cauchy distribution, `pcauchy()` is the distribution function and `qcauchy()` the quantile function. The corresponding random number generator is `rcauchy()`.

Special case: standard Cauchy distribution

The **standard Cauchy distribution** $\text{Cau}(0,1) = t_1(0,1) = t_1$ is obtained by setting $\mu = 0$ and $\tau^2 = 1$ (Cauchy distribution) or, equivalently, by setting $\nu = 1$ (Student's t-distribution).

It has pdf

$$p(x) = \frac{1}{\pi(1+x^2)}$$

Location-scale transformation

Let $\tau > 0$ be the positive square root of τ^2 and $w = 1/\tau$.

If $x \sim t_\nu(\mu,\tau^2)$ then $y = w(x-\mu) \sim t_\nu$. This location-scale transformation reduces a location-scale t-distributed random variable to a Student's t-distributed random variable.

Conversely, if $y \sim t_\nu$ then $x = \mu + \tau y \sim t_\nu(\mu,\tau^2)$. This location-scale transformation generates the location-scale t-distribution from the Student's t -distribution.

For the special case of the Cauchy distribution (corresponding to $\nu = 1$) similar relations hold between it and the standard Cauchy distribution. If $x \sim \text{Cau}(\mu,\tau^2)$ then $y = w(x-\mu) \sim \text{Cau}(0,1)$. Conversely, if $y \sim \text{Cau}(0,1)$ then $x = \mu + \tau y \sim \text{Cau}(\mu,\tau^2)$.

Convolution property

The location-scale t-distribution is not generally closed under convolution, with the exception of two special cases, the normal distribution ($\nu = \infty$), see Section 4.3, and the Cauchy distribution ($\nu = 1$).

For the Cauchy distribution with $\tau_i^2 = a_i^2\tau^2$, where $a_i > 0$ are positive scalars,

$$\sum_{i=1}^n \text{Cau}(\mu_i, a_i^2\tau^2) \sim \text{Cau}\left(\sum_{i=1}^n \mu_i, \left(\sum_{i=1}^n a_i\right)^2 \tau^2\right)$$

Location-scale t-distribution as compound distribution

The location-scale t-distribution can be obtained as mixture of normal distributions with identical mean and varying variance. Specifically, let z be a univariate inverse Wishart random variable

$$z \sim \text{IW}(\psi = \nu, k = \nu) = \text{IG}\left(\alpha = \frac{\nu}{2}, \beta = \frac{\nu}{2}\right)$$

and let $x|z$ be normal

$$x|z \sim N(\mu, \sigma^2 = z\tau^2)$$

Then the resulting marginal (scale mixture) distribution for x is the location-scale t-distribution

$$x \sim t_\nu\left(\mu,\tau^2\right)$$

An alternative way to arrive at $t_\nu\left(\mu,\tau^2\right)$ is to include τ^2 as parameter in the inverse Wishart distribution

$$z \sim \tau^2\text{IW}(\psi = \nu, k = \nu) = \text{IW}(\psi = \nu\tau^2, k = \nu)$$

and let

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

Note that τ^2 is now the biased mean parameter of the univariate inverse Wishart distribution. This characterisation is useful in Bayesian analysis.

5 Multivariate distributions

5.1 Multinomial distribution

The **multinomial distribution** Mult$(n, \boldsymbol{\theta})$ is the multivariate generalisation of the binomial distribution Bin(n, θ) (Section 4.1) from two classes to K classes.

A special case is the **categorical distribution** Cat$(\boldsymbol{\theta})$ that generalises the Bernoulli distribution Ber(θ).

Standard parametrisation

A multinomial random variable $\boldsymbol{x}$ describes describes the allocation of n items to K classes. We write

$$\boldsymbol{x} \sim \text{Mult}(n, \boldsymbol{\theta})$$

where the parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^T$ specifies the probability of each of K classes, with $\theta_i \in [0, 1]$ and $\boldsymbol{\theta}^T \mathbf{1}_K = \sum_{i=1}^K \theta_i = 1$. Thus there are $K - 1$ independent elements in $\boldsymbol{\theta}$. The number of classes K is implicitly given by the dimension of the vector $\boldsymbol{\theta}$. Each element of the vector $\boldsymbol{x} = (x_1, \ldots, x_K)^T$ is an integer $x_i \in \{0, 1, \ldots, n\}$ and $\boldsymbol{x}$ satisfies the constraint $\boldsymbol{x}^T \mathbf{1}_K = \sum_{i=1}^K x_i = n$. Therefore the support of $\boldsymbol{x}$ is a $K - 1$ dimensional space and it notably depends on n.

The multinomial distribution is best illustrated by an urn model distributing n items into K bins where $\boldsymbol{\theta}$ contains the corresponding bin probabilities (Figure 5.1).

The expected value is

$$\text{E}(\boldsymbol{x}) = n\boldsymbol{\theta}$$

The covariance matrix is

$$\text{Var}(\boldsymbol{x}) = n(\text{Diag}(\boldsymbol{\theta}) - \boldsymbol{\theta}\boldsymbol{\theta}^T)$$

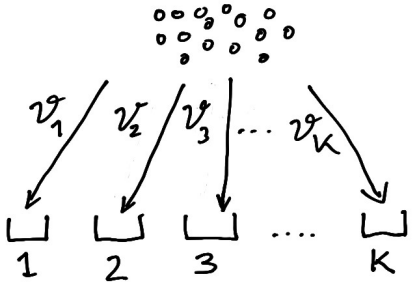

Figure 5.1: Multinomial urn model.

The covariance matrix is singular by construction because of the dependencies among the elements of $\boldsymbol{x}$.

The corresponding pmf is

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \binom{n}{x_1, \ldots, x_n} \prod_{i=1}^K \theta_i^{x_i}$$

where $\binom{n}{x_1, \ldots, x_n}$ is the multinomial coefficient. While all K elements of $\boldsymbol{x}$ appear in the pmf recall that due the dependencies among the x_i the pmf is defined over a $K - 1$ dimensional support.

For $K = 2$ the multinomial distribution reduces to the binomial distribution (Section 4.1).

> R code
>
> The pmf of the multinomial distribution is given by `dmultinom()`. The corresponding random number generator is `rmultinom()`.

