MATH2011: Statistical Distribution Theory

Dr Helen Ogden

2019/20

Contents

Part I

Distributions and their properties

Chapter 1

Probability distributions

1.1 Random variables

A random variable Y is described by its domain (or sample space) D together with the probabilities assigned to subsets of the domain. These define the *probability distribution* of the random variable. We distinguish between discrete and continuous random variables.

Discrete probability distributions are defined by a probability (mass) function

$$p(y) \equiv P(Y = y), \quad \text{for } y \in D$$

where

$$\sum_{y \in D} p(y) = 1.$$

The distribution function $F(\cdot)$ is defined for all $y \in \mathbb{R}$ by

$$F(y) \equiv P(Y \leq y) = \sum_{x \in D:\, x \leq y} p(x).$$

Continuous probability distributions are defined by a probability density function (pdf) $f(\cdot)$ where

$$P(y_1 < Y \leq y_2) = \int_{y_1}^{y_2} f(y)dy.$$

The domain of Y is the set $D = \{y \in \mathbb{R} : f(y) > 0\}$ Hence

$$\int_{-\infty}^{\infty} f(y)dy = \int_{D} f(y)dy = 1.$$

The distribution function $F(\cdot)$ is then given by

$$F(y) \equiv P(Y \leq y) = \int_{-\infty}^{y} f(x)dx.$$

Therefore

$$P(y_1 < Y \leq y_2) = F(y_2) - F(y_1)$$

and

$$f(y) = \frac{d}{dy}F(y).$$

1.2 Examples of discrete distributions

1.2.1 The Bernoulli distribution

A *Bernoulli trial* is an experiment with just two possible outcomes 'success' and 'failure' which occur with probabilities θ and $1-\theta$ respectively, where θ is the success probability. The indicator of success in a Bernoulli trial has Bernoulli distribution.

Definition 1.1. A discrete random variables Y has *Bernoulli* distribution if it has probability function of the form

$$p(y) = \theta^y(1-\theta)^{1-y}, \quad y = 0, 1,$$

for some $0 < \theta < 1$. We write $Y \sim \text{Bernoulli}(\theta)$.

1.2.2 The binomial distribution

Suppose we undertake a fixed number, n, of independent Bernoulli trials, each with success probability θ. Let Y be the number of successes in these n trials. Then Y has *binomial* distribution.

Definition 1.2. A discrete random variables Y has *binomial* distribution if it has probability function of the form

$$p(y) = \binom{n}{y}\theta^x(1-\theta)^{n-y}, \quad y = 0, 1, \ldots, n,$$

for some $n \in \mathbb{N}$ and $0 < \theta < 1$. We write $Y \sim \text{binomial}(n, \theta)$.

1.2.3 The negative binomial distribution

Suppose we undertake a sequence of independent Bernoulli trials, each with success probability θ. Let X be the number failures that occur before the kth success. Then X has *negative binomial* distribution.

Definition 1.3. A discrete random variables Y has *negative binomial* distribution if it has probability function of the form

$$p(y) = \binom{k+y-1}{y}(1-\theta)^y\theta^k, \quad y = 0, 1, \ldots,$$

for some $k \in \mathbb{N}$, and $0 < \theta < 1$. We write $Y \sim \text{negbin}(k, \theta)$.

The *geometric* distribution is the special case of the negative binomial distribution with $k = 1$: the number of failures that occur before the first success.

Definition 1.4. A discrete random variables Y has *geometric* distribution if it has probability function of the form

$$p(y) = (1-\theta)^y\theta, \quad y = 0, 1, \ldots,$$

for some $0 < \theta < 1$. We write $Y \sim \text{geometric}(\theta)$.

1.2.4 The Poisson distribution

The *Poisson* distribution arises in a variety of practical situations where we are interested in modelling counts of how often an 'event' occurs.

Definition 1.5. A discrete random variable Y has *Poisson* distribution if it has probability function of the form

$$p(y) = \frac{e^{-\theta}(\theta)^y}{y!}, \quad y = 0, 1, \ldots,$$

for some *rate parameter* $\theta > 0$. We write $Y \sim \text{Poisson}(\theta)$.

In a *Poisson process*, events occur at random at constant rate θ per unit time, independent of all other events. If we define Y as the number of events of a Poisson process in an interval of fixed length t, then $Y \sim \text{Poisson}(t\theta)$.

1.3 Examples of continuous distributions

1.3.1 The exponential distribution

In Section 1.2.4, we considered the Poisson process, in which events occur at random, at a rate θ per unit time. The actual number of events which take

place in any given unit of time has Poisson(θ) distribution. The *exponential* distribution represents the time between consecutive events in this process.

Let Y represent the time interval between two events. Clearly this variable cannot be negative, but can take any positive value. The domain of Y is $(0, \infty)$. We have

$$\begin{aligned} P(Y > y) &= P(\text{no events in an interval of length } y) \\ &= \frac{e^{-\theta y}(\theta y)^0}{0!} \\ &= e^{-\theta y} \end{aligned}$$

so

$$F(y) = P(Y \leq y) = 1 - e^{-\theta y}, \quad y > 0.$$

Differentiating, we obtain the pdf

$$f(y) = \frac{d}{dy} F(y) = \theta e^{-\theta y}, \quad y > 0.$$

Definition 1.6. A random variable Y has *exponential* distribution if it has pdf of the form

$$f(y) = \theta e^{-\theta y}, \quad y > 0,$$

for some *rate parameter* $\theta > 0$. We write $Y \sim \text{exponential}(\theta)$.

Example 1.1. Suppose the lifetime in hours of a certain type of electronic component is described by an Exponential random variable with rate parameter $\theta = 0.01$. What is the probability such a component will have a lifetime of between 100 and 200 hours?

The probability is the area under the curve $f(y) = 0.01e^{-0.01y}$ between $y = 100$ and $y = 200$, so

$$\begin{aligned} P(100 < Y \leq 200) &= \int_{100}^{200} 0.01 e^{-0.01y} dy \\ &= e^{-1} - e^{-2} = 0.37 - 0.14 = 0.23. \end{aligned}$$

We could find this in R with

```
pexp(200, rate = 0.01) - pexp(100, rate = 0.01)
```

```
## [1] 0.2325442
```

1.3.2 The uniform distribution

The uniform distribution is one of the simplest probability distributions: it just places a constant density between some range (a, b):

Definition 1.7. A random variable Y has *uniform* distribution if it has pdf of the form

$$f(y) = \frac{1}{b-a}, a < y < b,$$

for some parameters $a, b \in \mathbb{R}$, with $b > a$. We write $Y \sim U(a, b)$.

1.3.3 The normal distribution

The normal distribution is probably the single most important distribution in statistics. The main reason for its importance is the central limit theorem, which you have seen before in MATH1024, and which we will prove in Section 4.4.

Definition 1.8. A random variable Y has *normal* distribution if it has pdf of the form

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

for some parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. We write $Y \sim N(\mu, \sigma^2)$.

Example 1.2. Suppose that daily water use at a factory varies about a mean of 77500 litres with standard deviation 5700 litres. If demand is normally distributed

1. What proportion of days does the demand fall short of 70000 litres?
2. What proportion of days does demand exceed 90000 litres?
3. What is your reaction to a demand of 175000 gallons?

Writing X for the daily water use in litres , we have $X \sim N(77500, 5700^2)$ We can use `R` to compute the probability $P(X < 70000) = F(70000)$

```
pnorm(70000, mean = 77500, sd = 5700)
```

```
## [1] 0.09412236
```

so the daily water use will be less than 70000 litres about 9.4% of the time.

We can find $P(x > 90000) = 1 - F(90000)$

```
1 - pnorm(90000, mean = 77500, sd = 5700)
```

```
## [1] 0.01415432
```

so the daily water use will be more than 90000 litres about 1.4% of the time.

We can find $P(x > 175000) = 1 - F(175000)$

```
1 - pnorm(175000, mean = 77500, sd = 5700)
```

```
## [1] 0
```

In fact, the number is not exactly zero, but it is so small that the computer is rounding it to zero. Such an extreme water use is therefore surprising, and an explanation should be sought. It is possible that a error has occurred in recording the water use, such as two days data being taken together. Alternatively, perhaps our model which assumes that $X \sim N(77500, 5700^2)$ is incorrect. This idea of surprise at an extreme result of low probability, as predicted by a statistical model, will be important later in this module and also in modules such as MATH2010 Statistical Modelling I.

Chapter 2

Moments

2.1 Expected value

Suppose we have a discrete random variable Y with probability function $p(.)$ with domain D. Although $p(.)$ tells us everything about the properties of Y, it is often useful to summarise the properties Y using a few simple quantities.

A simple summary of the probabilistic properties of Y is the *expected value* or (population) *mean* of Y, denoted by $E(Y)$ or μ, depending on the context.

Definition 2.1. If Y is a discrete random variable with probability function $p(.)$ and domain D, the expected value of Y is

$$\mu = E(Y) = \sum_{y \in D} yp(y).$$

If Y is a continuous random variable with probability density function $f(.)$ and domain D, the expected value of Y is:

$$\mu = E(Y) = \int_{-\infty}^{\infty} yf(y)dy = \int_{D} yf(y)dy.$$

Example 2.1 (Expected value of the Bernoulli distribution)**.** Suppose $Y \sim$ Bernoulli(θ). Then

$$E(Y) = \sum_{y \in \{0,1\}} yp(y) = 0 \times (1 - \theta) + 1 \times \theta = \theta.$$

Example 2.2 (Expected value of the exponential distribution)**.** Suppose $Y \sim$ exponential(θ). Then

$$\begin{aligned} E(Y) &= \int_0^\infty y\theta e^{-\theta y} dy \\ &= \frac{1}{\theta} \int_0^\infty te^{-t} dt \\ &= \frac{1}{\theta} \left\{ [-te^{-t}]_0^\infty + \int_0^\infty e^{-t} dt \right\} \\ &= \frac{1}{\theta} \{0 + 1\} = \frac{1}{\theta}. \end{aligned}$$

Note that we have shown $\int_0^\infty te^{-t} dt = 1$: this result will be useful later on.

We can find the expected value of new random variable $h(Y)$ with

$$E[h(Y)] = \int_{-\infty}^{\infty} h(y) f(y) dy = \int_D h(y) f(y) dy$$

if Y is a continuous random variable with pdf $f(.)$. We may obtain a similar expression in the discrete case.

2.2 Variance

Definition 2.2. The (population) variance of Y is

$$\text{Var}(Y) = E\left\{ [Y - E(Y)]^2 \right\}.$$

It is easy to show that

$$\text{Var}(Y) = E(Y^2) - [E(Y)]^2.$$

$\text{Var}(Y)$ is a measure of spread, and is often denoted by σ^2.

We sometimes use the (population) standard deviation, which is just the square root of the variance, to return to the original scale of measurement of Y.

Example 2.3 (Variance of the Bernoulli distribution)**.** Suppose $Y \sim$ Bernoulli(θ). We have seen in Example 2.1 that $E(Y) = \theta$. We have

$$E(Y^2) = \sum_{y \in \{0,1\}} y^2 p(y) = 0^2 \times (1 - \theta) + 1^2 \times \theta = \theta,$$

so

$$\text{Var}(Y) = E(Y^2) - [E(Y)]^2 = \theta - \theta^2 = \theta(1 - \theta).$$

Example 2.4 (Variance of the exponential distribution)**.** Suppose $Y \sim$ exponential(θ). We have seen in Example 2.2 that $E(Y) = 1/\theta$. We have

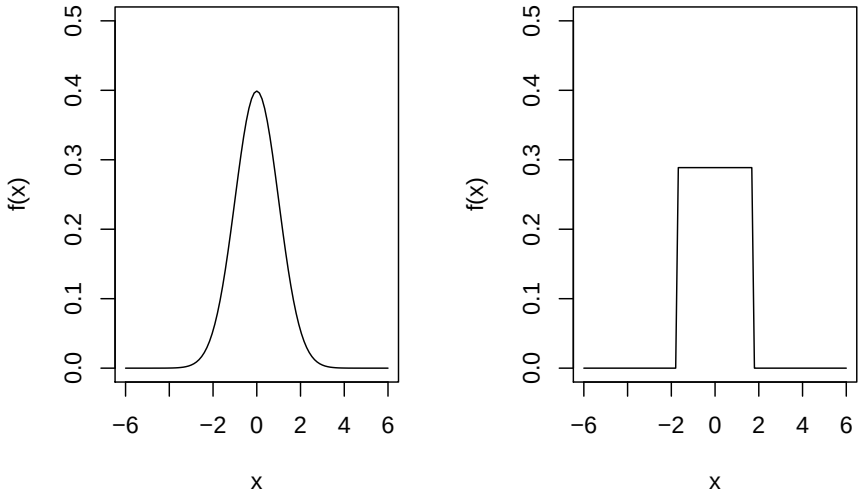
$$\begin{aligned} E(Y^2) &= \int_0^\infty y^2 \theta e^{-\theta y} dy \\ &= \frac{1}{\theta^2} \int_0^\infty t^2 e^{-t} dt \\ &= \frac{1}{\theta} \left\{ [-t^2 e^{-t}]_0^\infty + 2 \int_0^\infty t e^{-t} dt \right\} \\ &= \frac{1}{\theta^2} \{0 + 2\} = \frac{2}{\theta^2}, \end{aligned}$$

since we saw in Example 2.2 that $\int_0^\infty t e^{-t} dt = 1$. So

$$\text{Var}(Y) = E(Y^2) - [E(Y)]^2 = \frac{2}{\theta^2} - \frac{1}{\theta^2} = \frac{1}{\theta^2}.$$

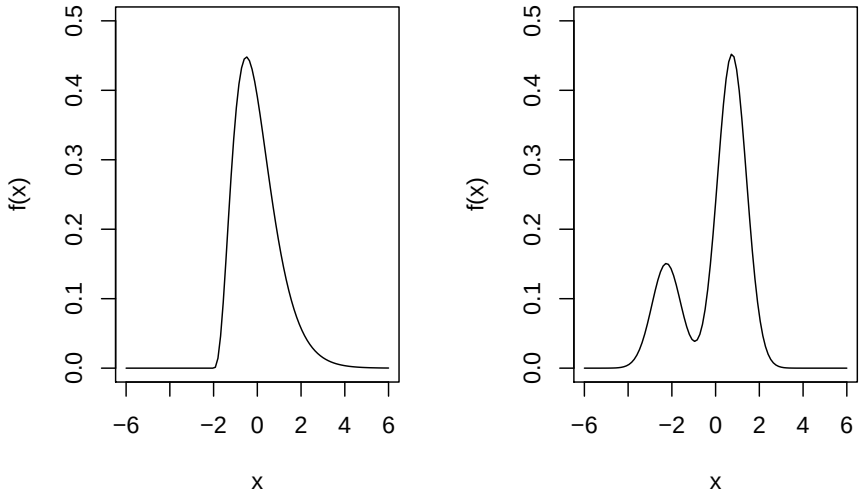
2.3 Higher-order moments

So far we have summarised random variables using the mean and variance (or standard deviation), which measure the "location" and the "spread". Why would we need anything more? Consider the following two density functions of two different continuous distributions:



In fact, the plot on the left is the pdf of a $N(0,1)$ distribution, and the plot on the right is the pdf of a Uniform$(-\sqrt{3}, \sqrt{3})$ distribution. Both have mean 0 and variance 1, yet they look very different. They are both symmetric about 0 (the mean) but differ in terms of "shape".

Consider two more density functions:



Again both have mean 0 and variance 1, yet they look very different. Neither is symmetric and they have different "shapes".

How can we capture something useful about "shape"? We use so-called *higher-order moments* — particularly the third and fourth moments. So we need a general definition of moments and it will be useful to obtain relationships between them.

In what follows, we will assume our random variable Y has continuous distribution. The discrete case follows by replacing the pdf by the probability function and the integral by a sum.

