Distribution Refresher

Korbinian Strimmer

12 December 2023

Contents

Welcome

These notes intended to provide a quick refresher about commonly used univariate and multivariate distributions.

As such the notes are offered as supporting information along with the lecture notes of the statistical modules I am or have been teaching at the Department of Mathematics of the University of Manchester.

This includes the current modules:

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

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

- MATH20802 Statistical Methods.

The Distribution Refresher notes were written by Korbinian Strimmer from 2018–2023. This version is from 12 December 2023.

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

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

About the author

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

License

Chapter 1

Univariate distributions

1.1 Bernoulli distribution

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

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

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

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

and

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

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

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

The log PMF is

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

If a random variable x follows the Bernoulli distribution we write

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

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

1.2 Binomial distribution

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

Its probability mass function is:

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

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

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

If a random variable y follows the binomial distribution we write

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

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

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

1.3 Normal distribution

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

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

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

with mean

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

and variance

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

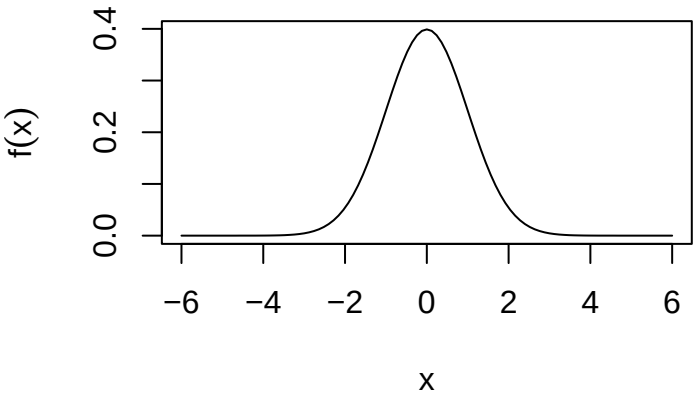
Probability density function (PDF):

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

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

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

Plot of the PDF of the standard normal:

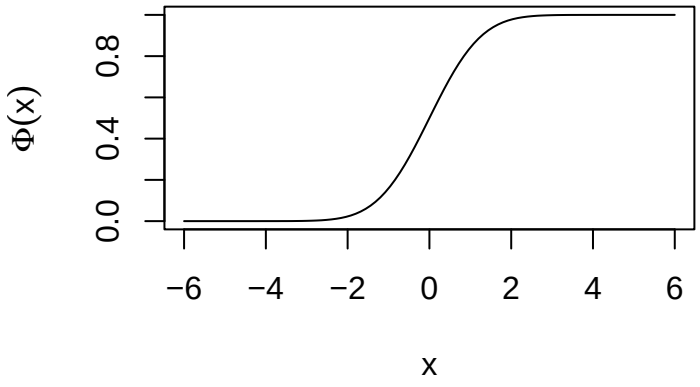


The cumulative distribution function (CDF) of the standard normal $N(0, 1)$ is

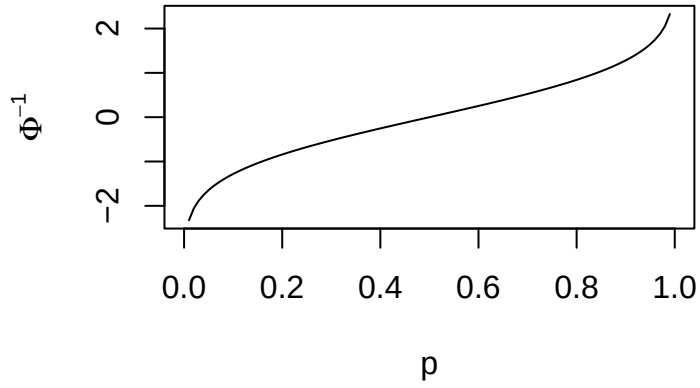
$$\Phi(x) = \int_{-\infty}^{x} p(x'|\mu = 0, \sigma^2 = 1)dx'$$

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

Plot of the CDF of the standard normal:



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



1.4 Gamma distribution (aka Wishart and scaled chi-squared distribution) and special cases (chi-squared and exponential distribution)

The gamma distribution is widely used in statistics, and appears in various parameterisations and under different names, which may be confusing at times.

1.4.1 Standard parameterisation

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

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

with mean

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

and variance

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

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

Probability density function (PDF):

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

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

1.4.2 Wishart parameterisation and scaled chi-squared distribution

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

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

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

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

and the variance

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

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

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

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

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

and variance

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

1.4.3 Construction as sum of squared normals

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

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

Then the sum of the squares

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

follows

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

or equivalently

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

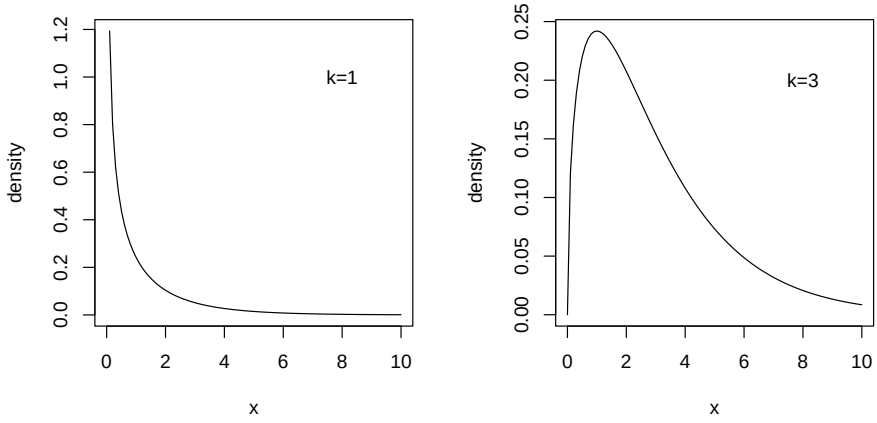
1.4.4 Special cases of the gamma distribution (chi-squared and exponential distribution)

1.4.4.1 Chi-squared distribution

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

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

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



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

1.4.4.2 Exponential distribution

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

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

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

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

1.5 Location-scale t-distribution and special cases (Student's t and Cauchy distribution)

1.5.1 Location-scale t-distribution

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

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

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

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

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

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

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

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

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

1.5.2 Student's t-distribution

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

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

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

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

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

1.5.3 Cauchy and standard Cauchy distribution

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

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

Chapter 2

Multivariate distributions

2.1 Categorical distribution

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

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

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

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

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

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

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

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

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

and the log PMF as

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

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

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

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

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

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

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

2.2 Multinomial distribution

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

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

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

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

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

2.3 Multivariate normal distribution

The univariate normal distribution for a random scalar x generalises to the **multivariate normal distribution** for a random vector $\boldsymbol{x} = (x_1, x_2, ..., x_d)^T \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with with mean $\mathrm{E}(\boldsymbol{x}) = \boldsymbol{\mu}$ and covariance matrix $\mathrm{Var}(\boldsymbol{x}) = \boldsymbol{\Sigma}$. The corresponding density is

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

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

For $d = 1$ we get $\boldsymbol{x} = x$, $\boldsymbol{\mu} = \mu$ and $\boldsymbol{\Sigma} = \sigma^2$ so that the multivariate normal density reduces to the univariate normal density.