Mean parametrisation

Instead of $\boldsymbol{\theta}$ one may also use a mean parameter $\boldsymbol{\mu}$, with elements $\mu_i \in [0, n]$ and $\boldsymbol{\mu}^T \mathbf{1}_K = \sum_{i=1}^K \mu_i = n$, so that

$$\boldsymbol{x} \sim \text{Mult}\left(n, \boldsymbol{\theta} = \frac{\boldsymbol{\mu}}{n}\right)$$

The mean parameter $\boldsymbol{\mu}$ can be obtained from $\boldsymbol{\theta}$ and n by $\boldsymbol{\mu} = n\boldsymbol{\theta}$. Note that the parameter space for $\boldsymbol{\mu}$ and the support of $\boldsymbol{x}$ are both of dimension $K - 1$.

The mean and variance of the multinomial distribution expressed in terms of $\boldsymbol{\mu}$ and n are

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

and

$$\mathrm{Var}(x) = \mathrm{Diag}(\boldsymbol{\mu}) - \frac{\boldsymbol{\mu}\boldsymbol{\mu}^T}{n}$$

The covariance matrix is singular by construction because of the dependencies among the elements of $\boldsymbol{x}$.

Special case: categorical distribution

For $n = 1$ the multinomial distribution reduces to the **categorical distribution** $\mathrm{Cat}(\boldsymbol{\theta})$ which in turn is the multivariate generalisation of the Bernoulli distribution $\mathrm{Ber}(\theta)$ from two classes to K classes.

If a random variable $\boldsymbol{x}$ follows the categorical distribution we write

$$\boldsymbol{x} \sim \mathrm{Cat}(\boldsymbol{\theta})$$

with class probabilities $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^T\mathbf{1}_K = 1$. The support is $x_i \in \{0, 1\}$ and $\boldsymbol{x}^T\mathbf{1}_K = 1$ and is a $K - 1$ dimensional space.

The random vector $\boldsymbol{x}$ takes the form of an indicator vector $\boldsymbol{x} = (x_1, \ldots, x_K)^T = (0, 0, \ldots, 1, \ldots, 0)^T$ containing zeros everywhere except for a single element $x_k = 1$ indicating the class k to which the item has been allocated. This is called "one hot encoding", as opposed to "integer encoding", i.e. stating the class number k.

The expected value is

$$\mathrm{E}(\boldsymbol{x}) = \boldsymbol{\theta}$$

The covariance matrix is

$$\mathrm{Var}(\boldsymbol{x}) = \mathrm{Diag}(\boldsymbol{\theta}) - \boldsymbol{\theta}\boldsymbol{\theta}^T$$

The covariance matrix is singular by construction because of the dependencies among the elements of $\boldsymbol{x}$. This 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 $i \neq j$.

The corresponding pmf is

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

Recall that the pmf is defined over the $K - 1$ dimensional support of $\boldsymbol{x}$.

For $K = 2$ the categorical distribution reduces to the Bernoulli $\mathrm{Ber}(\theta)$ distribution, with $\theta_1 = \theta$, $\theta_2 = 1 - \theta$ and $x_1 = x$ and $x_2 = 1 - x$.

Convolution property

The convolution of n multinomial distributions, each with identical bin probabilities $\boldsymbol{\theta}$ but possibly different number of items n_i, yields another multinomial distribution with the same parameter $\boldsymbol{\theta}$:

$$\sum_{i=1}^{n} \mathrm{Mult}(n_i, \boldsymbol{\theta}) \sim \mathrm{Mult}\left(\sum_{i=1}^{n} n_i, \boldsymbol{\theta}\right)$$

It follows that the multinomial distribution with n items is the result of the convolution of n categorical distributions:

$$\sum_{i=1}^{n} \mathrm{Cat}(\boldsymbol{\theta}) \sim \mathrm{Mult}(n, \boldsymbol{\theta})$$

Thus, repeating the same categorical trial n times and counting the total number of allocations in each bin yields a multinomial random variable.

5.2 Dirichlet distribution

The **Dirichlet distribution** $\mathrm{Dir}(\boldsymbol{\alpha})$ is the multivariate generalisation of the beta distribution $\mathrm{Beta}(\alpha_1, \alpha_2)$ (Section 4.2) that is useful to model proportions or probabilities for $K \geq 2$ classes. It is named after Peter Gustav Lejeune Dirichlet (1805–1859).

It includes the **uniform distribution** over the $K - 1$ unit simplex as special case.

Standard parametrisation

A Dirichlet distributed random vector is denoted by

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

with shape parameter $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_K)^T > 0$ and $K \geq 2$. Let $m = \boldsymbol{\alpha}^T \mathbf{1}_K = \sum_{i=1}^K \alpha_i$. The support of $\boldsymbol{x}$ is the $K-1$ dimensional unit simplex given by $x_i \in [0, 1]$ and $\boldsymbol{x}^T \mathbf{1}_K = \sum_{i=1}^K x_i = 1$. Thus, the Dirichlet distribution is defined over a $K-1$ dimensional space.

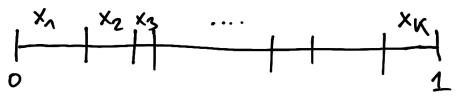

Figure 5.2: Stick breaking visualisation of a Dirichlet random variable.

A Dirichlet random variable can be visualised as breaking a unit stick into K individual pieces of lengths $x_1, x_2, \ldots, x_K$ adding up to one (Figure 5.2). Thus, the x_i may be used as the exclusive proportions or probabilities for K classes.

The mean is

$$\text{E}(\boldsymbol{x}) = \frac{\boldsymbol{\alpha}}{m}$$

and the variance is

$$\text{Var}(\boldsymbol{x}) = \frac{m\text{Diag}(\boldsymbol{\alpha}) - \boldsymbol{\alpha}\boldsymbol{\alpha}^T}{m^2(m+1)}$$

The covariance matrix is singular by construction because of the dependencies among the elements of $\boldsymbol{x}$. In component notation it is

$$\text{Cov}(x_i, x_j) = \frac{[i=j]m\alpha_i - \alpha_i\alpha_j}{m^2(m+1)}$$

where the indicator function $[i = j]$ equals 1 if $i = j$ and 0 otherwise.

The pdf 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 $\boldsymbol{\alpha}$ defined as

$$B(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma(m)}$$

While all K elements of $\boldsymbol{x}$ appear in the pdf recall that due the dependencies among the x_i the pdf is defined over a $K-1$ dimensional support.

For $K = 2$ the Dirichlet distribution reduces to the beta distribution (Section 4.2).

> **R code**
>
> The `extraDistr` package implements the Dirichlet distribution. The pmf of the Dirichlet distribution is given by `extraDistr::ddirichlet()`. The corresponding random number generator is `extraDistr::rdirichlet()`.