Definition 2.3. The *rth moment about the origin* is

$$\mu'_r = E(Y^r) = \int_{-\infty}^{\infty} y^r f(y) dy$$

and the *rth moment about the mean* is

$$\mu_r = E\left\{[Y - E(Y)]^r\right\} = \int_{-\infty}^{\infty} (y-\mu)^r f(y) dy.$$

We have

$$\begin{aligned}
\mu'_1 &= \mu = E(Y) \\
\mu_1 &= 0 \\
\mu_2 &= \mathrm{Var}(Y).
\end{aligned}$$

How about the third and fourth moments about the mean? We have

$$\mu_3 = E\left\{[Y - E(Y)]^3\right\} = \int_{-\infty}^{\infty} (y-\mu)^3 f(y) dy.$$

Theorem 2.1. *We have*

$$\mu_3 = \mu_3' - 3\mu_2'\mu + 2\mu^3.$$

Remark. This formula allows us to find the third moment about the mean from the first three moments about the origin.

Proof. We have

$$\mu_3 = E\{[Y - E(Y)]^3\} = E\{(Y-\mu)^3\},$$

writing $\mu = E(Y)$. Expanding, we have

$$(Y-\mu)^3 = Y^3 - 3Y^2\mu + 3Y\mu^2 - \mu^3$$

and using the linearity of expectation gives

$$\begin{aligned}\mu_3 &= E(Y^3) - 3\mu E(Y^2) + 3\mu^2 E(Y) - \mu^3 \\ &= \mu_3' - 3\mu\mu_2' + 3\mu^2\mu - \mu^3 \\ &= \mu_3' - 3\mu\mu_2' + 2\mu^3,\end{aligned}$$

as required. □

If Y has symmetric distribution then $\mu_3 = 0$. If Y has a heavier right tail than left tail then $\mu_3 > 0$, and conversely if Y has a heavier left tail than right, then $\mu_3 < 0$.

Similarly, the fourth moment about the mean is

$$\mu_4 = E\left\{[Y - E(Y)]^4\right\} = \int_{-\infty}^{\infty} (y-\mu)^4 f(y) dy.$$

Theorem 2.2. *We have*

$$\mu_4 = \mu_4' - 4\mu_3'\mu + 6\mu_2'\mu^2 - 3\mu^4.$$

Remark. This formula allows us to find the fourth moment about the mean from the first fourth moments about the origin.

The proof is very similar to that of 2.1, and we leave it as an exercise.

For symmetric distributions, roughly speaking thick tails lead to higher values of μ_4 than light tails.

2.4 Standardised moments

Remember that the basic idea is to describe location and spread via the mean and variance and the describe "shape" in terms of the third and fourth moments. So we don't mix up spread with shape we usually use standardised third and fourth moments about the mean.

Definition 2.4. The *skewness* is

$$\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}},$$

the standardised third moment about the mean.

Definition 2.5. The *kurtosis* is

$$\gamma_2 = \frac{\mu_4}{\mu_2^2}$$

the standardised fourth moment about the mean

The skewness and kurtosis are unchanged by linear transformations of Y.

Of course we could consider yet higher order moments to fine tune our understanding of Y. However, we often stop at $r = 4$. Even so, if we just want to obtain the first four moments of a distribution, this may involve a lot of (difficult!) integration. In Chapter 3 we will find a method that allows us to find as many moments as we like but with only one integration required.

We will see in Example 3.4 that any $N(\mu, \sigma^2)$ distribution has skewness $\gamma_1 = 0$ and kurtosis $\gamma_2 = 3$. The kurtosis of other distributions is often compared with that of a normal distribution: if $\gamma_2 < 3$, a distribution has lighter tails than the normal distribution, while if $\gamma_2 > 3$, a distribution has heavier tails than the normal distribution.

Example 2.5 (Higher order Bernoulli moments)**.** Let $Y \sim \text{Bernoulli}(\theta)$. It is easy to see that $\mu'_r = \theta$, for $r = 1, 2, 3, \ldots$.

So, using Theorem 2.1, the third moment about the mean is

$$\begin{aligned}
\mu_3 &= \mu'_3 - 3\mu'_2\mu + 2\mu^3 \\
&= \theta - 3\theta.\theta + 2\theta^3 \\
&= \theta - 3\theta^2 + 2\theta^3 \\
&= \theta(1-\theta)(1-2\theta).
\end{aligned}$$

So the skewness is

$$\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\theta(1-\theta)(1-2\theta)}{[\theta(1-\theta)]^{3/2}} = \frac{1-2\theta}{\sqrt{\theta(1-\theta)}}.$$

Note that the skewness is positive if $\theta < 0.5$, zero if $\theta = 0.5$, and negative if $\theta > 0.5$.

Using Theorem 2.2, the fourth moment about the mean is

$$\begin{aligned}\mu_4 &= \mu_4' - 4\mu_3'\mu + 6\mu_2'\mu^2 - 3\mu^4 \\ &= \theta - 4\theta.\theta + 6\theta.\theta^2 - 3\theta^4 \\ &= \theta(1 - 4\theta + 6\theta^2 - 3\theta^3) \\ &= \theta(1-\theta)(1 - 3\theta + 3\theta^2).\end{aligned}$$

So the kurtosis is

$$\gamma_2 = \frac{\mu_4}{\mu_2^2} = \frac{1 - 3\theta + 3\theta^2}{\theta(1-\theta)}.$$

Note that $\gamma_2 = 1$ for $\theta = 0.5$.

Example 2.6 (Higher order exponential moments)**.** Let $Y \sim$ exponential(θ), with pdf $f(y) = \theta \exp(-\theta y)$ for $y > 0$. We have seen that $E(Y) = \theta^{-1}$ (Example 2.2) and $E(Y^2) = 2\theta^{-2}$ (Example 2.4).

The third moment is

$$\begin{aligned}\mu_3' = E(Y^3) &= \int_0^\infty y^3 \theta e^{-\theta y} dy \\ &= \frac{1}{\theta^3} \int_0^\infty t^3 e^{-t} dt \\ &= \frac{1}{\theta^3} \left\{ \left[-t^3 e^{-t} \right]_0^\infty + 3 \int_0^\infty t^2 e^{-t} dt \right\} \\ &= \frac{1}{\theta^3} \{0 + 3 \times 2\} = \frac{6}{\theta^3},\end{aligned}$$

where we have used the fact that $\int_0^\infty t^2 e^{-t} dt = 2$, from Example 2.4. So, using Theorem 2.1, the third moment about the mean is

$$\mu_3 = \mu_3' - 3\mu_2'\mu + 2\mu^3 = \frac{6}{\theta^3} - \frac{6}{\theta^2} \cdot \frac{1}{\theta} + \frac{2}{\theta^3} = \frac{2}{\theta^3}.$$

So the skewness is

$$\gamma_1 = \frac{(2/\theta^3)}{(1/\theta^2)^{3/2}} = 2,$$

so the exponential distribution has the same positive skewness for all values of the rate parameter θ.

The fourth moment is

$$\begin{aligned}\mu_4' = E(Y^4) &= \int_0^\infty y^4 \theta e^{-\theta y} dy \\ &= \frac{1}{\theta^4} \int_0^\infty t^4 e^{-t} dt \\ &= \frac{1}{\theta^4} \left\{ [-t^4 e^{-t}]_0^\infty + 4 \int_0^\infty t^3 e^{-t} dt \right\} \\ &= \frac{1}{\theta^4} \{0 + 4 \times 6\} = \frac{24}{\theta^4},\end{aligned}$$

where we have used that $\int_0^\infty t^3 e^{-t} dt = 6$, from above. So, using Theorem 2.2, the fourth moment about the mean is

$$\begin{aligned}\mu_4 &= \mu_4' - 4\mu_3'\mu + 6\mu_2'\mu^2 - 3\mu^4 \\ &= \frac{24}{\theta^4} - 4 \cdot \frac{6}{\theta^3} \cdot \frac{1}{\theta} + 6 \cdot \frac{2}{\theta^2} \cdot \frac{1}{\theta^2} - \frac{3}{\theta^4} \\ &= \frac{9}{\theta^4}.\end{aligned}$$

So the kurtosis is

$$\gamma_2 = \frac{9/\theta^4}{(1/\theta^2)^2} = 9,$$

which does not depend on θ. All exponential random variables are positively skewed ($\gamma_1 = 2$), with a high kurtosis ($\gamma_2 = 9$), meaning the exponential distribution has heavier tails than the normal distribution.

Chapter 3

Generating functions

3.1 The moment generating function

We now define an entity — the *moment generating function* (*mgf*) — that enables us to find as many moments as we wish using just a single integral (or sum in the discrete case).

We define the *moment generating function* for Y as

$$M_Y(t) = E(e^{tY}) = \int_{-\infty}^{\infty} e^{ty} f(y) dy,$$

if Y has continuous distribution, or

$$M_Y(t) = E(e^{tY}) = \sum_{y \in D} e^{ty} p(y)$$

if Y has discrete distribution.

A problem with this definition is that in some cases the expectation $E(e^{tY})$ might not be well-defined for some values of t. Provided that there is some value $h > 0$ such that the expectation $E(e^{tY})$ exists for all $-h < t < h$, we say the mgf is well-defined.

Consider e^{ty} expanded as a power series in t:

$$e^{ty} = 1 + ty + \frac{(ty)^2}{2!} + \frac{(ty)^3}{3!} + \ldots = \sum_{k=0}^{\infty} \frac{(ty)^k}{k!}.$$

We can use this power series expansion to see that

$$M_Y(t) = E(e^{tY}) = E\left\{\sum_{k=0}^{\infty} \frac{t^k Y^k}{k!}\right\} = \sum_{k=0}^{\infty} \frac{t^k E(Y^k)}{k!} = \sum_{k=0}^{\infty} \frac{t^k \mu'_k}{k!}. \qquad (3.1)$$

The moment generating function allows us to easily find any moment μ'_r, by using the following result.

Theorem 3.1. *If a random variable Y has moment generating function $M_Y(t)$, for $-h < t < h$, then*

$$\mu'_r = E(Y^r) = M_Y^{(r)}(0),$$

where $M_Y^{(r)}(.)$ is the rth derivative of the moment generating function, for any $r = 0, 1, 2, \ldots$.

Proof. We first claim that

$$M_Y^{(r)}(t) = \sum_{k=0}^{\infty} \frac{t^k \mu'_{k+r}}{k!}. \tag{3.2}$$

for $r = 0, 1, 2, \ldots$. To show (3.2), we proceed by induction. For $r = 0$, this is just the power series (3.1). Now, assuming (3.2) holds for $r - 1$,

$$\begin{aligned}
M_Y^{(r)}(t) &= \frac{d}{dt} M_Y^{(r-1)}(t) \\
&= \frac{d}{dt} \sum_{k=0}^{\infty} \frac{t^k \mu'_{k+r-1}}{k!} \\
&= \frac{d}{dt} \mu'_{k+r-1} + \sum_{k=1}^{\infty} \frac{d}{dt} \frac{t^k \mu'_{k+r-1}}{k!} \\
&= 0 + \sum_{k=1}^{\infty} \frac{t^{k-1} \mu'_{k-1+r}}{(k-1)!} \\
&= \sum_{l=0}^{\infty} \frac{t^l \mu'_{l+r}}{l!},
\end{aligned}$$

so (3.2) is proved. So

$$M_Y^{(r)}(0) = \lim_{t \to 0} \sum_{k=0}^{\infty} \frac{t^k \mu'_{k+r}}{k!} = \mu'_r,$$

as required. □

Example 3.1 (Binomial mgf)**.** Suppose that $Y \sim$ binomial(n, θ), with probability function

$$p(y) = \binom{n}{y} \theta^y (1 - \theta)^{n-y}, \quad y = 0, \ldots n.$$

Then the moment generating function is

$$\begin{aligned} M_Y(t) = E(e^{tY}) &= \sum_{y=0}^{n} e^{ty} \binom{n}{y} \theta^y (1-\theta)^{n-y} \\ &= \sum_{y=0}^{n} \binom{n}{y} (\theta e^t)^y (1-\theta)^{n-y} \\ &= (\theta e^t + 1 - \theta)^n, \end{aligned}$$

by the binomial theorem.

To find $E(Y)$, we first differentiate $M_Y(t)$, to find

$$M_Y^{(1)}(t) = \frac{d}{dt} M_Y(t) = \frac{d}{dt}(\theta e^t + 1 - \theta)^n = n(\theta e^t + 1 - \theta)^{n-1} \theta e^t.$$

We get

$$E(Y) = M_Y^{(1)}(0) = n(\theta + 1 - \theta)^{n-1} \theta = n\theta,$$

as expected.

To find $\text{Var}(Y)$, we first find $E(Y^2)$ by differentiating the mgf a section time. We have

$$\begin{aligned} M_Y^{(2)}(t) &= \frac{d}{dt} n(\theta e^t + 1 - \theta)^{n-1} \theta e^t \\ &= n(n-1)(\theta e^t + 1 - \theta)^{n-2} (\theta e^t)^2 + n(\theta e^t + 1 - \theta)^{n-1} \theta e^t. \end{aligned}$$

We get

$$E(Y^2) = M_Y^{(2)}(0) = n(n-1)\theta^2 + n\theta,$$

so

$$\begin{aligned} \text{Var}(Y) &= E(Y^2) - [E(Y)]^2 \\ &= n(n-1)\theta^2 + n\theta - (n\theta)^2 \\ &= -n\theta^2 + n\theta = n\theta(1-\theta), \end{aligned}$$

as expected.

Example 3.2 (Normal mgf)**.** Suppose that $Y \sim N(\mu, \sigma^2)$, with pdf

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(y-\mu)^2}{2\sigma^2} \right\}, \quad y \in \mathbb{R}.$$

The moment generating function is

$$
\begin{aligned}
M_Y(t) = E(e^{tY}) &= \int_{-\infty}^{\infty} e^{ty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\} dy \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{y^2 - 2\mu y + \mu^2}{2\sigma^2} + ty\right\} dy \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{y^2 - 2(\mu + \sigma^2 t)y + \mu^2}{2\sigma^2}\right\} dy \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{\left[y - (\mu + \sigma^2 t)\right]^2 - (\mu - \sigma^2 t)^2 + \mu^2}{2\sigma^2}\right\} dy \\
&= \exp\left\{\frac{(\mu + \sigma^2 t)^2 - \mu^2}{2\sigma^2}\right\} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{\left[y - (\mu + \sigma^2 t)\right]^2}{2\sigma^2}\right\} dy \\
&= \exp\left\{\frac{\mu^2 + 2\mu\sigma^2 t + \sigma^4 t^2 - \mu^2}{2\sigma^2}\right\} \times 1 \\
&= \exp\left\{\mu t + \frac{\sigma^2 t^2}{2}\right\}.
\end{aligned}
$$

where we have used the fact that the integral of a $N(\mu + \sigma^2 t, \sigma^2)$ pdf over the whole real line is one.

To find $E(Y)$, we first differentiate $M_Y(t)$, to find

$$
\begin{aligned}
M_Y^{(1)}(t) &= \frac{d}{dt} \exp\left\{\mu t + \frac{\sigma^2 t^2}{2}\right\} \\
&= (\mu + \sigma^2 t) \exp\left\{\mu t + \frac{\sigma^2 t^2}{2}\right\}.
\end{aligned}
$$

So $E(Y) = M_Y^{(1)}(0) = \mu$, as expected.