Mean parametrisation

Instead of employing $\boldsymbol{\alpha}$ as parameter vector another useful reparametrisation of the Dirichlet distribution is in terms of a mean parameter $\boldsymbol{\mu}$, with elements $\mu_i \in [0, 1]$ and $\boldsymbol{\mu}^T \mathbf{1}_K = \sum_{i=1}^K \mu_i = 1$, and a concentration parameter $m > 0$ so that

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

The concentration and mean parameters can be obtained from $\boldsymbol{\alpha}$ by $m = \boldsymbol{\alpha}^T \mathbf{1}_K$ and $\boldsymbol{\mu} = \boldsymbol{\alpha}/m$. The space of possible values for the mean parameter $\boldsymbol{\mu}$ and the support of $\boldsymbol{x}$ are both of dimension $K-1$.

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

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

and

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

The covariance matrix is singular by construction because of the dependencies among the elements of $\boldsymbol{x}$. In component notation it is

$$\begin{aligned}\mathrm{Cov}(x_i, x_j) &= \frac{[i=j]\mu_i - \mu_i\mu_j}{m+1} \\ &= \begin{cases}\mu_i(1-\mu_i)/(m+1) & \text{if } i = j \\ -\mu_i\mu_j/(m+1) & \text{if } i \neq j\end{cases}\end{aligned}$$

Special case: symmetric Dirichlet distribution

For $\boldsymbol{\alpha} = \alpha\mathbf{1}_K$ the Dirichlet distribution becomes the **symmetric beta distribution** with a single shape parameters $\alpha > 0$. In mean parametrisation the symmetric Dirichlet distribution corresponds to $\boldsymbol{\mu} = \mathbf{1}_K/K$ and $m = \alpha K$.

Special case: uniform distribution

For $\boldsymbol{\alpha} = \mathbf{1}_K$ the Dirichlet distribution becomes the uniform distribution over the $K-1$ unit simplex with pdf $p(\boldsymbol{x}) = 1/\Gamma(K)$. In mean parametrisation the symmetric Dirichlet distribution corresponds to $\boldsymbol{\mu} = \mathbf{1}_K/K$ and $m = K$.

5.3 Multivariate normal distribution

The **multivariate normal distribution** $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ generalises the univariate normal distribution $N(\mu, \sigma^2)$ (Section 4.3) from one to d dimensions.

Special cases are the **multivariate standard normal distribution** $N(\mathbf{0}, \boldsymbol{I})$ and the **multivariate delta distribution** δ.

Standard parametrisation

The multivariate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ has a mean or location parameter $\boldsymbol{\mu}$ (a d dimensional vector), a variance parameter $\boldsymbol{\Sigma}$ (a $d \times d$ positive definite symmetric matrix) and support $\boldsymbol{x} \in \mathbb{R}^d$.

If a random vector $\boldsymbol{x} = (x_1, x_2, ..., x_d)^T$ follows a multivariate normal distribution we write

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

with mean

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

and variance

$$\mathrm{Var}(\boldsymbol{x}) = \boldsymbol{\Sigma}$$

In the above notation the dimension d is implicitly given by the dimensions of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ but for clarity one often also writes $N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to explicitly indicate the dimension.

The pdf is given by

$$\begin{aligned}p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \det(2\pi\boldsymbol{\Sigma})^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \\ &= \det(\boldsymbol{\Sigma})^{-1/2}(2\pi)^{-d/2}e^{-\Delta^2/2}\end{aligned}$$

Here $\Delta^2 = (\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ is the **squared Mahalanobis distance** between $\boldsymbol{x}$ and $\boldsymbol{\mu}$ taking into account the variance $\boldsymbol{\Sigma}$. Note that this pdf is a joint pdf over the d elements $x_1, \ldots, x_d$ of the random vector $\boldsymbol{x}$.

The multivariate normal distribution is sometimes also used by specifying the precision matrix $\boldsymbol{\Sigma}^{-1}$ instead of the variance $\boldsymbol{\Sigma}$.

For $d = 1$ the random vector $\boldsymbol{x} = x$ is a scalar, $\boldsymbol{\mu} = \mu$, $\boldsymbol{\Sigma} = \sigma^2$ and thus the multivariate normal distribution reduces to the univariate normal distribution (Section 4.3).

R code

The `mnormt` package implements the multivariate normal distribution. The function `mnormt::dmnorm()` provides the pdf and `mnormt::pmnorm()` returns the distribution function. The function `mnormt::rmnorm()` is the corresponding random number generator.
The `mniw` package also implements the multivariate normal distribution. The pdf of the Wishart distribution is given by `mniw::dmNorm()`. The corresponding random number generator is `mniw::rmNorm()`.

Scale parametrisation

In the univariate case it is straightforward to use the standard deviation σ as scale parameter instead of the variance σ^2, and similarly the inverse standard deviation $w = 1/\sigma$ instead of the precision σ^{-2}. However, in

the multivariate setting with a matrix variance parameter $\boldsymbol{\Sigma}$ it is less obvious how to define a suitable matrix scale parameter.

Let $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$ be the eigendecomposition of the positive definite matrix $\boldsymbol{\Sigma}$. Then $\boldsymbol{\Sigma}^{1/2} = \boldsymbol{U}\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}^T$ is the principal matrix square root and $\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{U}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}^T$ the inverse principal matrix square root. Furthermore, let $\boldsymbol{Q}$ be an arbitrary orthogonal matrix with $\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{Q}^T = \boldsymbol{I}$.

Then $\boldsymbol{W} = \boldsymbol{Q}\boldsymbol{\Sigma}^{-1/2}$ is called a **whitening matrix** based on $\boldsymbol{\Sigma}$ and $\boldsymbol{L} = \boldsymbol{W}^{-1} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{Q}^T$ is the corresponding **inverse whitening matrix**. By construction, the matrix $\boldsymbol{L}$ provides a factorisation of the covariance matrix by $\boldsymbol{L}\boldsymbol{L}^T = \boldsymbol{\Sigma}$. Similarly, $\boldsymbol{W}$ factorises the precision matrix by $\boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{\Sigma}^{-1}$. The two matrices thus provide the basis for the scale parametrisation of the multivariate normal distribution.

Specifically, the matrix $\boldsymbol{L}$ is used in place of $\boldsymbol{\Sigma}$ and plays the role of the matrix scale parameter (corresponding to σ in the univariate setting) and $\boldsymbol{W}$ is used in place of the precision matrix $\boldsymbol{\Sigma}^{-1}$ and plays the role of the inverse matrix scale parameter (corresponding to $1/\sigma$ in the univariate case). The determinants occurring in the multivariate normal pdf can be rewritten in terms of $\boldsymbol{L}$ and $\boldsymbol{W}$ using the identities $|\det(\boldsymbol{W})| = \det(\boldsymbol{\Sigma})^{-1/2}$ and $|\det(\boldsymbol{L})| = \det(\boldsymbol{\Sigma})^{1/2}$ as $\det(\boldsymbol{Q}) = \pm 1$.

Since $\boldsymbol{Q}$ can be freely chosen the matrices $\boldsymbol{W}$ and $\boldsymbol{L}$ are not fully determined by $\boldsymbol{\Sigma}$ alone but there is rotational freedom due to $\boldsymbol{Q}$. Standard choices are

- $\boldsymbol{Q}^{\text{ZCA}} = \boldsymbol{I}$ for ZCA-type factorisation with $\boldsymbol{W}^{\text{ZCA}} = \boldsymbol{\Sigma}^{-1/2}$ and
- $\boldsymbol{Q}^{\text{PCA}} = \boldsymbol{U}^T$ for PCA-type factorisation with $\boldsymbol{W}^{\text{PCA}} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}^T$. Note that the matrix $\boldsymbol{U}$ is not unique because its columns (eigenvectors) can have different signs (directions), hence $\boldsymbol{W}^{\text{PCA}}$ and $\boldsymbol{L}^{\text{PCA}}$ are also not unique without further constraints, such as positive diagonal elements of the (inverse) whitening matrix.
- A third common choice is to compute $\boldsymbol{L}$ directly by Cholesky decomposition of $\boldsymbol{\Sigma}$, which yields an $\boldsymbol{L}^{\text{Chol}}$ (and also a $\boldsymbol{W}^{\text{Chol}}$) in the form of a lower-triangular matrix with a positive diagonal, and a corresponding underlying $\boldsymbol{Q}^{\text{Chol}} = (\boldsymbol{L}^{\text{Chol}})^T\boldsymbol{\Sigma}^{-1/2}$.

Finally, the whitening matrix $\boldsymbol{W}$ and its inverse may also be constructed from the correlation matrix $\boldsymbol{P}$ and the diagonal matrix containing the variances $\boldsymbol{V}$ (with $\boldsymbol{\Sigma} = \boldsymbol{V}^{1/2}\boldsymbol{P}\boldsymbol{V}^{1/2}$) in the form $\boldsymbol{W} = \boldsymbol{Q}\boldsymbol{P}^{-1/2}\boldsymbol{V}^{-1/2}$ and $\boldsymbol{L} = \boldsymbol{V}^{1/2}\boldsymbol{P}^{1/2}\boldsymbol{Q}^T$.

Special case: multivariate standard normal distribution

The **multivariate standard normal distribution** $N(\boldsymbol{0}, \boldsymbol{I})$ has mean $\boldsymbol{\mu} = \boldsymbol{0}$ and variance $\boldsymbol{\Sigma} = \boldsymbol{I}$. The corresponding pdf is

$$p(\boldsymbol{x}) = (2\pi)^{-d/2} e^{-\boldsymbol{x}^T\boldsymbol{x}/2}$$

with the squared Mahalanobis distance reduced to $\Delta^2 = \boldsymbol{x}^T\boldsymbol{x} = \sum_{i=1}^{d} x_i^2$.

The density of the multivariate standard normal distribution is the product of the corresponding univariate standard normal densities

$$p(\boldsymbol{x}) = \prod_{i=1}^{d} (2\pi)^{-1/2} e^{-x_i^2/2}$$

and therefore the elements x_i of $\boldsymbol{x} = (x_1, \ldots, x_d)^T$ are independent of each other.

Special case: multivariate delta distribution

The **multivariate delta distribution** δ is obtained as the limit of $N(\boldsymbol{0}, \varepsilon\boldsymbol{A})$ for $\varepsilon \to 0$ and where $\boldsymbol{A}$ is a positive definite matrix (e.g. $\boldsymbol{A} = \boldsymbol{I}$). Thus δ is a distribution that behaves like an infinite spike at zero.

The corresponding pdf $\delta(\boldsymbol{x})$ is called the **multivariate Dirac delta function**, even though it is not an ordinary function. It satisfies $\delta(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \neq \boldsymbol{0}$ and integrates to one, thus representing a point mass at zero.

Location-scale transformation

Let $\boldsymbol{W}$ be a whitening matrix for $\boldsymbol{\Sigma}$ and $\boldsymbol{L}$ the corresponding inverse whitening matrix.

If $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then $\boldsymbol{y} = \boldsymbol{W}(\boldsymbol{x} - \boldsymbol{\mu}) \sim N(\boldsymbol{0}, \boldsymbol{I})$. This location-scale transformation corresponds to centring and whitening (i.e. standardisation and decorrelation) of a multivariate normal random variable.

Conversely, if $\boldsymbol{y} \sim N(\boldsymbol{0}, \boldsymbol{I})$ then $\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. This location-scale transformation generates the multivariate normal distribution from the multivariate standard normal distribution.

Note that under the location-scale transformation $\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{y}$ with $\text{Var}(\boldsymbol{y}) = \boldsymbol{I}$ we get $\text{Cov}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{L}$. This provides a means to choose

between different (inverse) whitening transformation and the corresponding factorisations of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^{-1}$. For example, if positive correlation between corresponding elements in $\boldsymbol{x}$ and $\boldsymbol{y}$ is desired then the diagonal elements in $\boldsymbol{L}$ must be positive.

Convolution property

The convolution of n independent, but not necessarily identical, multivariate normal distributions of the same dimension d results in another d-dimensional multivariate normal distribution with corresponding mean and variance:

$$\sum_{i=1}^{n} N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \sim N\left(\sum_{i=1}^{n} \boldsymbol{\mu}_i, \sum_{i=1}^{n} \boldsymbol{\Sigma}_i\right)$$

Hence, any multivariate normal random variable can be constructed as the sum of n suitable independent multivariate normal random variables.

Since n is an arbitrary positive integer the multivariate normal distribution is said to be **infinitely divisible**.

5.4 Wishart distribution

The Wishart distribution $\text{Wis}(\boldsymbol{S}, k)$ is a multivariate generalisation of the gamma distribution $\text{Gam}(\alpha, \theta)$ (Section 4.4) from one to d dimensions.

Standard parametrisation

If the symmetric random matrix $\boldsymbol{X}$ of dimension $d \times d$ is Wishart distributed we write

$$\boldsymbol{X} \sim \text{Wis}(\boldsymbol{S}, k)$$

where $\boldsymbol{S} = (s_{ij})$ is the scale parameter (a symmetric $d \times d$ positive definite matrix with elements s_{ij}). The dimension d is implicit in the scale parameter $\boldsymbol{S}$.