To find $\mathrm{Var}(Y)$, we first find $E(Y^2)$ by differentiating the mgf a section time. We have

$$
\begin{aligned}
M_Y^{(2)}(t) &= \frac{d}{dt}(\mu + \sigma^2 t) \exp\left\{\mu t + \frac{\sigma^2 t^2}{2}\right\} \\
&= \sigma^2 \exp\left\{\mu t + \frac{\sigma^2 t^2}{2}\right\} + (\mu + \sigma^2 t)^2 \exp\left\{\mu t + \frac{\sigma^2 t^2}{2}\right\}
\end{aligned}
$$

We get $E(Y^2) = M_Y^{(2)}(0) = \sigma^2 + \mu^2$, so

$$\mathrm{Var}(Y) = E(Y^2) - [E(Y)]^2 = \sigma^2 + \mu^2 - \mu^2 = \sigma^2,$$

as expected.

3.2 The cumulant generating function

Define the *cumulant generating function* (*cgf*) by

$$K_Y(t) = \log M_Y(t).$$

Define the rth cumulant κ_r by

$$\kappa_r = K_Y^{(r)}(0).$$

What are these cumulants in terms of the more familiar moments?

We have

$$K_Y^{(1)}(t) = \frac{d}{dt} \log M_Y(t) = \frac{M_Y^{(1)}(t)}{M_Y(t)},$$

so

$$\kappa_1 = K_Y^{(1)}(0) = \frac{M_Y^{(1)}(0)}{M_Y(0)} = \frac{\mu_1'}{1},$$

so $\kappa_1 = \mu_1' = \mu$.

We have

$$K_2^{(1)}(t) = \frac{d}{dt} \frac{M_Y^{(1)}(t)}{M_Y(t)} = \frac{M_Y^{(2)}(t)}{M_Y(t)} - \frac{\left[M_Y^{(1)}(t)\right]^2}{\left[M_Y(t)\right]^2},$$

so

$$\kappa_2 = K_Y^{(2)}(0) = \frac{\mu_2'}{1} - \frac{\left(\mu_1'\right)^2}{1^2} = \mu_2' - \mu^2 = \mu_2 = \mathrm{Var}(Y).$$

So we can find $\mathrm{Var}(Y)$ directly by differentiating the cumulant generating function.

In a similar manner, we can show that $\kappa_3 = \mu_3$, so we can find the third central moment directly from the cgf.

It is tempting to assume that $\kappa_4 = \mu_4$, but this is **not** the case. In fact, we can show that

$$\kappa_4 = \mu_4 - 3\mu_2^2,$$

so $\mu_4 = \kappa_4 + 3\mu_2^2 = \kappa_4 + 3\kappa_2^2$.

So we see that if we are just interested in finding the mean, variance, skewness and kurtosis of a distribution, then the cgf is particularly useful.

Example 3.3 (Binomial cgf)**.** Suppose that $Y \sim \mathrm{binomial}(n, \theta)$. From Example 3.1, we know that $M_Y(t) = (\theta e^t + 1 - \theta)^n$, so $K_Y(t) = n \log(\theta e^t + 1 - \theta)$.

Example 3.4 (Normal cgf)**.** Suppose that $Y \sim N(\mu, \sigma^2)$. From Example 3.2, we know that $M_Y(t) = \exp\{\mu t + \frac{1}{2}\sigma^2 t^2\}$, so $K_Y(t) = \mu t + \frac{1}{2}\sigma^2 t^2$.

By differentiating this, we get $\kappa_1 = \mu$, $\kappa_2 = \sigma^2$, and $\kappa_r = 0$ for $r = 3, 4, 5, \ldots$.

The third moment about the mean is $\mu_3 = 0$, so the skewness is $\gamma_1 = 0$ for all normal random variables.

The fourth moment about the mean is $\mu_4 = \kappa_4 + 3\kappa_2^2 = 0 + 3\sigma^4$, so the kurtosis is $\gamma_2 = 3$ for all normal random variables.

3.3 Generating functions under linear transformation

Theorem 3.2. *Let Y be a random variable with mgf $M_Y(t)$ and cgf $K_Y(t)$, and let $Z = a+bY$, where a and b are constants. Then Z has mgf $M_Z(t) = e^{at}M_Y(bt)$ and cgf $K_Z(t) = at + K_Y(bt)$.*

Proof. The moment generating function of Z is

$$\begin{aligned} M_Z(t) &= E(e^{tZ}) = E(e^{t(a+bY)}) = E(e^{at}e^{btY}) \\ &= e^{at}E(e^{(bt)Y}) = e^{at}M_Y(t). \end{aligned}$$

So the cumulant generating function of Z is

$$K_Z(t) = \log M_Z(t) = \log\left\{e^{at}M_Y(t)\right\} = at + \log M_Y(t) = at + K_Y(bt)$$

as required. □

Using this result, we see that some distributions are "closed" under certain types of linear transformation.

Clearly, if two random variables have the same distribution, then they have the same mgf (cgf). We also state without proof that if two random variables have the same mgf (cgf), then they have the same distribution. We say that the mgf and cgf *characterise* the distribution: if we find the mgf (cgf) of a random variable, and find that it matches the mgf (cgf) of a known distribution, we can immediately conclude that the random variable has that distribution.

Example 3.5 (Scaling an exponential distribution)**.** Suppose $Y \sim$ exponential(θ). Let $Z = bY$ where $b > 0$. Then we claim that $Z \sim$ exponential(θ/b).

The mgf of Y is

$$\begin{aligned} M_Y(t) = E(e^{tY}) &= \int_0^\infty e^t y\theta e^{-\theta y} \\ &= \theta \int_0^\infty e^{-(\theta-t)y} dy \\ &= \frac{\theta}{\theta - t} \text{if } \theta - t > 0, \text{ i.e. if } t < \theta. \end{aligned}$$

By Theorem 3.2 with $a = 0$, we have

$$M_Z(t) = M_Y(bt) = \frac{\theta}{\theta - bt} = \frac{\theta/b}{\theta/b - t},$$

which is the exponential(θ/b) mgf. Since the mgf characterises the distribution, we conclude $Z \sim$ exponential(θ/b). A scale change of any exponential random variable gives another exponential random variable.

Example 3.6 (Linear transformation of a Normal distribution)**.** Suppose $Y \sim N(\mu, \sigma^2)$. Let $Z = a + bY$ where $a \in \mathbb{R}$ and $b > 0$. Then we claim that $Z \sim N(a + b\mu, b^2\sigma^2)$.

From Example 3.4, we know that $K_Y(t) = \mu t + \frac{1}{2}\sigma^2 t^2$. By Theorem 3.2, we have

$$K_Z(t) = at + K_Y(bt) = at + \mu bt + \frac{1}{2}\sigma^2(bt)^2 = (a + b\mu)t + \frac{1}{2}(b\sigma)^2 t^2,$$

which is the $N(a+b\mu, b^2\sigma^2)$ cgf. Since the cgf characterises the distribution, we conclude $Z \sim N(a + b\mu, b^2\sigma^2)$. A linear transformation of any normal random variable gives another normal random variable.

Chapter 4

Sums of random variables

4.1 Generating functions of a sum

In many situations in statistics we end up with situations where we are interested in understanding the properties of (possibly weighted) sums of random variables. For example, the sample mean is frequently used as a summary measure for a set of observations and it is simply a weighted sum of the observations (each observation has weight $1/n$, where n is the number of observations).

It is very easy to find the moment generating function of a sum of independent random variables, given the moment generating function for each of those random variables, by using the following result:

Theorem 4.1. *Suppose $Y_1, Y_2, \ldots Y_n$ are independent random variables, where Y_i has moment generating function $M_{Y_i}(t)$, for $i = 1, 2, \ldots, n$. Then $Z = \sum_{i=1}^{n} Y_i$ has moment generating function*

$$M_Z(t) = \prod_{i=1}^{n} M_{Y_i}(t).$$

Remark. A special case of this is that if $Y_1, Y_2, \ldots Y_n$ all have the same distribution with moment generating function $M_Y(t)$, then $M_Z(t) = [M_Y(t)]^n$.

Proof. We have

$$\begin{aligned} M_Z(t) &= E(e^{tZ}) \\ &= E\left[\exp\left(t\sum_{i=1}^{n} Y_i\right)\right] \\ &= E\left[\prod_{i=1}^{n} \exp(tY_i)\right] \\ &= \prod_{i=1}^{n} E\left[\exp(tY_i)\right], \quad \text{by independence of the } Y_i \\ &= \prod_{i=1}^{n} M_{Y_i}(t) \end{aligned}$$

as required. □

A corollary of Theorem 4.1 gives a similar result for cumulant generating functions:

Theorem 4.2. *Suppose $Y_1, Y_2, \ldots Y_n$ are independent random variables, where Y_i has cumulant generating function $K_{Y_i}(t)$, for $i = 1, 2, \ldots, n$. Then $Z = \sum_{i=1}^{n} Y_i$ has cumulant generating function*

$$K_Z(t) = \sum_{i=1}^{n} K_{Y_i}(t).$$

Proof. We have

$$K_Z(t) = \log M_Y(t) = \log\left(\prod_{i=1}^{n} M_{Y_i}(t)\right)$$

by Theorem 4.1. Simplifying, we get

$$K_Z(t) = \sum_{i=1}^{n} \log M_{Y_i}(t) = \sum_{i=1}^{n} K_{Y_i}(t)$$

as required. □

4.2 Closure results for some standard distributions

We can use Theorems 4.1 and 4.2 to prove some useful "closure" results for sums of independent random variables.

Example 4.1 (Sum of binomial random variables)**.** Suppose that $Y_1, \ldots, Y_n$ are independent, with $Y_i \sim$ binomial(n_i, θ). Then we claim that $Z = \sum_{i=1}^n Y_i \sim$ binomial$(\sum_{i=1}^n n_i, \theta)$.

From Example 3.1, the mgf of each Y_i is

$$M_{Y_i}(t) = (\theta e^t + 1 - \theta)^{n_i}.$$

By Theorem 4.1,

$$M_Z(t) = \prod_{i=1}^n M_{Y_i}(t) = \prod_{i=1}^n (\theta e^t + 1 - \theta)^{n_i} = (\theta e^t + 1 - \theta)^{\sum_{i=1}^n n_i},$$

which is the binomial$(\sum_{i=1}^n n_i, \theta)$ mgf. Since the mgf characterises the distribution, we conclude that $Z \sim$ binomial$(\sum_{i=1}^n n_i, \theta)$.

As a corollary, note that if the Y_i are independent Bernoulli(θ) random variables (in which case $Y_i \sim$ binomial$(1, \theta)$), then $Z \sim$ binomial(n, θ).

Example 4.2 (Sum of normal random variables)**.** Suppose that $Y_1, \ldots, Y_n$ are independent, with $Y_i \sim N(\mu_i, \sigma_i^2)$. Then we claim that $Z = \sum_{i=1}^n Y_i \sim N(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2)$.

From Example 3.4, we know that $K_{Y_i}(t) = \mu_i t + \frac{1}{2}\sigma_i^2 t^2$. By Theorem 4.2, we have

$$\begin{aligned} K_Z(t) &= \sum_{i=1}^n K_{Y_i}(t) \\ &= \sum_{i=1}^n \left\{ \mu_i t + \frac{1}{2}\sigma_i^2 t^2 \right\} \\ &= \left(\sum_{i=1}^n \mu_i \right) t + \frac{1}{2} \left(\sum_{i=1}^n \sigma_i^2 \right) t^2 \end{aligned}$$

which is the $N(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2)$ cgf. Since the cgf characterises the distribution, we conclude that $Z \sim N(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2)$.

As a corollary, note that if the Y_i are independent $N(\mu, \sigma^2)$ random variables, then $Z \sim N(n\mu, n\sigma^2)$.

Example 4.3 (Sum of negative binomial random variables)**.** Suppose that $Y_1, \ldots, Y_n$ are independent, with $Y_i \sim$ negbin(k_i, θ). Then $Z = \sum_{i=1}^n Y_i \sim$ negbin$(\sum_{i=1}^n k_i, \theta)$. The proof of this is left as an exercise.

Example 4.4 (Sum of Poisson random variables)**.** Suppose that $Y_1, \ldots, Y_n$ are independent, with $Y_i \sim$ Poisson(θ_i). Then $Z = \sum_{i=1}^n Y_i \sim$ Poisson$(\sum_{i=1}^n \theta_i)$. The proof of this is left as an exercise.

4.3 Properties of the sample mean of normal observations

We may use cumulant generating functions to derive a well-known result about the distribution of the sample mean of normal observations:

Proposition 4.1. *Suppose that $Y_1, \ldots, Y_n$ are independent and identically distributed, with $Y_i \sim N(\mu, \sigma^2)$, and let*

$$\bar{Y} = \frac{\sum_{i=1}^n Y_i}{n}$$

be the sample mean. Then $\bar{Y} \sim N(\mu, \sigma^2/n)$.

Proof. Let $Z = \sum_{i=1}^n Y_i$. We have seen in Example 4.2 that the cgf of Z is

$$K_Z(t) = n\mu t + \frac{1}{2} n\sigma^2 t^2.$$

We have $\bar{Y} = \frac{1}{n} Z$. By Theorem 3.2 with $a = 0$ and $b = 1/n$, $\bar{Y}$ has cgf

$$\begin{aligned} K_{\bar{Y}}(t) &= K_Z\left(\frac{1}{n}t\right) \\ &= n\mu\left(\frac{1}{n}t\right) + \frac{1}{2}n\sigma^2\left(\frac{1}{n}t\right)^2 \\ &= \mu t + \frac{1}{2}\frac{\sigma^2}{n}t^2, \end{aligned}$$

which is the $N(\mu, \sigma^2/n)$ cgf. Since the cgf characterises the distribution, we conclude that $\bar{Y} \sim N(\mu, \sigma^2/n)$. □

4.4 The central limit theorem

If we have n independent and identically distributed $N(\mu, \sigma^2)$ random variables, we have just shown that the sample mean has $N(\mu, \sigma^2/n)$ distribution. The central limit theorem says that if n is large, the sample mean has approximately this distribution even if the Y_i are not normally distributed.

Theorem 4.3 (Central limit theorem)**.** *Suppose that $Y_1, \ldots, Y_n$ are independent and identically distributed, with $E(Y_i) = \mu$ and $Var(Y_i) = \sigma^2 < \infty$, and with cumulants $\kappa_r < \infty$ for all $r = 1, 2, \ldots$. Let*

$$Z = \frac{\sqrt{n}(\bar{Y} - \mu)}{\sigma}$$

Then $Z \to N(0, 1)$ as $n \to \infty$.

Remark. If n is large, the limiting distribution may be used as an approximation. If n is large, then Z has approximately $N(0,1)$ distribution, or equivalently $\bar{Y}$ has approximately $N(\mu, \sigma^2/n)$ distribution.

Proof. From Example 3.4, a $N(0,1)$ random variable has second cumulant 1 and all other cumulants are 0. So in our sketch proof, we aim to show that the cumulants of Z have the same structure as $n \to \infty$. Let the rth cumulant of Z be κ_r^*. We will show that $\kappa_1^* = 0$, $\kappa_2^* = 1$, and $\kappa_r^* \to 0$ as $n \to \infty$.

Since the Y_i are identically distributed, we write $K_{Y_i}(t) = K_Y(t)$. Then, by Theorem 4.2, we have

$$K_{\sum_{i=1}^n Y_i}(t) = nK_Y(t).$$

By Theorem 3.2 with $a = 0$, $b = 1/n$,

$$K_{\bar{Y}}(t) = K_{\sum_{i=1}^n Y_i}\left(\frac{t}{n}\right) = nK_Y\left(\frac{t}{n}\right).$$

Now let $c = \sqrt{n}/\sigma$ and $d = -\sqrt{n}\mu/\sigma$, so that $Z = d + c\bar{Y}$. By Theorem 3.2, we have

$$K_Z(t) = K\bar{Y}(ct) + dt = nK_Y(ct/n) + dt,$$

and we have written the cgf of Z in terms of the cgf of the original random variables.