The shape parameter k takes on real values in the range $k > d - 1$ and integer values in the range $k \in 1, \ldots, d-1$ for $d > 1$. For $k > d - 1$ the matrix $\boldsymbol{X}$ is positive definite and invertible (see also Section 5.5), otherwise $\boldsymbol{X}$ is singular and positive semi-definite.

The distribution has mean

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

and variances of the elements of $\boldsymbol{X}$ are

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

The pdf is (for $k > d - 1$)

$$p(\boldsymbol{X}|\boldsymbol{S}, k) = \frac{1}{\Gamma_d(k/2)\det(2\boldsymbol{S})^{k/2}} \det(\boldsymbol{X})^{(k-d-1)/2} \exp\left(-\text{Tr}(\boldsymbol{S}^{-1}\boldsymbol{X})/2\right)$$

with the multivariate gamma function defined as

$$\Gamma_d(k/2) = \pi^{d(d-1)/4} \prod_{j=1}^{d} \Gamma((k - j + 1)/2)$$

Note that this pdf is a joint pdf over the d diagonal elements x_{ii} and the $d(d-1)/2$ off-diagonal elements x_{ij} of the symmetric random matrix $\boldsymbol{X}$.

If $\boldsymbol{S}$ is a scalar rather than a matrix (and hence $d = 1$) then the multivariate Wishart distribution reduces to the univariate Wishart aka gamma distribution (Section 4.4).

The Wishart distribution is closely related to the multivariate normal distribution with mean zero. Specifically, if $\boldsymbol{z} \sim N(\boldsymbol{0}, \boldsymbol{S})$ then $\boldsymbol{z}\boldsymbol{z}^T \sim \text{Wis}(\boldsymbol{S}, 1)$.

> R code
>
> The `mniw` package implements the Wishart distribution. The pdf of the Wishart distribution is given by `mniw::dwish()`. The corresponding random number generator is `mniw::rwish()`.

Mean parametrisation

It is useful to employ the Wishart distribution in **mean parametrisation**

$$\text{Wis}\left(\boldsymbol{S} = \frac{\boldsymbol{M}}{k}, k\right)$$

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

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

and variances of the elements of $\boldsymbol{X}$ are

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

Special case: standard Wishart distribution

For $\boldsymbol{S} = \boldsymbol{I}$ the Wishart distribution reduces to the **standard Wishart distribution**

$$\boldsymbol{X} \sim \mathrm{Wis}\,(\boldsymbol{I}, k)$$

with a single shape parameter k. The mean is

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

and variances of the elements of $\boldsymbol{X}$ are

$$\mathrm{Var}(x_{ij}) = \begin{cases} 2k & \text{if } i = j \\ k & \text{if } i \neq j \end{cases}$$

The pdf is (for $k > d - 1$)

$$p(\boldsymbol{X}|k) = \frac{1}{\Gamma_d(k/2)2^{dk/2}} \det(\boldsymbol{X})^{(k-d-1)/2} \exp\left(-\mathrm{Tr}(\boldsymbol{X})/2\right)$$

The standard Wishart distribution is closely related to the standard multivariate normal distribution with mean zero. Specifically, if $\boldsymbol{z} \sim N(\boldsymbol{0}, \boldsymbol{I})$ then $\boldsymbol{z}\boldsymbol{z}^T \sim \mathrm{Wis}(\boldsymbol{I}, 1)$.

The **Bartlett decomposition** of the standard multivariate Wishart $\mathrm{Wis}(\boldsymbol{I}, k)$ distribution for any real $k > d - 1$ is obtained by Cholesky factorisation of the random matrix $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{Z}^T$. By construction $\boldsymbol{Z}$ is a lower-triangular matrix with positive diagonal elements z_{ii} and lower off-diagonal elements z_{ij} with $i > j$ and $i, j \in \{1, \ldots, d\}$. The corresponding upper off-diagonal elements are set to zero ($z_{ji} = 0$).

The $d(d+1)/2$ elements of $\boldsymbol{Z}$ are independent and allow to generate a standard Wishart variate as follows:

1) the *squared* diagonal elements follow a univariate standard Wishart distribution $z_{ii}^2 \sim \mathrm{Wis}(1, k - i + 1)$ and
2) the off-diagonal elements follow the univariate standard normal distribution $z_{ij} \sim N(0, 1)$.
3) Then $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{Z}^T \sim \mathrm{Wis}(\boldsymbol{I}, k)$.

Scale transformation

If $\boldsymbol{X} \sim \mathrm{Wis}(\boldsymbol{S}, k)$ then the scaled symmetric random matrix $\boldsymbol{A}\boldsymbol{X}\boldsymbol{A}^T$ is also Wishart distributed with $\boldsymbol{A}\boldsymbol{X}\boldsymbol{A}^T \sim \mathrm{Wis}(\boldsymbol{A}\boldsymbol{S}\boldsymbol{A}^T, k)$ where the matrix $\boldsymbol{A}$ must be full rank and $\boldsymbol{A}\boldsymbol{S}\boldsymbol{A}^T$ remains positive definite. The matrix $\boldsymbol{A}$ may be rectangular, hence the size of $\boldsymbol{A}\boldsymbol{X}\boldsymbol{A}^T$ and $\boldsymbol{A}\boldsymbol{S}\boldsymbol{A}^T$ may be smaller compared to $\boldsymbol{X}$ and $\boldsymbol{S}$.

The transformations between the Wishart distribution and the standard Wishart distribution are two important special cases:

1) With $\boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{S}^{-1}$ and $\boldsymbol{X} \sim \mathrm{Wis}(\boldsymbol{S}, k)$ then $\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{X}\boldsymbol{W}^T \sim \mathrm{Wis}(\boldsymbol{I}, k)$ as $\boldsymbol{W}\boldsymbol{S}\boldsymbol{W}^T = \boldsymbol{I}$. This transformation reduces the Wishart distribution to the standard Wishart distribution.
2) Conversely, with $\boldsymbol{L}\boldsymbol{L}^T = \boldsymbol{S}$ and $\boldsymbol{Y} \sim \mathrm{Wis}(\boldsymbol{I}, k)$ then $\boldsymbol{X} = \boldsymbol{L}\boldsymbol{Y}\boldsymbol{L}^T \sim \mathrm{Wis}(\boldsymbol{S}, k)$ as $\boldsymbol{L}\boldsymbol{I}\boldsymbol{L}^T = \boldsymbol{S}$. This transformation generates the Wishart distribution from the standard Wishart distribution.

Convolution property

The convolution of n Wishart distributions with the same scale parameter $\boldsymbol{S}$ but possible different shape parameters k_i yields another Wishart distribution:

$$\sum_{i=1}^{n} \mathrm{Wis}(\boldsymbol{S}, k_i) \sim \mathrm{Wis}\left(\boldsymbol{S}, \sum_{i=1}^{n} k_i\right)$$

Note that the shape parameter k is restricted to be an integer in the range $1, \ldots, d-1$ for $d > 1$ but is a real number in the range $k > d - 1$. Thus, if the k_i are all valid shape parameters (for dimension d) then $\sum_{i=1}^{n} k_i$ is also a valid shape parameter.

Due the partial restriction of the shape parameter k to integer values the multivariate Wishart distribution is **not infinitely divisible** for $d > 1$.

The above includes the following construction of the multivariate Wishart distribution $\mathrm{Wis}(S, k)$ for integer-valued k. The sum of k independent Wishart random variables $\mathrm{Wis}(\boldsymbol{S}, 1)$ with one degree of freedom and identical scale parameter yields a Wishart random variable $\mathrm{Wis}(\boldsymbol{S}, k)$ with degree of freedom k and the same scale parameter. Thus, if $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_k \sim N(\boldsymbol{0}, \boldsymbol{S})$ are k independent samples from $N(\boldsymbol{0}, \boldsymbol{S})$ then $\sum_{i_1}^{k} \boldsymbol{z}_i \boldsymbol{z}_i^T \sim \mathrm{Wis}(\boldsymbol{S}, k)$.

5.5 Inverse Wishart distribution

The **inverse Wishart distribution** $\text{IW}(\boldsymbol{\Psi}, k)$ is a multivariate generalisation of the inverse gamma distribution $\text{IG}(\alpha, \beta)$ (Section 4.5) from one to d dimensions. It is linked to the Wishart distribution $\text{Wis}(\boldsymbol{S}, k)$ (Section 5.4).

Standard parametrisation

A symmetric positive definite random matrix $\boldsymbol{X}$ of dimension $d \times d$ following an inverse Wishart distribution is denoted by

$$\boldsymbol{X} \sim \text{IW}(\boldsymbol{\Psi}, k)$$

where $\boldsymbol{\Psi} = (\psi_{ij})$ is the scale parameter (a $d \times d$ positive definite symmetric matrix) and $k > d-1$ is the shape parameter. The dimension d is implicit in the scale parameter $\boldsymbol{\Psi}$.

The mean is (for $k > d+1$)

$$\text{E}(\boldsymbol{X}) = \frac{\boldsymbol{\Psi}}{k-d-1}$$

and the variances of elements of $\boldsymbol{X}$ are (for $k > d+3$)

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

The inverse Wishart distribution $\text{IW}(\boldsymbol{\Psi}, k)$ has pdf

$$p(\boldsymbol{X}|\boldsymbol{\Psi}, k) = \frac{\det(\boldsymbol{\Psi}/2)^{k/2}}{\Gamma_d(k/2)} \det(\boldsymbol{X})^{-(k+d+1)/2} \exp\left(-\text{Tr}(\boldsymbol{\Psi}\boldsymbol{X}^{-1})/2\right)$$

As with the Wishart distribution his pdf is a joint pdf over the d diagonal elements x_{ii} and the $d(d-1)/2$ off-diagonal elements x_{ij} of the symmetric random matrix $\boldsymbol{X}$.

The inverse Wishart and the Wishart distributions are linked. If $\boldsymbol{X} \sim \text{IW}(\boldsymbol{\Psi}, k)$ then the inverse of $\boldsymbol{X}$ is Wishart distributed with inverted scale parameter:

$$\boldsymbol{X}^{-1} \sim \text{Wis}\left(\boldsymbol{S} = \boldsymbol{\Psi}^{-1}, k\right)$$

where k is the shape parameter and $\boldsymbol{S}$ the scale parameter of the Wishart distribution.

If $\boldsymbol{\Psi}$ is a scalar ψ (and $d = 1$) then the multivariate inverse Wishart distribution reduces to the univariate inverse Wishart distribution (Section 4.5).

> R code
>
> The `mniw` package implements the Wishart distribution. The pdf of the Wishart distribution is given by `mniw::diwish()`. The corresponding random number generator is `mniw::riwish()`.

Mean parametrisation

Instead of $\boldsymbol{\Psi}$ and k we may also equivalently use $\boldsymbol{M} = \boldsymbol{\Psi}/(k-d-1)$ and $\kappa = k-d-1$ as parameters for the inverse Wishart distribution, so that

$$\boldsymbol{X} \sim \text{IW}(\boldsymbol{\Psi} = \kappa\boldsymbol{M}\,,\ k = \kappa + d + 1)$$

with mean (for $\kappa > 0$)

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

and variances (for $\kappa > 2$)

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

For $\boldsymbol{M}$ equal to scalar μ with $d = 1$ the above reduces to the univariate inverse Wishart distribution in mean parametrisation.

Biased mean parametrisation

Using $\boldsymbol{T} = (t_{ij}) = \boldsymbol{\Psi}/(k-d+1) = \boldsymbol{\Psi}/\nu$ as biased mean parameter together with $\nu = k-d+1$ we arrive at the **biased mean parametrisation**

$$\boldsymbol{X} \sim \text{IW}(\boldsymbol{\Psi} = \nu\boldsymbol{T}\,,\ k = \nu + d - 1)$$

The corresponding mean is (for $\nu > 2$)

$$\text{E}(\boldsymbol{X}) = \frac{\nu}{\nu-2}\boldsymbol{T} = \boldsymbol{M}$$

and the variances of elements of $\boldsymbol{X}$ are (for $\nu > 4$)

$$\operatorname{Var}(x_{ij}) = \left(\frac{\nu}{\nu-2}\right)^2 \frac{(\nu-2)\,t_{ii}t_{jj} + \nu\,t_{ij}^2}{(\nu-1)(\nu-4)}$$

As $\boldsymbol{T} = \boldsymbol{M}(\nu-2)/\nu$ for large ν the parameter $\boldsymbol{T}$ will become identical to the true mean $\boldsymbol{M}$.

For $\boldsymbol{T}$ equal to scalar τ^2 with $d = 1$ the above reduces to the univariate inverse Wishart distribution in biased mean parametrisation.