Differentiating with respect to t, we get

$$K_Z^{(1)}(t) = nK_Y^{(1)}\left(\frac{ct}{n}\right)\frac{c}{n} + d$$

and

$$K_Z^{(2)}(t) = nK_Y^{(2)}\left(\frac{ct}{n}\right)\frac{c^2}{n^2}.$$

So the first cumulant of Z is

$$\begin{aligned}\kappa_1^* &= K_Z^{(1)}(0) = nK_Y^{(1)}(0)\frac{c}{n} + d \\ &= n\kappa_1\frac{c}{n} + d = \kappa_1 c + d \\ &= \frac{\mu\sqrt{n}}{\sigma} - \frac{\sqrt{n}\mu}{\sigma} = 0,\end{aligned}$$

where we have used that $\kappa_1 = E(Y) = \mu$.

The second cumulant of Z is

$$\begin{aligned}\kappa_2^* &= K_Z^{(2)}(0) = nK_Y^{(2)}(0)\frac{c^2}{n^2} \\ &= \kappa_2 \frac{c^2}{n} \\ &= \sigma^2 \frac{n}{\sigma^2 n} = 1,\end{aligned}$$

where we have used that $\kappa_2 = \mathrm{Var}(Y) = \sigma^2$.

For $r \geq 3$ we have

$$K_Z^{(r)}(t) = nK^{(r)}\left(\frac{ct}{n}\right)\frac{c^r}{n^r},$$

so the rth cumulant of Z is

$$\begin{aligned}\kappa_r^* &= K_Z^{(r)}(0) = n\kappa_r \frac{c^r}{n^r} \\ &= n\kappa_r \frac{n^{r/2}}{n^r \sigma^r} \\ &= \frac{\kappa_r}{n^{r/2-1}\sigma^r} \to 0\end{aligned}$$

as $r \geq 3$, so $r/2 - 1 > 0$.

Hence $Z \to N(0,1)$ as $n \to \infty$.

In order to make the proof fully rigorous, we would need to define what we mean by "convergence in distribution" more carefully, and to show that convergence of all cumulants of Z to the cumulants of the limiting distribution implies this convergence in distribution. □

Chapter 5

Maxima and minima

5.1 Order Statistics

Suppose that $Y_1, \ldots, Y_n$ represent n independently and identically distributed random variables each with cumulative distribution function F.

Suppose that the corresponding observed values are $y_1, \ldots, y_n$. Let these values, when ordered, be represented by

$$y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}.$$

The $y_{(i)}$, $i = 1, \ldots, n$, are called the *order statistics* corresponding to $y_1, \ldots, y_n$.

You have already met certain order statistics. For example, the sample median is an order statistic: for odd values of n the sample median is equal to $y_{(\{n+1\}/2)}$, while for even n the sample median is defined as

$$[y_{(n/2)} + y_{(n/2+1)}]/2.$$

We shall concentrate, however, on two particular order statistics: $y_{(1)}$, the sample minimum, and $y_{(n)}$, the sample maximum. We define the corresponding random variables $Y_{(1)} = \min\{Y_1, \ldots, Y_n\}$ and $Y_{(n)} = \max\{Y_1, \ldots, Y_n\}$.

There are many applications for which maxima or minima are of interest. For example:

- Understanding outliers in statistical data: an outlier will often be the largest or smallest data point.
- In reliability engineering, a system will tend to fail at its weakest point (which might be thought of as the point with the minimum "strength").
- In designing coastal defences one needs to understand the distribution of the wave heights of the highest tides.
- In insurance the behaviour of the largest claims is important.

There is a whole area of statistics devoted to the study of extremes, called extreme value theory. In this short chapter we just give a brief introduction to the subject.

5.2 The cdf of $Y_{(n)}$, the largest value in a random sample of size n

Since $Y_{(n)} = \max\{Y_1, \ldots, Y_n\}$, the probability that $Y_{(n)} \leq y$ gives the cumulative distribution function of $Y_{(n)}$, the sample maximum.

Now the event $\{Y_{(n)} \leq y\}$ is identical to the event

$$\{Y_1 \leq y \text{ and } Y_2 \leq y \ldots \text{ and } Y_n \leq y\}.$$

So

$$G_n(y) = P(Y_{(n)} \leq y) = P(\text{all } Y_i \leq y) = P(Y_1 \leq y \text{ and } Y_2 \leq y \ldots \text{ and } Y_n \leq y).$$

Thus, by independence

$$G_n(y) = P(Y_1 \leq y)P(Y_2 \leq y) \ldots P(Y_n \leq y) = [F(y)]^n.$$

Example 5.1 (Maximum of dice rolls)**.** Suppose I roll a fair die twice. What is the probability function of the maximum of the two scores?

We know that $F(y) = y/6$ for $y = 1, 2, 3, 4, 5, 6$, and $n = 2$. So the distribution function of the maximum of the two scores is

$$G_2(y) = \left(\frac{y}{6}\right)^2, \quad \text{for } y = 1, 2, \ldots, 6.$$

Hence

$$P(Y_{(2)} = y) = \begin{cases} \left(\frac{1}{6}\right)^2 & \text{if } y = 1 \\ \left(\frac{y}{6}\right)^2 - \left(\frac{y-1}{6}\right)^2 & \text{if } y = 2, \ldots, 6. \end{cases}$$

5.3 The pdf of the maximum in the continuous case

If the Y_i are continuous, each with density function f, then the density function of $Y_{(n)}$ may be found by differentiating $G_n(y)$ with respect to y, to give

$$g_n(y) = \frac{d}{dy}G_n(y) = \frac{d}{dy}[F(y)]^n = n[F(y)]^{n-1}f(y)$$

where the domain of the maximum is the same as that of each of the Y_i.

Example 5.2 (Maximum of a uniform random sample)**.** Suppose that each $Y_i \sim U(0, \theta)$, so that

$$F(y) = \frac{y}{\theta} \quad \text{for } 0 < y < \theta.$$

So $Y_{(n)}$ has cdf

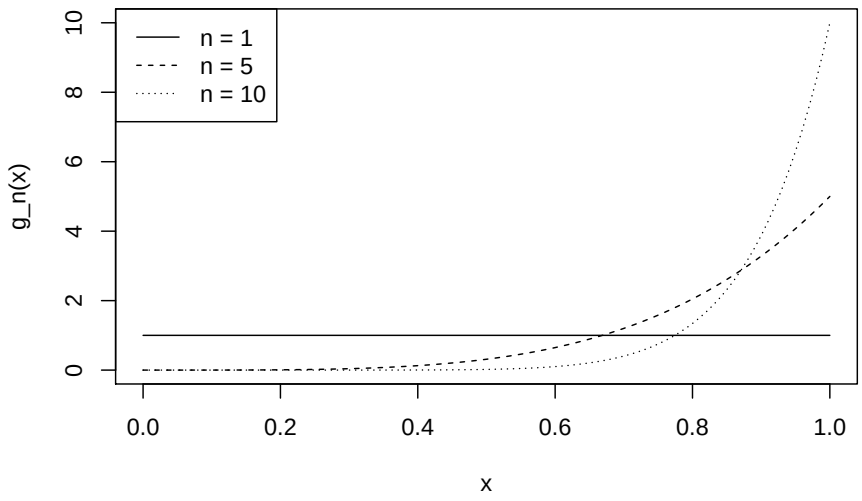
$$G_n(y) = \frac{y^n}{\theta^n} \quad \text{for } 0 < y < \theta,$$

and pdf

$$g_n(y) = \frac{d}{dy} G_n(y) = \frac{d}{dy} \frac{y^n}{\theta^n} = \frac{n y^{n-1}}{\theta^n}, \quad \text{for } 0 < y < \theta.$$

We can use `R` to plot out what these pdfs look like, for $\theta = 1$, and $n = 1, 5$, or 10:

```
g_n <- function(y, n, theta) {
    n * y^{n-1} / theta^n
}
curve(g_n(x, n = 10, theta = 1), from = 0, to = 1, lty = 3,
      ylab = "g_n(x)")
curve(g_n(x, n = 5, theta = 1), add = TRUE, lty = 2)
curve(g_n(x, n = 1, theta = 1), add = TRUE, lty = 1)
legend("topleft", lty = 1:3,
       legend = c("n = 1", "n = 5", "n = 10"))
```



As expected, we see that as n increases, it becomes more likely that the maximum of the n $U(0, 1)$ variables is close to 1.

5.4 The cdf of $Y_{(1)}$, the smallest value in a random sample of size n

Since $Y_{(1)} = \min\{Y_1, \ldots, Y_n\}$, the probability that $Y_{(1)} \leq y$ gives the cumulative distribution function of $Y_{(1)}$, the smallest value in the sample. Now

$$\begin{aligned} G_1(y) = P(Y_{(1)} \leq y) &= 1 - P(Y_{(1)} > y) \\ &= 1 - P(\text{all } Y_i > y) \\ &= 1 - P(Y_1 > y \text{ and } Y_2 > y \ldots \text{ and } Y_n > y) \\ &= 1 - P(Y_1 > y)P(Y_2 > y) \ldots P(Y_n > y) \\ &= 1 - [1 - F(y)]^n. \end{aligned}$$

by independence of the Y_i.

5.5 The pdf of the minimum in the continuous case

If the Y_i are continuous, each with probability density function f, then the pdf of $Y_{(1)}$ may be found by differentiating $G_1(y)$ with respect to y to give

$$g_1(y) = \frac{d}{dy} G_1(y) = n[1 - F(y)]^{n-1} f(y),$$

where the domain of the minimum is the same as that of each of the Y_i.

Example 5.3 (Minimum of an exponential random sample)**.** Suppose that each $Y_i \sim \text{Exp}(\theta)$, with pdf

$$f(y) = \theta e^{-\theta y}, \quad 0 < y < \infty$$

and cdf

$$F(y) = \int_0^y f(u)du = 1 - e^{-\theta y}.$$

Then $Y_{(1)} = \min\{Y_{(1)}, \ldots, Y_{(n)}\}$ has cdf

$$G_1(y) = 1 - [1 - F(y)]^n = 1 - e^{-n\theta y}$$

and pdf

$$g_1(y) = \frac{d}{dy} G_1(y) = n\theta e^{-n\theta y}.$$

So the distribution of the smallest value in an Exponential random sample of size n with rate parameter θ (i.e. with mean value $1/\theta$) is also an Exponential random variable but with rate parameter $n\theta$ (i.e. with mean value $1/(n\theta)$).

This hints at some interesting structure in the probabilistic behaviour of maxima and minima. The central limit theorem essentially says that under certain conditions the sum of n independent, identically distributed random variables is approximately normal as n grows large. There are corresponding results for maxima and minima, though the large-n distribution is not normal (it is the so-called generalised extreme value distribution).

Chapter 6

The gamma distribution

6.1 The gamma distribution

In this chapter we will introduce a family of distributions known as the gamma distribution. We will see that this family contains several sub-families of distributions that arise in practical problems, including the exponential, chi-squared and Erlang distributions. The gamma distribution is also related to normal, t and F distributions.

In practice we are often interested in modelling data which is positive, such as the time to the occurrence of an event of interest. We have seen that an Exponential distribution is one possibility to model such data, but it lacks flexibility — it only has a single parameter and we have seen that all exponential pdfs have the same shape (see Example 2.6). It is natural to try to generalise the exponential to a more flexible family of distributions, and the gamma distribution is one way of achieving this.

Before we introduce the gamma distribution, we first need to introduce a function which is used in the pdf of the gamma distribution, to make sure that the pdf integrates to one.

Definition 6.1 (Gamma function)**.** The *Gamma function* is defined by

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx.$$

When $\alpha = 1$, we have

$$\Gamma(1) = \int_0^\infty e^{-x} dx = 1.$$

Next, let's prove a useful property about the Gamma function.

Proposition 6.1. *For any* $\alpha > 1$, $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$.

Proof. We use integration by parts.

$$\begin{aligned}\Gamma(\alpha) &= \int_0^\infty x^{\alpha-1}e^{-x}dx \\ &= \left[-x^{\alpha-1}e^{-x}\right]_0^\infty + (\alpha-1)\int_0^\infty x^{\alpha-1}e^{-x}dx \\ &= 0 + (\alpha-1)\int_0^\infty x^{(\alpha-1)-1}e^{-x}dx \\ &= (\alpha-1)\Gamma(\alpha-1)\end{aligned}$$

since $\alpha - 1 > 0$ as $\alpha > 1$. □

As a consequence of Proposition 6.1 and the fact that $\Gamma(1) = 1$, if α is any positive integer then

$$\Gamma(\alpha) = (\alpha - 1)!,$$

so the Gamma function can be thought of as a continuous version of the factorial function.

It is also useful to know what happens to the Gamma function at half integer points, $\Gamma(1/2)$, $\Gamma(3/2)$ and so on. It can be shown that

$$\Gamma(1/2) = \sqrt{\pi},$$

and this may be used along with the recurrence relationship (Proposition 6.1) to compute any other $\Gamma(n/2)$.

Definition 6.2. We say a random variable Y has *gamma distribution* with *shape* parameter $\alpha > 0$ and *rate* parameter $\beta > 0$ if it has pdf

$$f(y) = \frac{\beta^\alpha}{\Gamma(\alpha)}y^{\alpha-1}e^{-\beta y}, \quad y > 0. \tag{6.1}$$

We write this as $Y \sim \text{gamma}(\alpha, \beta)$.

Note that the exponential distribution with rate parameter β is a special case of the gamma distribution, with $\alpha = 1$.

Example 6.1 (Time until the kth event in a Poisson process)**.** Suppose we have a Poisson process with events arriving at rate β per unit time, so that the number of events occurring in a time interval of length t has Poisson(βt) distribution. Let Y be the time until the kth event occurs. We claim that $Y \sim \text{gamma}(k, \beta)$.

Let N_y be the number of events in the time interval $(0, y]$. Then $N_y \sim \text{Poisson}(\beta y)$. We have

$$P(Y > y) = P(N_y \leq k-1) = \sum_{n=0}^{k-1} \frac{(\beta y)^n e^{-\beta y}}{n!},$$

so

$$\begin{aligned} F(y) = P(Y \leq y) &= 1 - P(Y > y) \\ &= 1 - \sum_{n=0}^{k-1} \frac{(\beta y)^n e^{-\beta y}}{n!} \\ &= 1 - e^{-\beta y} - \sum_{n=1}^{k-1} \frac{(\beta y)^n e^{-\beta y}}{n!}. \end{aligned}$$

So

$$\begin{aligned} f(y) = \frac{dF}{dy} &= \beta e^{-\beta y} - \sum_{n=1}^{k-1} \left\{ \frac{\beta^n y^{n-1} e^{-\beta y}}{(n-1)!} - \beta \frac{(\beta y)^n}{n!} e^{-\beta y} \right\} \\ &= \beta e^{-\beta y} - \beta e^{-\beta y} + \beta^2 y e^{-\beta y} - \beta^2 y e^{-\beta y} + \frac{\beta^3 y^2}{2!} e^{-\beta y} - \ldots + \frac{\beta (\beta y)^{k-1}}{(k-1)!} e^{-\beta y} \\ &= \frac{\beta^k}{\Gamma(k)} y^{k-1} e^{-\beta y}, \end{aligned}$$

which is the pdf of a gamma(k, β) random variable.

If α is an integer, as in this example, the gamma(α, β) distribution is also known as the *Erlang* distribution. The exponential distribution (with $\alpha = 1$) is a special case of the Erlang distribution.