Scale transformation

If $\boldsymbol{X} \sim \text{IW}(\boldsymbol{\Psi}, k)$ then the scaled symmetric random matrix $\boldsymbol{AXA}^T$ is also inverse Wishart distributed with $\boldsymbol{AXA}^T \sim \text{IW}(\boldsymbol{A\Psi A}^T, k)$ where the matrix $\boldsymbol{A}$ has full rank and both $\boldsymbol{AXA}^T$ and $\boldsymbol{A\Psi A}^T$ remain positive definite. The matrix $\boldsymbol{A}$ may be rectangular, hence the size of $\boldsymbol{AXA}^T$ and $\boldsymbol{A\Psi A}^T$ may be smaller compared to $\boldsymbol{X}$ and $\boldsymbol{\Psi}$.

5.6 Multivariate t-distribution

The **multivariate t-distribution** $t_\nu(\boldsymbol{\mu}, \boldsymbol{T})$ is a multivariate generalisation of the location-scale t-distribution $t_\nu(\mu, \tau^2)$ (Section 4.6) from one to d dimensions. It is a generalisation of the multivariate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{T})$ (Section 5.3) with an additional parameter $\nu > 0$ (degrees of freedom) controlling the probability mass in the tails.

Special cases include the **multivariate standard t-distribution** $t_\nu(\boldsymbol{0}, \boldsymbol{I})$, the **multivariate normal distribution** $N(\boldsymbol{\mu}, \boldsymbol{T})$ and the **multivariate Cauchy** distribution $\text{Cau}(\boldsymbol{\mu}, \boldsymbol{T})$.

Standard parametrisation

If $\boldsymbol{x} \in \mathbb{R}^d$ is a multivariate t-distributed random variable we write

$$\boldsymbol{x} \sim t_\nu(\boldsymbol{\mu}, \boldsymbol{T})$$

where the vector $\boldsymbol{\mu}$ is the location parameter (a d dimensional vector) and the dispersion parameter $\boldsymbol{T}$ is a symmetric positive definite matrix of dimension $d \times d$. The dimension d is implicit in both parameters. The parameter $\nu > 0$ prescribes the degrees of freedom. For small values of

ν the distribution is heavy-tailed and as a result only moments of order smaller than ν are finite and defined.

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

$$\operatorname{E}(\boldsymbol{x}) = \boldsymbol{\mu}$$

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

$$\operatorname{Var}(\boldsymbol{x}) = \frac{\nu}{\nu-2}\boldsymbol{T}$$

The pdf of $t_\nu(\boldsymbol{\mu}, \boldsymbol{T})$ is

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{T}, \nu) = \det(\boldsymbol{T})^{-1/2} \frac{\Gamma(\frac{\nu+d}{2})}{(\pi\nu)^{d/2}\,\Gamma(\frac{\nu}{2})} \left(1 + \frac{\Delta^2}{\nu}\right)^{-(\nu+d)/2}$$

with $\Delta^2 = (\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{T}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ the squared Mahalanobis distance between $\boldsymbol{x}$ and $\boldsymbol{\mu}$. Note that this pdf is a joint pdf over the d elements $x_1, \ldots, x_d$ of the random vector $\boldsymbol{x}$.

For $d = 1$ the random vector $\boldsymbol{x} = x$ is a scalar, $\boldsymbol{\mu} = \mu$, $\boldsymbol{T} = \tau^2$ and thus the multivariate t-distribution reduces to the location-scale t-distribution (Section 4.6).

> R code
>
> The `mnormt` package implements the multivariate t-distribution. The function `mnormt::dmt()` provides the pdf and `mnormt::pmt()` returns the distribution function. The function `mnormt::rmt()` is the corresponding random number generator.

Scale parametrisation

The multivariate t-distribution, like the multivariate distribution, can also be represented with a matrix scale parameter $\boldsymbol{L}$ in place of a matrix dispersion parameter $\boldsymbol{T}$.

Let $\boldsymbol{L}$ be a matrix scale parameter such that $\boldsymbol{LL}^T = \boldsymbol{T}$ and $\boldsymbol{W} = \boldsymbol{L}^{-1}$ be the corresponding inverse matrix scale parameter with $\boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{T}^{-1}$. By construction $|\det(\boldsymbol{W})| = \det(\boldsymbol{T})^{-1/2}$ and $|\det(\boldsymbol{L})| = \det(\boldsymbol{T})^{1/2}$.

Note that $\boldsymbol{T}$ alone does not fully determine $\boldsymbol{L}$ and $\boldsymbol{W}$ due to rotational freedom, see the discussion in Section 5.3 for details.

Special case: multivariate standard t-distribution

With $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{T} = \boldsymbol{I}$ the multivariate t-distribution reduces to the **multivariate standard t-distribution** $t_\nu(\mathbf{0}, \boldsymbol{I})$. It is a generalisation of the multivariate standard normal distribution $N(\mathbf{0}, \boldsymbol{I})$ to allow for heavy tails.

The distribution has mean $\mathrm{E}(\boldsymbol{x}) = \mathbf{0}$ (for $\nu > 1$) and variance $\mathrm{Var}(\boldsymbol{x}) = \frac{\nu}{\nu-2}\boldsymbol{I}$ (for $\nu > 2$).

The pdf of $t_\nu(\mathbf{0}, \boldsymbol{I})$ is

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

with the squared Mahalanobis distance reducing to $\Delta^2 = \boldsymbol{x}^T\boldsymbol{x}$.

For scalar x (and hence $d = 1$) the multivariate standard t-distribution reduces to the Student's t-distribution $t_\nu = t_\nu(0, 1)$.

Unlike the multivariate standard normal distribution, the density of the multivariate standard t-distribution cannot be written as product of corresponding univariate standard densities.

Special case: multivariate normal distribution

For $\nu \to \infty$ the multivariate t-distribution $t_\nu(\boldsymbol{\mu}, \boldsymbol{T})$ reduces to the **multivariate normal distribution** $N(\boldsymbol{\mu}, \boldsymbol{T})$ (Section 5.3). Correspondingly, for $\nu \to \infty$ the multivariate standard t-distribution $t_\nu(\mathbf{0}, \boldsymbol{I})$ becomes equal to the **multivariate standard normal distribution** $N(\mathbf{0}, \boldsymbol{I})$.

This can be seen from the corresponding limits of the two factors in the pdf of the multivariate t-distribution that depend on ν:

1) Following Sterling's approximation for large x we can approximate $\log\Gamma(x) \approx (x-1)\log(x-1)$. For large ν this implies that
$$\frac{\Gamma((\nu+d)/2)}{(\pi\nu)^{d/2}\,\Gamma(\nu/2)} \to (2\pi)^{-d/2}$$

2) For small x we can approximate $\log(1+x) \approx x$. Thus for large $\nu \gg d$ (and hence small Δ^2/ν) this yields $(\nu+d)\log(1+\Delta^2/\nu) \to \Delta^2$ and hence $\left(1+\Delta^2/\nu\right)^{-(\nu+d)/2} \to e^{-\Delta^2/2}$.

Hence, the pdf of $t_\infty(\boldsymbol{\mu}, \boldsymbol{T})$ is the multivariate normal pdf

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{T}, \nu = \infty) = \det(\boldsymbol{T})^{-1/2}(2\pi)^{-d/2}e^{-\Delta^2/2}$$

Special case: multivariate Cauchy distribution

For $\nu = 1$ the multivariate t-distribution becomes the **multivariate Cauchy distribution** $\mathrm{Cau}(\boldsymbol{\mu}, \boldsymbol{T}) = t_1(\boldsymbol{\mu}, \boldsymbol{T})$.

Its mean, variance and other higher moments are all undefined.

It has pdf

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

For scalar x (and hence $d = 1$) the multivariate Cauchy distribution $\mathrm{Cau}(\boldsymbol{\mu}, \boldsymbol{T})$ reduces to the univariate Cauchy distribution $\mathrm{Cau}(\mu, \tau^2)$.

Special case: multivariate standard Cauchy distribution

The **multivariate standard Cauchy distribution** $\mathrm{Cau}(\mathbf{0}, \boldsymbol{I}) = t_1(\mathbf{0}, \boldsymbol{I})$ is obtained by setting $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{T} = \boldsymbol{I}$ in the multivariate Cauchy distribution or, equivalently, by setting $\nu = 1$ in the multivariate standard t-distribution.

It has pdf

$$p(\boldsymbol{x}) = \Gamma\left(\frac{d+1}{2}\right)\left(\pi(1+\boldsymbol{x}^T\boldsymbol{x})\right)^{-(d+1)/2}$$

For scalar x (and hence $d = 1$) the multivariate standard Cauchy distribution $\mathrm{Cau}(\mathbf{0}, \boldsymbol{I})$ reduces to the standard univariate Cauchy distribution $\mathrm{Cau}(0, 1)$.

Location-scale transformation

Let $\boldsymbol{L}$ be a scale matrix for $\boldsymbol{T}$ and $\boldsymbol{W}$ the corresponding inverse scale matrix.

If $\boldsymbol{x} \sim t_\nu(\boldsymbol{\mu}, \boldsymbol{T})$ then $\boldsymbol{y} = \boldsymbol{W}(\boldsymbol{x} - \boldsymbol{\mu}) \sim t_\nu(\mathbf{0}, \boldsymbol{I})$. This location-scale transformation reduces a multivariate t-distributed random variable to a standard multivariate t-distributed random variable.

Conversely, if $\boldsymbol{y} \sim t_\nu(\mathbf{0}, \boldsymbol{I})$ then $\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{y} \sim t_\nu(\boldsymbol{\mu}, \boldsymbol{T})$. This location-scale transformation generates the multivariate t-distribution from the multivariate standard t-distribution.

Note that for $\nu > 2$ under the location-scale transformation $\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{y}$ with $\text{Var}(\boldsymbol{y}) = \nu/(\nu-2)\boldsymbol{I}$ we get $\text{Cov}(\boldsymbol{x}, \boldsymbol{y}) = \nu/(\nu-2)\boldsymbol{L}$. This provides a means to choose between different factorisations of $\boldsymbol{T}$ and $\boldsymbol{T}^{-1}$. For example, if positive correlation between corresponding elements in $\boldsymbol{x}$ and $\boldsymbol{y}$ is desired then the diagonal elements in $\boldsymbol{L}$ must be positive.

For the special case of the multivariate Cauchy distribution (corresponding to $\nu = 1$) similar relations hold between it and the multivariate standard Cauchy distribution. If $\boldsymbol{x} \sim \text{Cau}(\boldsymbol{\mu}, \boldsymbol{T})$ then $\boldsymbol{y} = \boldsymbol{W}(\boldsymbol{x} - \boldsymbol{\mu}) \sim \text{Cau}(\boldsymbol{0}, \boldsymbol{I})$. Conversely, if $\boldsymbol{y} \sim \text{Cau}(\boldsymbol{0}, \boldsymbol{I})$ then $\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{y} \sim \text{Cau}(\boldsymbol{\mu}, \boldsymbol{T})$.

Convolution property

The multivariate t-distribution is not generally closed under convolution, with the exception of two special cases, the multivariate normal distribution ($\nu = \infty$), see Section 5.3, and the multivariate Cauchy distribution ($\nu = 1$) with the additional restriction that the dispersion parameters are proportional.

For the Cauchy distribution with $\boldsymbol{T}_i = a_i^2\boldsymbol{T}$, where $a_i > 0$ are positive scalars,

$$\sum_{i=1}^{n} \text{Cau}(\boldsymbol{\mu}_i, a_i^2\boldsymbol{T}) \sim \text{Cau}\left(\sum_{i=1}^{n} \boldsymbol{\mu}_i, \left(\sum_{i=1}^{n} a_i\right)^2 \boldsymbol{T}\right)$$

Multivariate t-distribution as compound distribution

The multivariate t-distribution can be obtained as mixture of multivariate normal distributions with identical mean and varying covariance matrix. Specifically, let z be a univariate inverse Wishart random variable

$$z \sim \text{IW}(\psi = \nu, k = \nu) = \text{IG}\left(\alpha = \frac{\nu}{2}, \beta = \frac{\nu}{2}\right)$$

and let $\boldsymbol{x}|z$ be multivariate normal

$$\boldsymbol{x}|z \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma} = z\boldsymbol{T})$$

The resulting marginal (scale mixture) distribution for $\boldsymbol{x}$ is the multivariate t-distribution

$$\boldsymbol{x} \sim t_\nu\left(\boldsymbol{\mu}, \boldsymbol{T}\right)$$

An alternative way to arrive at $t_\nu\left(\boldsymbol{\mu}, \boldsymbol{T}\right)$ is to include $\boldsymbol{T}$ as parameter in the inverse Wishart distribution

$$\boldsymbol{Z} \sim \text{IW}(\boldsymbol{\Psi} = \nu\boldsymbol{T}, k = \nu + d - 1)$$

and let

$$\boldsymbol{x}|\boldsymbol{Z} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma} = \boldsymbol{Z})$$

Note that $\boldsymbol{T}$ is now the biased mean parameter of the multivariate inverse Wishart distribution. This characterisation is useful in Bayesian analysis.

Bibliography

Mardia, K. V., J. T. Kent, and J. M. Bibby. 1979. *Multivariate Analysis*. Academic Press.

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