6.2 Properties of the gamma distribution

Proposition 6.2. *The moment generating function of the gamma(α, β) distribution is*

$$M_Y(t) = \left(1 - \frac{t}{\beta}\right)^{-\alpha}, \quad \textit{for } t < \beta$$

Proof. We have

$$\begin{aligned} M_Y(t) = E(e^{tY}) &= \int_0^\infty e^{ty} \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y} dy \\ &= \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-y(\beta-t)} dy \\ &= \frac{\beta^\alpha}{(\beta-t)^\alpha} \int_0^\infty \frac{(\beta-t)^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-y(\beta-t)} dy \\ &= \left(1 - \frac{t}{\beta}\right)^{-\alpha} \times 1, \end{aligned}$$

as the final integrand is the pdf of a gamma$(\alpha, \beta - t)$ distribution, which integrates to 1 provided that the new rate parameter $\beta - t > 0$, i.e. if $t < \beta$. □

The following proposition tells us that the gamma distribution is always positively skewed, but the skewness decreases as α increases. The gamma distribution always has larger kurtosis than the normal distribution, but the kurtosis decreases towards the kurtosis of a normal distribution ($\gamma_2 = 3$) as α increases.

Proposition 6.3. *Suppose $Y \sim$ gamma(α, β). Then $E(Y) = \alpha\beta^{-1}$ and $Var(Y) = \alpha\beta^{-2}$. The skewness is $\gamma_1 = 2\alpha^{-1/2}$ and the kurtosis is $\gamma_2 = 3 + 6\alpha^{-1}$.*

Proof. From Proposition 6.2, the mgf is $M_Y(t) = (1 - t/\beta)^{-\alpha}$, so the cgf is $K_Y(t) = \log M_Y(t) = -\alpha \log(1 - t/\beta)$. Differentiating, we have

$$K_Y^{(1)}(t) = \frac{\alpha}{\beta\left(1 - \frac{t}{\beta}\right)} = \frac{\alpha}{\beta - t},$$

so $E(Y) = K_Y^{(1)}(0) = \alpha\beta^{-1}$.

Differentiating the cgf again,

$$K_Y^{(2)}(t) = \frac{d}{dt}\left[\alpha(\beta - t)^{-1}\right] = \alpha(\beta - t)^{-2},$$

so $\mu_2 = \mathrm{Var}(Y) = K_Y^{(2)}(0) = \alpha\beta^{-2}$.

Differentiating the cgf again,

$$K_Y^{(3)}(t) = \frac{d}{dt}\left[\alpha(\beta - t)^{-2}\right] = 2\alpha(\beta - t)^{-3},$$

so the third moment about the mean is $\mu_3 = K_Y^{(3)}(0) = 2\alpha\beta^{-3}$. So the skewness is

$$\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}} = \frac{2\alpha\beta^{-3}}{[\alpha\beta^{-2}]^{3/2}} = 2\alpha^{-1/2}.$$

Differentiating the cgf again,

$$K_Y^{(4)}(t) = \frac{d}{dt}\left[2\alpha(\beta - t)^{-3}\right] = 6\alpha(\beta - t)^{-4},$$

so the fourth cumulant is $\kappa_4 = 6\alpha\beta^{-4}$. The fourth moment about the mean is

$$\mu_4 = \kappa_4 + 3\mu_2^2 = 6\alpha\beta^{-4} + 3\alpha^2\beta^{-4},$$

so the kurtosis is

$$\gamma_2 = \frac{\mu_4}{\mu_2^2} = 3 + \frac{6\alpha\beta^{-4}}{\alpha^2\beta^{-4}} = 3 + 6\alpha^{-1},$$

as claimed. □

The gamma distribution is closed under scaling, and under summation of independent gamma distributions with a common rate parameter.

Proposition 6.4. *Suppose $Y \sim$ gamma(α, β), and let $Z = bY$, for some constant $b > 0$. Then $Z \sim$ gamma$(\alpha, \beta/b)$.*

Proof. By Theorem 3.2, we have

$$M_Z(t) = M_Y(bt) = \left(1 - \frac{bt}{\beta}\right)^{-\alpha} = \left(1 - \frac{t}{\beta/b}\right)^{-\alpha},$$

which is the mgf of a gamma$(\alpha, \beta/b)$ distribution. Since the mgf characterises the distribution, we have $Z \sim$ gamma$(\alpha, \beta/b)$. □

Proposition 6.5. *If $Y_1, \ldots, Y_n$ are independent, with $Y_i \sim$ gamma(α_i, β) then $Z = \sum_{i=1}^n Y_i \sim$ gamma$(\sum_{i=1}^n \alpha_i, \beta)$.*

Proof. By Theorem 4.1, we have

$$\begin{aligned} M_Z(t) &= \prod_{i=1}^n M_{Y_i}(t) \\ &= \prod_{i=1}^n \left(1 - \frac{t}{\beta}\right)^{-\alpha_i} \qquad\qquad = \left(1 - \frac{t}{\beta}\right)^{-\sum_{i=1}^n \alpha_i}, \end{aligned}$$

which is the mgf of a gamma$(\sum_{i=1}^n \alpha_i, \beta)$ distribution. Since the mgf characterises the distribution, we have $Z \sim$ gamma$(\sum_{i=1}^n \alpha_i, \beta)$. □

6.3 The chi-squared distribution

The chi-squared distribution is a useful special case of the gamma distribution.

Definition 6.3. We say a random variable Y has *chi-squared* distribution with n degrees of freedom if $Y \sim \text{gamma}(n/2, 1/2)$. We write $Y \sim \chi^2_n$.

We can write down the properties of the chi-squared distribution by using results we have already proved about the gamma distribution.

If $Y \sim \chi^2_n$, we have

$$E(Y) = \frac{n/2}{1/2} = n; \qquad \text{Var}(Y) = \frac{n/2}{(1/2)^2} = 2n.$$

By Proposition 6.5, if $Y_1, \ldots, Y_n$ are independent random variables, with $Y_i \sim \chi^2_{n_i}$ then

$$\sum_{i=1}^{n} Y_i \sim \chi^2_{\sum_{i=1}^{n} n_i}. \tag{6.2}$$

The *sum of independent chi-squared random variables* has chi-squared distribution with degrees of freedom given by the *sum of the individual degrees of freedom*.

The chi-squared distribution is also related to the normal distribution.

Proposition 6.6. *Let $Z \sim N(0,1)$, and let $Y = Z^2$. Then $Y \sim \chi^2_1$.*

Proof. Write $\Phi(.)$ for the cumulative distribution function of Z, and $\phi(.)$ for the pdf of Z, so that

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right).$$

Let Y have cdf $F(.)$ and pdf $f(.)$. Then, for $y > 0$,

$$\begin{aligned} F(y) &= P(Y \leq y) = P(Z^2 \leq y) \\ &= P\left(-\sqrt{y} \leq Z \leq \sqrt{y}\right) \\ &= \Phi\left(\sqrt{y}\right) - \Phi\left(-\sqrt{y}\right). \end{aligned}$$

So

$$\begin{aligned} f(y) = \frac{d}{dy}F(y) &= \frac{1}{2\sqrt{y}}\phi\left(\sqrt{y}\right) + \frac{1}{2\sqrt{y}}\phi\left(-\sqrt{y}\right) \\ &= \frac{1}{2\sqrt{y}}\left\{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y}{2}\right) + \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y}{2}\right)\right\} \\ &= \frac{1}{\sqrt{2}\sqrt{y}\sqrt{\pi}}\exp\left(-\frac{y}{2}\right) \\ &= \frac{y^{\frac{1}{2}-1}}{\Gamma(1/2)}\left(\frac{1}{2}\right)^{1/2}\exp\left(-\frac{y}{2}\right), \end{aligned}$$

since $\Gamma(1/2) = \sqrt{\pi}$. This is the pdf of a gamma$(1/2, 1/2) \equiv \chi^2_1$ distribution, as claimed. □

Putting together Proposition 6.6 and Equation (6.2) gives us a way to construct a random variable with any chi-squared distribution, given a supply of independent and identically distributed standard normal random variables:

Proposition 6.7. *Let $Z_1, Z_2, \ldots Z_n$ be independent and identically distributed, with $Z_i \sim N(0,1)$, and let $Y = \sum_{i=1}^n Z_i^2$. Then $Y \sim \chi^2_n$.*

Proof. By Proposition 6.6, each $Z_i^2 \sim \chi^2_1$, and $Z_1^2, \ldots, Z_n^2$ are independent. So by Equation (6.2) with $n_i = 1$, $Y \sim \chi^2_{\sum_{i=1}^n 1} \equiv \chi^2_n$, as claimed. □

6.4 Distribution of the sample variance

Suppose $Y_1, Y_2, \ldots, Y_n$ are independent identically distributed $N(\mu, \sigma^2)$ random variables. We can use observations of $y_1, y_2, \ldots, y_n$ of $Y_1, \ldots, Y_n$ to estimate μ and σ^2. We can use the sample mean

$$\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$$

to estimate μ, and the sample variance

$$s^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}$$

to estimate σ^2.

For any specific sample of observations, $\bar{y}$ and s^2 will take specific numerical values, but when they are defined in terms of the random variables $Y_1, Y_2, \ldots, Y_n$, they too are random variables and will have distributions.

We have already seen (in Proposition 4.1) that $\bar{Y} \sim N(\mu, \sigma^2/n)$. We would like to also know the distribution of the random version of the sample variance

$$S^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1}. \tag{6.3}$$

Theorem 6.1. *$Y_1, Y_2, \ldots, Y_n$ are independent identically distributed $N(\mu, \sigma^2)$ random variables, and let S^2 be as defined in Equation (6.3). Then*

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1},$$

and S^2 and $\bar{Y}$ are independent.

In order to prove this important result, we will need to use Cochran's theorem, which we state here without proof.

Theorem 6.2 (Cochran's Theorem)**.** *Suppose that $U \sim$ gamma(α_1, β) and $W \sim$ gamma$(\alpha_1 + \alpha_2, \beta)$. If $V = W - U$, then any one of the following implies the other two:*

i) $V \sim$ gamma(α_2, β).
ii) V is independent of U
iii) V is non-negative.

We use Cochran's theorem to prove Theorem 6.1:

Theorem ref(thm:S2dist). The key to the proof is to write $\sum_{i=1}^n (Y_i - \bar{Y})^2$ in terms of $\sum_{i=1}^n (Y_i - \mu)^2$. We have

$$\begin{aligned}
\sum_{i=1}^n (Y_i - \mu)^2 &= \sum_{i=1}^n (Y_i - \bar{Y} + \bar{Y} - \mu)^2 \\
&= \sum_{i=1}^n \left\{ (Y_i - \bar{Y})^2 + 2(Y_i - \bar{Y})(\bar{Y} - \mu) + (\bar{Y} - \mu)^2 \right\} \\
&= \sum_{i=1}^n (Y_i - \bar{Y})^2 + 2(\bar{Y} - \mu) \sum_{i=1}^n (Y_i - \bar{Y}) + n(\bar{Y} - \mu)^2 \\
&= \sum_{i=1}^n (Y_i - \bar{Y})^2 + n(\bar{Y} - \mu)^2,
\end{aligned}$$

as $\sum_{i=1}^n (Y_i - \bar{Y}) = 0$. So

$$\sum_{i=1}^n (Y_i - \mu)^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2 - n(\bar{Y} - \mu)^2.$$

Multiplying through by $\frac{1}{\sigma^2}$ gives

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i-\mu)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i-\bar{Y})^2 - \frac{n}{\sigma^2}(\bar{Y}-\mu)^2,$$

so

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{Y_i-\mu}{\sigma}\right)^2 - \left(\frac{\sqrt{n}(\bar{Y}-\mu)}{\sigma}\right)^2.$$

But $(Y_i-\mu)/\sigma \sim N(0,1)$, so

$$\sum_{i=1}^{n}\left(\frac{Y_i-\mu}{\sigma}\right)^2 \sim \chi^2_n$$

by Proposition 6.7. We have $\sqrt{n}(\bar{Y}-\mu)/\sigma \sim N(0,1)$ by Proposition 4.1, so

$$\left(\frac{\sqrt{n}(\bar{Y}-\mu)}{\sigma}\right)^2 \sim \chi^2_1$$

by Proposition 6.6.

Now we use Cochran's Theorem (6.2), with

$$\begin{aligned} V &= \frac{(n-1)S^2}{\sigma^2} \\ W &= \sum_{i=1}^{n}\left(\frac{Y_i-\mu}{\sigma}\right)^2 \sim \chi^2_n \equiv \text{gamma}(n/2,1/2) \\ U &= \left(\frac{\sqrt{n}(\bar{Y}-\mu)}{\sigma}\right)^2 \sim \chi^2_1 \equiv \text{gamma}(1/2,1/2), \end{aligned}$$

so $\alpha_1 = 1/2$ and $\alpha_1+\alpha_2 = n/2$, so $\alpha_2 = (n-1)/2$. We have $\sum_{i=1}^{n}(Y_i-\bar{Y})^2 \geq 0$, so $V \geq 0$, so condition (iii) in Cochran's Theorem is met.

So by (ii) of Cochran's Theorem, $V = \frac{(n-1)S^2}{\sigma^2}$ is independent of $U = \left(\frac{\sqrt{n}(\bar{Y}-\mu)}{\sigma}\right)^2$, and hence S^2 is independent of $\bar{Y}$, as claimed.

By (i) of Cochran's Theorem,

$$V = \frac{(n-1)S^2}{\sigma^2} \sim \text{gamma}(\alpha_2,1/2) \equiv \text{gamma}((n-1)/2,1/2) \equiv \chi^2_{n-1},$$

as claimed. □

A consequence of Theorem 6.1 is that

$$\begin{aligned} E(S^2) &= E\left(\frac{\sigma^2}{n-1}\frac{(n-1)S^2}{\sigma^2}\right) \\ &= \frac{\sigma^2}{n-1}E\left(\frac{(n-1)S^2}{\sigma^2}\right) \\ &= \frac{\sigma^2}{n-1}\cdot(n-1) = \sigma^2, \end{aligned}$$

where we have used that the expected value of a chi-squared random variable is its degrees of freedom. So S^2 is an *unbiased* estimator of σ^2: a concept which we will revisit in Chapter 10.

Chapter 7

Univariate transformations

7.1 Transformed random variables

In this chapter, starting with a random variable Y, we define a new random variable $U = U(Y)$. We want to be able to determine the distribution of the transformed random variable U.

We have already studied one method for finding the distribution of a transformed random variable, when that transformation is linear ($U = a + bY$ for suitable constants a and b). In that case, if we have the moment generating function of Y, we can use Theorem 3.2 to find the moment generating function of U.

However, we are often interested in non-linear transformations. For example, we have seen (in Proposition 6.6) an example where we squared a $N(0, 1)$ random variables to obtain a new random variable with χ^2_1 distribution, i.e. we had a transformation of the form $U = Y^2$.

We begin with two simple examples.

Example 7.1. Suppose Y has pdf

$$f(y) = 2y, \quad \text{for } 0 < y < 1.$$

Let $U = Y^2$. What is the distribution of U?

Let $G(.)$ and $g(.)$ be the cdf and pdf of U. We have

$$G(u) = P(U \leq u) = P(Y^2 \leq u) = P(Y \leq \sqrt{u}) = F(\sqrt{u}).$$

So

$$g(u) = \frac{dG(u)}{du} = \frac{dF(\sqrt{u})}{du} = f(\sqrt{u})\frac{1}{2\sqrt{u}}$$
$$= 2\sqrt{u}\frac{1}{2\sqrt{u}} = 1 \quad \text{for } 0 < u < 1,$$

which is the pdf of a $U(0,1)$ distribution. So $U \sim U(0,1)$.

Example 7.2. Suppose $Y \sim U(0,1)$. Let $U = -\log Y$. What is the distribution of U?

Let $G(.)$ and $g(.)$ be the cdf and pdf of U. We have

$$\begin{aligned} G(u) &= P(U \leq u) = P(-\log Y \leq u) \\ &= P(\log Y \geq -u) = P(Y \geq e^{-u}) \\ &= 1 - P(Y < e^{-u}) = 1 - F(e^{-u}) \\ &= 1 - e^{-u}, \quad \text{for } 0 < u < \infty. \end{aligned}$$

So

$$g(u) = \frac{dG(u)}{du} = e^{-u}, \quad \text{for } u > 0,$$

which we recognise as the pdf of an exponential distribution with rate parameter 1, so $U \sim$ exponential(1).

7.2 One-to-one transformations of continuous random variables

Theorem 7.1. *If U is a one-to-one transformation of a continuous random variable Y, the pdf of U is*

$$g(u) = f(w(u))\left|\frac{dy}{du}\right|$$

where $Y = w(U)$ is the inverse transformation.

Remark. In Example 7.1, $U = Y^2$, so the inverse transformation is $Y = U^{1/2} = w(U)$. In Example 7.2, $U = -\log Y$, so the inverse transformation is $Y = \exp(-U) = w(U)$.
Remark. If the domain of Y is (a, b), then the domain of U is all values of u such that $a < w(u) < b$ (where a and b might be infinite).

Theorem ref(thm:unitrans). Let $F(.)$ and $f(.)$ be the cdf and pdf of Y, and let $G(.)$ and $g(.)$ be the cdf and pdf of U. Suppose $U = h(Y)$ is a one-to-one function with inverse $Y = w(U)$. Since $h(.)$ is one-to-one, we may conclude that is is either an increasing or a decreasing function, and we split the proof into two cases.

Case 1: $h(.)$ is an increasing function. Then

$$G(u) = P(U \leq u) = P(h(Y) \leq u) = P(Y \leq w(u)),$$

since $h(.)$ is increasing. So $G(u) = F(w(u))$, and

$$g(u) = f\left(w(u)\right) \frac{dw(u)}{du} = f\left(w(u)\right) \frac{dy}{du} = f\left(w(u)\right) \left|\frac{dy}{du}\right|,$$

since $\frac{dy}{du} > 0$.

Case 2: $h(.)$ is a decreasing function. Then

$$G(u) = P(U \leq u) = P(h(Y) \leq u) = P(Y \geq w(u)),$$

since $h(.)$ is decreasing. So $G(u) = 1 - F(w(u))$, and

$$g(u) = -f\left(w(u)\right) \frac{dw(u)}{du} = f\left(w(u)\right) \left(-\frac{dy}{du}\right) = f\left(w(u)\right) \left|\frac{dy}{du}\right|,$$

since $\frac{dy}{du} < 0$. □

Example 7.3 (Kinetic energy of gas molecules)**.** Gas molecules move about with varying velocity which has, according to the Maxwell-Boltzmann law, a pdf given by

$$f(v) = cv^2 e^{-\beta v^2}, \quad v > 0,$$

where c is a constant to ensure the pdf integrates to 1. The kinetic energy is given by $U = \frac{1}{2}mV^2$, where m is the mass of the molecule. What is the pdf of U, the kinetic energy of the molecule?

The domain of U is $0 < u < \infty$. For $v > 0$, $h(v) = \frac{1}{2}mv^2$ is a one-to-one function, as it is increasing with v. So we may apply Theorem 7.1 to find the pdf $g(.)$ of U.

We have $u = \frac{1}{2}mv^2$, so the inverse transformation is $v = \sqrt{\frac{2u}{m}}$, and

$$\frac{dv}{du} = \frac{\sqrt{2}}{2\sqrt{um}} = \frac{1}{\sqrt{2um}} > 0 \quad \text{for } u > 0.$$

So U has pdf

$$\begin{aligned} g(u) &= f\left(\sqrt{\frac{2u}{m}}\right)\frac{1}{\sqrt{2um}} \\ &= c\cdot\frac{2u}{m}e^{-\beta\frac{2u}{m}}\frac{1}{\sqrt{2um}} \\ &= c\sqrt{2u}m^{-\frac{3}{2}}e^{-\beta\frac{2u}{m}} \quad \text{for } u>0. \end{aligned}$$

We see that

$$g(u) \propto u^{\frac{1}{2}}e^{-\beta\frac{2u}{m}} = u^{\frac{3}{2}-1}e^{-\theta u}$$

with $\theta = 2\beta/m$. So, by comparison with Equation (6.1), $U \sim \text{gamma}(3/2, 2\beta/m)$.

7.3 Generating samples from any distribution

Proposition 7.1. *Suppose that Y is a continuous random variable with cumulative distribution function $F(.)$. Let $U = F(Y)$. Then $U \sim U(0,1)$.*

Proof. U has cdf

$$\begin{aligned} G(u) &= P(U \le u) = P(F(Y) \le u) \\ &= P(Y \le F^{-1}(u)) \quad \text{since } F(.) \text{ is increasing} \\ &= F\left(F^{-1}(u)\right) = u \quad \text{for } 0<u<1, \end{aligned}$$

which is the cdf of a $U(0,1)$ distribution. So $U \sim U(0,1)$, as required. □

An immediate corollary of Proposition 7.1 allows us to use a sample from a $U(0,1)$ distribution to generate a sample from any other continuous distribution:

Corollary 7.1. *Let $F(.)$ be the cumulative distribution function of a continuous distribution, with inverse $F^{-1}(.)$. Suppose that $U \sim U(0,1)$. Then $Y = F^{-1}(U)$ has cumulative distribution function $F(.)$.*

Example 7.4. Suppose that you are given $U \sim N(0,1)$, and would like to generate $Y \sim$ exponential(2). The exponential(2) distribution has cdf

$$F(y) = 1 - e^{-2y}, \quad \text{for } y>0,$$

and inverting (solving $u = F(y)$ for y) gives

$$F^{-1}(u) = -\frac{1}{2}\log(1-u).$$

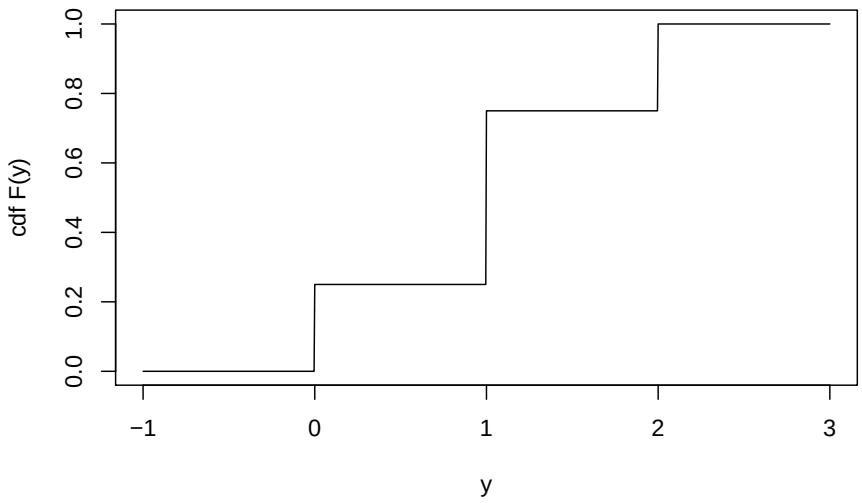
So

$$Y = -\frac{1}{2}\log(1-u) \sim \text{exponential}(2).$$

In fact, we can use the same method to find samples from a discrete distribution. To do this, we first need to define what we mean by $F^{-1}(.)$, as the cdf $F(.)$ is constant between points in the domain, and so is not an invertible function.

For instance, if $Y \sim \text{binomial}(2, 0.5)$, then we can plot the cdf of Y:

```
curve(pbinom(x, 2, 0.5), from = -1, to = 3,
      xlab = "y", ylab = "cdf F(y)", n = 1001)
```



In this example, if we ask for $F^{-1}(0.25)$, then it is not immediately clear what this means, as $F(y) = 0.25$ for all $0 \leq y < 1$.

We define $F^{-1}(.)$ by the *quantile* function

$$F^{-1}(u) = \inf\{y \in \mathbb{R} : F(y) \geq u\}, \tag{7.1}$$

for $u \in (0, 1)$. In our example, we now have $F^{-1}(0.25) = 0$. If $F(.)$ is an invertible function, then this definition agrees with the usual inverse we saw before.

Equipped with this definition of $F^{-1}(.)$, we may now rewrite Corollary 7.1 to apply to any distribution:

Corollary 7.2. *Let $F(.)$ be a cumulative distribution function, and let $F^{-1}(.)$ be the corresponding quantile function, defined by (7.1). Suppose that $U \sim U(0, 1)$. Then $Y = F^{-1}(U)$ has cumulative distribution function $F(.)$.*

Example 7.5 (Generating a binomial random variable)**.** Suppose that you are given $U \sim N(0,1)$, and would like to generate $Y \sim$ binomial$(2, 0.5)$. The binomial(2) distribution has cdf

$$F(y) = \begin{cases} 0 & \text{for } y < 0 \\ 0.25 & \text{for } 0 \leq y < 1 \\ 0.75 & \text{for } 1 \leq y < 2 \\ 1 & \text{for } y > 1. \end{cases}$$

The corresponding quantile function is

$$F^{-1}(u) = \begin{cases} -\infty & \text{for } u \leq 0 \\ 0 & \text{for } 0 < u \leq 0.25 \\ 1 & \text{for } 0.25 < u \leq 0.75 \\ 2 & \text{for } 0.75 < u \leq 1 \\ \infty & \text{for } u > 1. \end{cases}$$

So $Y = F^{-1}(U) \sim$ binomial$(2, 0.5)$.

Chapter 8

Bivariate distributions

8.1 Joint distributions

There are many situations where random variables vary simultaneously, for example:

- height and weight of individuals in a population
- systolic and diastolic blood pressure of individuals in a population
- value of Sterling and the Euro in US Dollars today at 12.00 GMT

In these cases and in many other situations, the variables are not independent so we need to consider their *joint behaviour*.

Under some circumstances it might be possible to assume or to deduce that the variables do not depend on each other, i.e. that they are independent. We need to define the joint probabilistic behaviour of two random variables. We could define these terms for either discrete or continuous random variables. However, we give the definitions in terms of continuous random variables with obvious extensions to the discrete or other cases.

We could generalise these results when we have several random variables (giving so-called multivariate models) but here we shall concentrate on the simplest case of two random variables (the bivariate case).

Suppose Y_1 and Y_2 vary together with joint probability density function $f(y_1, y_2)$. The function f has the following properties:

1. $f(y_1, y_2) \geq 0$ for all y_1, y_2.
2. $\int_{a_2}^{b_2} \int_{a_1}^{b_1} f(y_1, y_2) dy_1 dy_2 = P(a_1 < Y_1 \leq b_1 \text{ and } a_2 < Y_2 \leq b_2)$.

An immediate corollary is that $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(y_1, y_2) dy_1 dy_2 = 1$.

We will give some examples shortly, but first we set up some more functions of interest.

The *marginal* probability density function of Y_1 is given by integrating out Y_2 , i.e.

$$f_1(y_1) = \int_{-\infty}^{\infty} f(y_1, y_2) dy_2.$$

This essentially gives the probabilistic behaviour of Y_1 ignoring Y_2. Similarly, the marginal pdf of Y_2 is

$$f_2(y_2) = \int_{-\infty}^{\infty} f(y_1, y_2) dy_1.$$

We define the *conditional* probability density function of Y_2 given that $Y_1 = y_1$ as

$$f(y_2|y_1) = \frac{f(y_1, y_2)}{f_1(y_1)},$$

assuming that $f_1(y_1) > 0$.

If $f(y_1, y_2) = f_1(y_1) f_2(y_2)$ for all y_1 and y_2, then Y_1 and Y_2 are said to be *independent*. In that case, $f(y_2|y_1) = f_2(y_2)$, for all y_2, and all y_1 with $f_1(y_1) > 0$, which is an equivalent definition of independence.

Example 8.1. Suppose that Y_1 and Y_2 have joint pdf

$$f(y_1, y_2) = 14, \quad \text{for } -1 < y_1 < 1 \text{ and } -1 < y_2 < 1$$

. We now derive the marginal and conditional pdfs.

The marginal pdf of Y_1 is

$$f_1(y_1) = \int_{-\infty}^{\infty} f(y_1, y_2) dy_2 = \int_{-1}^{1} \frac{1}{4} dy_2 = \frac{1}{2} \quad \text{for } -1 < y_1 < 1,$$

so $Y_1 \sim U(-1, 1)$. By symmetry, $Y_2 \sim U(-1, 1)$.

The conditional pdf of $Y_2|Y_1 = y_1$ is

$$\begin{aligned} f(y_2|y_1) &= \frac{f(y_1, y_2)}{f_1(y_1)} \quad \text{for } -1 < y_1 < 1, -1 < y_2 < 1 \\ &= \frac{1/4}{1/2} = \frac{1}{2}. \end{aligned}$$

Hence if $-1 < y_1 < 1$, then $Y_2|Y_1 = y_1 \sim U(-1, 1)$. Knowing the value of Y_1 does not change the distribution of Y_2. This means that Y_1 and Y_2 are independent.

Example 8.2. Suppose that Y_1 and Y_2 have joint pdf

$$f(y_1, y_2) = \frac{1}{\pi}, \quad \text{for } y_1^2 + y_2^2 < 1.$$

We now derive the marginal and conditional pdfs.

The marginal pdf of Y_1 is

$$\begin{aligned} f_1(y_1) &= \int_{-\infty}^{\infty} f(y_1, y_2) dy_2 \\ &= \int_{-\sqrt{1-y_1^2}}^{\sqrt{1-y_1^2}} \frac{1}{\pi} dy_2 \\ &= \frac{2}{\pi}\sqrt{1-y_1^2}, \quad \text{for } -1 < y_1 < 1. \end{aligned}$$

Similarly, the marginal pdf of Y_2 is

$$f_2(y_2) = \frac{2}{\pi}\sqrt{1-y_2^2}, \quad \text{for } -1 < y_2 < 1.$$

The conditional pdf of $Y_2|Y_1 = y_1$ is

$$f(y_2|y_1) = \frac{1/\pi}{2\sqrt{1-y_1^2}/\pi} = \frac{1}{2\sqrt{1-y_1^2}} \quad \text{for } -\sqrt{1-y_1^2} < y_2 < \sqrt{1-y_1^2},$$

provided that $-1 < y_1 < 1$, so $Y_2|Y_1 = y_1 \sim U(-\sqrt{1-y_1^2}, \sqrt{1-y_1^2})$. Knowing that $Y_1 = y_1$ gives us information about the behaviour of Y_2. This means that Y_1 and Y_2 are not independent, as $f(y_2|y_1) \neq f(y_2)$.

8.2 Moments of jointly distributed random variables

For any general function $g(Y_1, Y_2)$ of Y_1 and Y_2, the expectation of $g(Y_1, Y_2)$ is defined as

$$E\{g(Y_1, Y_2)\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(y_1, y_2) f(y_1, y_2) dy_1 dy_2.$$

Applying this with $g(Y_1, Y_2) = Y_1$, we have

$$E(Y_1) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y_1 f(y_1, y_2) dy_1 dy_2,$$

and similarly

$$E(Y_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y_2 f(y_1, y_2) dy_1 dy_2.$$

Since $f(y_1, y_2) = f_1(y_1) f(y_2|y_1)$,

$$\begin{aligned} E(Y_1) &= \int_{-\infty}^{\infty} y_1 f(y_1) \left\{ \int_{-\infty}^{\infty} f(y_2|y_1) dy_2 \right\} dy_1 \\ &= \int_{-\infty}^{\infty} y_1 f_1(y_1) dy_1, \end{aligned}$$

which is our usual definition of the expected value of a single random variable with (marginal) pdf $f_1(.)$.

In general

$$E\left\{g(Y_1)\right\} = \int_{-\infty}^{\infty} g(y_1) f_1(y_1) dy_1,$$

and

$$E\left\{g(Y_2)\right\} = \int_{-\infty}^{\infty} g(y_2) f_2(y_2) dy_2.$$

Letting $g(Y_1, Y_2) = Y_1 Y_2$, we have

$$\begin{aligned} E(Y_1 Y_2) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y_1 y_2 f(y_1, y_2) dy_1 dy_2 \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y_1 y_2 f_1(y_1) f(y_2|y_1) dy_1 dy_2 \\ &= \int_{-\infty}^{\infty} y_1 f_1(y_1) \int_{-\infty}^{\infty} y_2 f(y_2|y_1) dy_2 dy_1. \end{aligned}$$

This involves the *conditional expectation* of $Y_2|Y_1 = y_1$,

$$E(Y_2|Y_1 = y_1) = \int_{-\infty}^{\infty} y_2 f(y_2|y_1) dy_2.$$

If Y_1 and Y_2 are *independent*,

$$\begin{aligned} E(Y_1 Y_2) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y_1 y_2 f_1(y_1) f_2(y_2) dy_1 dy_2 \\ &= \int_{-\infty}^{\infty} y_1 f_1(y_1) dy_1 \int_{-\infty}^{\infty} y_2 f_2(y_2) dy_2 \\ &= E(Y_1) E(Y_2). \end{aligned}$$

If Y_1 and Y_2 are independent, then this relationship holds. It might also hold in special circumstances even when the variables are not independent, for example when both $E(Y_1Y_2)$ and $E(Y_1)$ are zero.

Definition 8.1. The *covariance* of Y_1 and Y_2 is

$$\mathrm{Cov}(Y_1, Y_2) = E\left\{[Y_1 - E(Y_1)][Y_2 - E(Y_2)]\right\}.$$

We may rewrite the expression for the covariance as

$$\begin{aligned}\mathrm{Cov}(Y_1, Y_2) &= E\left\{[Y_1 - E(Y_1)][Y_2 - E(Y_2)]\right\} \\ &= E\left\{Y_1Y_2 - E(Y_1)Y_2 - E(Y_2)Y_1 + E(Y_1)E(Y_2)\right\} \\ &= E(Y_1Y_2) - E(Y_1)E(Y_2) - E(Y_2)E(Y_1) + E(Y_1)E(Y_2) \\ &= E(Y_1Y_2) - E(Y_1)E(Y_2).\end{aligned}$$

The covariance of a variable with itself is

$$\mathrm{Cov}(Y_1, Y_1) = E\left\{[Y_1 - E(Y_1)]^2\right\} = \mathrm{Var}(Y_1).$$

Definition 8.2. The *correlation* of Y_1 and Y_2 is

$$\mathrm{Corr}(Y_1, Y_2) = \frac{\mathrm{Cov}(Y_1, Y_2)}{\sqrt{\mathrm{Var}(Y_1)\mathrm{Var}(Y_2)}}.$$

Example 8.3. Returning to Example 8.2 with $f(y_1, y_2) = 1/\pi$ for $y_1^2 + y_2^2 < 1$, we already showed that Y_1 and Y_2 are not independent. By symmetry, we have $E(Y_1) = E(Y_2) = E(Y_2|Y_1 = y_1) = 0$, so

$$\mathrm{Cov}(Y_1, Y_2) = E(Y_1Y_2) - E(Y_1)E(Y_2) = E(Y_1Y_2).$$

Now

$$\begin{aligned}
E(Y_1Y_2) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y_1y_2f(y_1,y_2)dy_2dy_1 \\
&= \frac{1}{\pi}\int_{-1}^{1}\int_{-\sqrt{1-y_1^2}}^{\sqrt{1-y_1^2}} y_1y_2dy_2dy_1 \\
&= \frac{1}{\pi}\int_{-1}^{1} y_1\left(\int_{-\sqrt{1-y_1^2}}^{\sqrt{1-y_1^2}} y_2dy_2\right)dy_1 \\
&= \frac{1}{\pi}\int_{-1}^{1} y_1\left[\frac{y_2^2}{2}\right]_{-\sqrt{1-y_1^2}}^{\sqrt{1-y_1^2}} dy_1 \\
&= \frac{1}{\pi}\int_{-1}^{1} y_1\left[\frac{y_2^2}{2}\right]_{-\sqrt{1-y_1^2}}^{\sqrt{1-y_1^2}} dy_1 \\
&= \frac{1}{\pi}\int_{-1}^{1} y_1\times 0dy_1 \\
&= 0,
\end{aligned}$$

so $\mathrm{Cov}(Y_1,Y_2)=0$, even though Y_1 and Y_2 are **not** independent. This example shows that even though

$$Y_1 \text{ and } Y_2 \text{ independent } \Rightarrow \mathrm{Cov}(Y_1,Y_2)=0,$$

in general the reverse does not hold, so

$$\mathrm{Cov}(Y_1,Y_2)=0 \not\Rightarrow Y_1 \text{ and } Y_2 \text{ independent.}$$

Proposition 8.1. *For any two random variables Y_1 and Y_2*

$$Var(Y_1+Y_2) = Var(Y_1) + Var(Y_2) + 2Cov(Y_1,Y_2).$$

Proof. We have

$$\mathrm{Var}(Y_1+Y_2) = E\left[(Y_1+Y_2)^2\right] - \left[E(Y_1+Y_2)\right]^2,$$

where

$$\begin{aligned}
E\left[(Y_1+Y_2)^2\right] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(y_1+y_2)^2f(y_1,y_2)dy_1dy_2 \\
&= E(Y_1^2) + 2E(Y_1Y_2) + E(Y_2^2)
\end{aligned}$$

and

$$E(Y_1+Y_2) = E(Y_1) + E(Y_2).$$

So

$$\begin{aligned}\mathrm{Var}(Y_1+Y_2) &= E(Y_1^2)+2E(Y_1Y_2)+E(Y_2^2)-[E(Y_1)+E(Y_2)]^2\\ &= E(Y_1^2)-E(Y_1)^2+E(Y_2^2)-E(Y_2)^2+2\left[E(Y_1Y_2)-E(Y_1)E(Y_2)\right]\\ &= \mathrm{Var}(Y_1)+\mathrm{Var}(Y_2)+2\mathrm{Cov}(Y_1,Y_2)\end{aligned}$$

as claimed. □

8.3 The bivariate normal distribution

Definition 8.3. The random variables Y_1 and Y_2 are said to have a *bivariate normal distribution* if they have joint probability density function

$$f(y_1,y_2)=(2\pi)^{-1}\det(\boldsymbol{\Sigma})^{-1/2}\exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right\},\quad \text{for } y_1,y_2\in\mathbb{R},$$

where we write $\boldsymbol{y}=(y_1,y_2)^T$, and where $\boldsymbol{\mu}=(\mu_1,\mu_2)^T$ is a vector of means, and $\boldsymbol{\Sigma}$ is a 2×2 symmetric positive definite matrix. We write $\boldsymbol{Y}=(Y_1,Y_2)^T\sim N_2(\boldsymbol{\mu},\boldsymbol{\Sigma})$.

We often write out the elements of the 2×2 matrix $\boldsymbol{\Sigma}$ as

$$\boldsymbol{\Sigma}=\begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}. \tag{8.1}$$

The marginal pdf of Y_1 is

$$f_1(y_1)=\int_{-\infty}^{\infty} f(y_1,y_2)dy_2,$$

which reduces to

$$f_1(y_1)=\frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left\{-\frac{1}{2\sigma_1^2}(y_1-\mu_1)^2\right\},$$

so marginally $Y_1\sim N(\mu_1,\sigma_1^2)$. Similarly $Y_2\sim N(\mu_2,\sigma_2^2)$.

We may interpret the parameters of the bivariate normal distribution, as $\mu_1=E(Y_1)$, $\mu_2=E(Y_2)$, $\sigma_1^2=\mathrm{Var}(Y_1)$, $\sigma_2^2=\mathrm{Var}(Y_2)$. We may also show that $\mathrm{Cov}(Y_1,Y_2)=\rho\sigma_1\sigma_2$, so $\rho=\mathrm{Corr}(Y_1,Y_2)$.

The conditional distribution of Y_1 given $Y_2=y_2$ is

$$f(y_1|y_2)=\frac{f(y_1,y_2)}{f_2(y_2)},$$

which reduces to

$$f(y_1|y_2) = \frac{1}{\sqrt{2\pi\sigma_1^2(1-\rho)^2}} \exp\left\{-\frac{1}{2\sigma_1^2(1-\rho)^2}\left(y_1 - \mu_1 - \frac{\rho\sigma_1(y_2-\mu_2)}{\sigma_2}\right)^2\right\}$$

This means that

$$Y_1|Y_2 = y_2 \sim N\left(\mu_1 + \frac{\rho\sigma_1(y_2-\mu_2)}{\sigma_2}, \sigma_1^2(1-\rho)^2\right).$$

If Y_1 and Y_2 are uncorrelated ($\rho = 0$), knowing the value of Y_2 does not change the distribution of Y_1, so Y_1 and Y_2 are independent. If Y_1 and Y_2 are correlated ($\rho \neq 0$), the distribution of $Y_1|Y_2 = y_2$ is different from the distribution of Y_1.

8.4 Bivariate moment generating functions

Definition 8.4. The moment generating function of the bivariate distribution of Y_1, Y_2 is

$$M_{Y_1,Y_2}(t_1, t_2) = E\left\{\exp(t_1Y_1 + t_2Y_2)\right\}$$

As in the univariate case, the moment generating function is useful for proving properties about what happens to the distribution of random variables under linear transformations.

Example 8.4 (Bivariate normal mgf)**.** If $\boldsymbol{Y} = (Y_1, Y_2)^T \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then

$$M_{Y_1,Y_2}(t_1, t_2) = \exp(\boldsymbol{\mu}^T\boldsymbol{t} + \frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}),$$

where $\boldsymbol{t} = (t_1, t_2)^T$.

Now let $X_1 = aY_1 + bY_2$, where a and b are given constants. Then

$$\begin{aligned} M_{X_1}(t) &= E\left\{t(aY_1 + bY_2)\right\} \\ &= M_{Y_1,Y_2}(at, bt) \\ &= \exp\left\{(a\mu_1 + b\mu_2)t + \frac{1}{2}(a^2\sigma_1^2 + 2ab\rho\sigma_1\sigma_2 + b^2\sigma_2^2)t^2\right\}, \end{aligned}$$

where we have used the components of $\boldsymbol{\Sigma}$ as in Equation (8.1). So

$$X_1 \sim N(\mu_1 + \mu_2, a^2\sigma_1^2 + 2ab\rho\sigma_1\sigma_2 + b^2\sigma_2^2).$$

With $a = 1$ and $b = 1$,

$$Y_1 + Y_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + 2\rho\sigma_1\sigma_2 + \sigma_2^2).$$

With $a = 1$ and $b = -1$,

$$Y_1 - Y_2 \sim N(\mu_1 - \mu_2, \sigma_1^2 - 2\rho\sigma_1\sigma_2 + \sigma_2^2).$$

8.5 A useful property of covariances

Theorem 8.1. *Suppose V_i, $i = 1, \ldots, m$ and W_j, $j = 1, \ldots, n$ are random variables, and a_i, $i = 1, \ldots, m$ and $j = 1, \ldots, n$ are constants. Then*

$$Cov\left(\sum_{i=1}^{m} a_i V_i, \sum_{j=1}^{n} b_j W_j\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j \, Cov(V_i, W_j).$$

Proof.

$$\begin{aligned}
\mathrm{Cov}\left(\sum_{i=1}^{m} a_i V_i, \sum_{j=1}^{n} b_j W_j\right) &= E\left\{\left[\sum_{i=1}^{m} a_i V_i - E\left(\sum_{i=1}^{m} a_i V_i\right)\right]\left[\sum_{j=1}^{n} b_j W_j - E\left(\sum_{j=1}^{n} b_j W_j\right)\right]\right\} \\
&= E\left\{\left[\sum_{i=1}^{m} a_i \left(V_i - E(V_i)\right)\right]\left[\sum_{j=1}^{n} b_j \left(W_j - E(W_j)\right)\right]\right\} \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j E\left\{\left[V_i - E(V_i)\right]\left[W_j - E(W_j)\right]\right\} \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j \mathrm{Cov}(V_i, W_j),
\end{aligned}$$

as required. □

Example 8.5. Continuing from Example 8.4, we consider $X_1 = Y_1 + Y_2$ and $X_2 = Y_1 - Y_2$. Since X_1 and X_2 are normally distributed, we know that X_1 and X_2 are independent if $\mathrm{Cov}(X_1, X_2) = 0$. Applying Theorem 8.1 with $m = n = 2$, $V_i = W_i = Y_i$, $a_1 = a_2 = 1$, $b_1 = 1$ and $b_2 = -1$, we get

$$\mathrm{Cov}(X_1, X_2) = \mathrm{Cov}(Y_1, Y_1) - \mathrm{Cov}(Y_1, Y_2) + \mathrm{Cov}(Y_1, Y_2) - \mathrm{Cov}(Y_2, Y_2) = \sigma_1^2 - \sigma_2^2.$$

So X_1 and X_2 are independent if $\sigma_1^2 = \sigma_2^2$.

Chapter 9

Bivariate transformations

9.1 The transformation theorem

In Chapter 7 we considered transformations of a single random variable. In this chapter we will generalise to the case of transforming two random variables. As examples we will derive several important distributions distributions – the beta, Cauchy, t and F distributions.

We have already seen in Theorem 7.1 how to find the pdf of a one-to-one transformation of a random variable. We extend this result to find the pdf for a transformation of two random variables.

Theorem 9.1. *Suppose Y_1 and Y_2 have joint probability density function $f(y_1, y_2)$, and that we transform to two new variables $U_1 = U_1(Y_1, Y_2)$ and $U_2 = U_2(Y_1, Y_2)$ using a one-to-one transformation of (Y_1, Y_2) to (U_1, U_2). Then the joint probability density function of (U_1, U_2) is given by*

$$g(u_1, u_2) = f[w_1(u_1, u_2), w_2(u_1, u_2)] \times \left|\det(\boldsymbol{J})\right|,$$

where

$$\boldsymbol{J} = \begin{pmatrix} \frac{\partial y_1}{\partial u_1} & \frac{\partial y_1}{\partial u_2} \\ \frac{\partial y_2}{\partial u_1} & \frac{\partial y_2}{\partial u_2} \end{pmatrix}$$

is the Jacobian *matrix, and $Y_1 = w_1(U_1, U_2)$ and $Y_2 = w_2(U_1, U_2)$ are the inverse transformations.*

In addition to finding the inverse transformation, and using this in Theorem 9.1, we need to identify the domain of U_1 and U_2. We identify constraints on U_1 and U_2 in two passes, to double check we haven't missed any constraints:

- **forward** pass: plug in constraints on Y_1 and Y_2 directly into the definition of U_1 and U_2.

- **backward** pass: rewrite the constraints on Y_1 and Y_2 in terms of U_1 and U_2, by using the inverse transformation, and rearrange to derive additional constraints on U_1 and U_2.

We will work through several examples to see how this works.

9.2 The beta distribution

Definition 9.1. A random variables Y has a *beta* distribution if it has pdf of the form

$$f(y) = \frac{1}{B(m,n)} y^{m-1}(1-y)^{n-1}, \quad 0 < y < 1,$$

for some parameters n and m, where

$$B(m,n) = \int_0^1 u^{m-1}(1-u)^{n-1} du = \frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)}$$

is the *Beta function*. We write $Y \sim \text{beta}(m,n)$.

We will see one use of the beta distribution in Chapter 12. We may obtain the beta distribution by transforming a pair of gamma random variables with the same rate parameter:

Proposition 9.1. *Suppose that $Y_1 \sim$ gamma(m,β), $Y_2 \sim$ gamma(n,β), and that Y_1 and Y_2 are independent. Then we claim that*

$$U_1 = \frac{Y_1}{Y_1 + Y_2} \sim \textit{beta}(m,n).$$

Proof. To show this, we would like to use Theorem 9.1. In order to do this, we first need to define another random variable U_2, which is another transformation of Y_1 and Y_2. Here we will choose $U_2 = Y_1 + Y_2$, but many other choices would work too. We will derive the joint pdf of U_1 and U_2, then integrate this to find the marginal pdf of U_1, which we hope will be a beta(m,n) pdf.

The joint pdf of Y_1 and Y_2 is

$$f(y_1,y_2) = \frac{\beta^m}{\Gamma(m)} y_1^{m-1} e^{-\beta y_1} \cdot \frac{\beta^n}{\Gamma(n)} y_2^{n-1} e^{-\beta y_2},$$

for $y_1 > 0$, $y_1 > 0$, since Y_1 and Y_2 are independent.

The inverse transformations are

$$Y_1 = U_1 U_2, \quad Y_2 = U_2 - U_1 U_2 = U_2(1-U_1).$$

We know that $Y_1 > 0$ and $Y_2 > 0$. First, we derive the domain of (U_1, U_2):

- Forward pass: $U_1 = Y_1/(Y_1 + Y_2) > 0$, $U_2 = Y_1 + Y_2 > 0$.
- Backward pass: The inverse transformation gives us $Y_1 = U_1U_2 > 0$, which gives us no additional information as we already know $U_1 > 0$ and $U_2 > 0$. We also get $Y_2 = U_2(1-U_1) > 0$, so $U_2 > U_1U_2$, so $U_1 < 1$. We could have already seen this on the forward pass, but the backward pass is useful to catch any constraints we missed on the forwards pass.

The domain is $0 < U_1 < 1$ and $0 < U_2 < \infty$.

The Jacobian is

$$\boldsymbol{J} = \begin{pmatrix} \frac{\partial y_1}{\partial u_1} & \frac{\partial y_1}{\partial u_2} \\ \frac{\partial y_2}{\partial u_1} & \frac{\partial y_2}{\partial u_2} \end{pmatrix} = \begin{pmatrix} u_2 & u_1 \\ -u_2 & 1-u_1 \end{pmatrix},$$

so

$$\det \boldsymbol{J} = u_2(1-u_1) + u_1u_2 = u_2 > 0$$

Therefore, by Theorem 9.1, the joint pdf of U_1 and U_2 is

$$\begin{aligned} g(u_1, u_2) &= f(y_1, y_2) \times \left| \det J \right| \\ &= \frac{\beta^m}{\Gamma(m)} (u_1u_2)^{m-1} e^{-\beta u_1 u_2} \cdot \frac{\beta^n}{\Gamma(n)} \left[u_2(1-u_1)\right]^{n-1} e^{-\beta u_2(1-u_1)} \cdot u_2 \\ &= \frac{1}{\Gamma(m)\Gamma(n)} u_1^{m-1}(1-u_1)^{n-1} \beta^{m+n} u_2^{m+n-1} e^{-\beta u_2}, \end{aligned}$$

for $0 < u_1 < 1$ and $u_2 > 0$. Note that $g(.,.)$ factorises into a term involving only u_1 and a term only involving u_2. This means that U_1 and U_2 are independent.

We could find the marginal pdf of U_1 by integrating $g(u_1, u_2)$ over u_2. In this case the marginal pdf of U_1 can be obtained more simply. Gathering together all terms in $g(u_1, u_2)$ depending on u_2, we find

$$g_2(u_2) \propto u_2^{m+n-1} e^{-\beta u_2}, \quad u_2 > 0,$$

i.e.

$$g_2(u_2) = c u_2^{m+n-1} e^{-\beta u_2}$$

for some constant c which will ensure $g_2(.)$ integrates to one. We recognise this as the form of a gamma$(m+n, \beta)$ distribution, so

$$g_2(u_2) = \frac{\beta^{m+n}}{\Gamma(m+n)} u_2^{m+n-1} e^{-\beta u_2}.$$

So

$$\begin{aligned} g_1(u_1) &= \frac{g(u_1, u_2)}{g_2(u_2)} \quad \text{as } U_1, U_2 \text{ independent} \\ &= \frac{\Gamma(m+n)}{\Gamma(m)\Gamma(n)} u_1^{m-1}(1-u_1)^{n-1} \\ &= \frac{1}{B(m,n)} u_1^{m-1}(1-u_1)^{n-1} \quad 0 < u_1 < 1, \end{aligned}$$

so $U_1 \sim \text{beta}(m+n)$. □

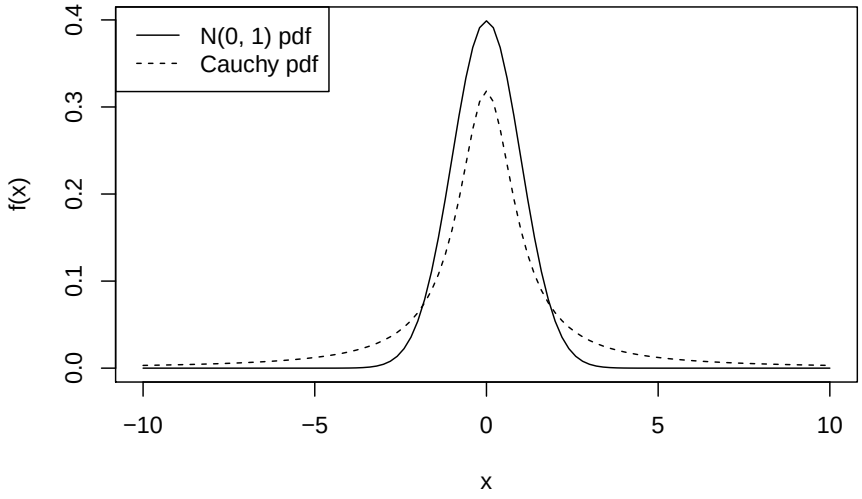
9.3 The Cauchy distribution

Definition 9.2. A random variables Y has *Cauchy* distribution if it has pdf

$$f(y) = \frac{1}{\pi(1+y^2)}, \quad y \in \mathbb{R}.$$

While the Cauchy distribution looks relatively innocuous — it is symmetric around zero, just like a standard Normal distribution — the thickness of its tails means that its moments do not exist (the required integrals do not converge).

```
curve(dnorm(x), from = -10, to = 10, xlab = "x", ylab = "f(x)")
curve(1/(pi*(1 + x^2)), add = TRUE, lty = 2)
legend("topleft", lty = c(1, 2),
       legend = c("N(0, 1) pdf", "Cauchy pdf"))
```



We obtain the Cauchy distribution as the ratio of standard normal random variables:

Proposition 9.2. *Suppose that* $Y_1 \sim N(0,1)$ *and* $Y_2 \sim N(0,1)$ *are independent standard normal random variables. Then* $U_1 = Y_1/Y_2$ *has Cauchy distribution.*

Proof. Again, in order to use Theorem 9.1 to show this, we need to first define a second random variable U_2. We will choose $U_2 = Y_2$.

By independence, the joint pdf of Y_1 and Y_2 is

$$\begin{aligned} f(y_1, y_2) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y_1^2}{2}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y_2^2}{2}\right) \\ &= \frac{1}{2\pi} \exp\left(-\frac{y_1^2 + y_2^2}{2}\right), \quad y_1, y_2 \in \mathbb{R}. \end{aligned}$$

The inverse transformations are

$$Y_1 = U_1 U_2, \quad Y_2 = U_2,$$

and the domain of U_1 and U_2 is clearly $U_1 \in \mathbb{R}$, $U_2 \in \mathbb{R}$.

The Jacobian matrix is

$$\boldsymbol{J} = \begin{pmatrix} u_2 & * \\ 0 & 1 \end{pmatrix},$$

where we do not need to evaluate the top right entry marked $*$, because it will not affect the determinant of $\boldsymbol{J}$. So $\det(\boldsymbol{J}) = u_2$.

So the joint pdf of U_1 and U_2 is

$$\begin{aligned} g(u_1, u_2) &= \frac{1}{2\pi} \exp\left(-\frac{u_1^2 u_2^2 + u_2^2}{2}\right) \cdot |u_2| \\ &= \frac{|u_2|}{2\pi} \exp\left(-\frac{u_2^2(1+u_2^2)}{2}\right) \end{aligned}$$

for $u_1, u_2 \in \mathbb{R}$.

We integrate out u_2 in order to obtain the marginal pdf of U_1

$$\begin{aligned} g_1(u_1) &= \int_{-\infty}^{\infty} g(u_1, u_2) du_2 \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} |u_2| \exp\left(-\frac{u_2^2(1+u_2^2)}{2}\right) du_2 \\ &= 2 \times \frac{1}{2\pi} \int_0^{\infty} u_2 \exp\left(-\frac{c u_2^2}{2}\right) du_2, \text{ with } c = 1 + u_1^2 \\ &= \frac{1}{\pi} \left[-\frac{1}{c} \exp\left(-\frac{c u_2^2}{2}\right)\right]_0^{\infty} \\ &= \frac{1}{\pi} \cdot \frac{1}{c} \\ &= \frac{1}{\pi(1+u_1)^2}, \quad u_1 \in \mathbb{R}, \end{aligned}$$

which is the Cauchy pdf, so U_1 has Cauchy distribution, as required. □

9.4 The t distribution

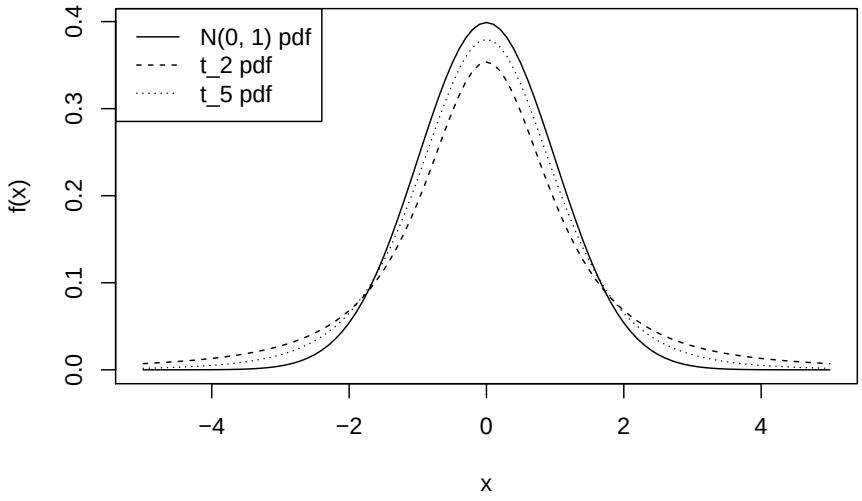
Definition 9.3. A random variable Y has t distribution with k degrees of freedom if it has pdf

$$f(y) = \frac{1}{\sqrt{k} B\left(\frac{1}{2}, \frac{k}{2}\right)} \left(1 + \frac{y^2}{k}\right)^{-\frac{k+1}{2}}, \quad y \in \mathbb{R}.$$

We write $Y \sim t_k$.

We can plot the pdfs of the t distribution with 2, and 5 degrees of freedom, compared with the $N(0, 1)$ pdf.

```r
curve(dnorm(x), from = -5, to = 5, xlab = "x", ylab = "f(x)")
curve(dt(x, df = 2), add = TRUE, lty = 2)
curve(dt(x, df = 5), add = TRUE, lty = 3)
legend("topleft", lty = c(1, 2, 3),
       legend = c("N(0, 1) pdf", "t_2 pdf", "t_5 pdf"))
```



The t_k distribution has heavier tails than the $N(0,1)$ distribution, but as $k \to \infty$, $t_k \to N(0,1)$. When $k = 1$, the t_1 distribution is the Cauchy distribution.

We obtain the t distribution as a ratio of a standard normal random variable, and a the square root of a chi-squared random variable divided by its degrees of freedom. Although this sounds complicated, this makes the t distribution important in practice, as we will soon see.

Proposition 9.3. *Suppose that $Y_1 \sim N(0,1)$ and $Y_2 \sim \chi^2_k$, and that Y_1 and Y_2 are independent. Then*

$$U_1 = \frac{Y_1}{\sqrt{Y_k/k}} \sim t_k.$$

Proof. In order to use Theorem 9.1 to show this, we need to first define a second random variable U_2. We will choose $U_2 = Y_2$.

The pdfs of Y_1 and Y_2 are

$$f_1(y_1) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-y_1^2}{2}\right), \quad y_1 \in \mathbb{R}$$

and

$$f_2(y_2) = \frac{y_2^{k/2-1}}{\Gamma(k/2)2^{k/2}} \exp\left(\frac{-y_2}{2}\right), \quad y_2 > 0.$$

So, by independence, the joint pdf of Y_1 and Y_2 is

$$f(y_1, y_2) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-y_1^2}{2}\right) \frac{y_2^{k/2-1}}{\Gamma(k/2)2^{k/2}} \exp\left(\frac{-y_2}{2}\right), \quad y_1 \in \mathbb{R}, y_2 > 0.$$

The inverse transformations are

$$Y_1 = U_1\sqrt{\frac{U_2}{k}}, \quad Y_2 = U_2,$$

and the domain of U_1 and U_2 is $U_2 \in \mathbb{R}$, $U_2 > 0$.

The Jacobian is

$$\boldsymbol{J} = \begin{pmatrix} \sqrt{\frac{u_2}{k}} & * \\ 0 & 1 \end{pmatrix},$$

so $\det \boldsymbol{J} = \sqrt{u_2/k} > 0$.

So the joint pdf of U_1 and U_2 is

$$\begin{aligned} g(u_1, u_2) &= \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{u_1^2 u_2}{2k}\Big) \frac{y_2^{k/2-1}}{\Gamma(k/2)2^{k/2}} \exp\Big(-\frac{u_2}{2}\Big)\sqrt{\frac{u_2}{k}} \\ &= \frac{{u_2}^{(k-1)/2}}{2^{(k+1)/2}\sqrt{\pi}\,\Gamma(k/2)\sqrt{k}} \exp\Big\{-\frac{u_2}{2}\Big(\frac{u_1^2}{k}+1\Big)\Big\}, \quad u_1 \in \mathbb{R}, u_2 > 0. \end{aligned}$$

We are interested in the marginal pdf of U_1:

$$g_1(u_1) = \frac{1}{2^{(k+1)/2}\sqrt{\pi}\,\Gamma(k/2)\sqrt{k}} \int_0^\infty {u_2}^{(k-1)/2} \exp\Big\{-\frac{u_2}{2}\Big(\frac{u_1^2}{k}+1\Big)\Big\} du_2.$$

The integrand is proportional to a gamma(α, β) pdf

$$h(u_2) = \frac{\beta^\alpha}{\Gamma(\alpha)} u_2^{\alpha-1} e^{-\beta u_2}, \quad u_2 > 0,$$

if we take $\alpha = \frac{k+1}{2}$ and $\beta = \frac{1}{2}\Big(\frac{u_1^2}{k}+1\Big)$. So

$$g_1(u_1) = \frac{1}{2^{(k+1)/2}\sqrt{\pi}\,\Gamma(k/2)\sqrt{k}} \frac{\Gamma((k+1)/2)}{\Big[\frac{1}{2}\big(\frac{u_1^2}{k}+1\big)\Big]^{(k+1)/2}} \int_0^\infty h(u_2) du_2.$$

But $\int_0^\infty h(u_2)du_2 = 1$ and $\sqrt{\pi} = \Gamma(1/2)$, so

$$\begin{aligned} g_1(u_1) &= \frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})\Gamma(\frac{1}{2})} \frac{\left(1+\frac{u_1^2}{k}\right)^{-\frac{k+1}{2}}}{\sqrt{k}} \\ &= \frac{1}{\sqrt{k}B\left(\frac{k}{2},\frac{1}{2}\right)}\left(1+\frac{u_1^2}{k}\right)^{-\frac{k+1}{2}}, \quad u_1 > 0, \end{aligned}$$

which is the pdf of a t_k distribution, so $U_1 \sim t_k$. □

The importance of the t distribution in practice comes from the following proposition, which we will use to construct confidence interval and hypothesis tests for the mean of normal random variables in Chapter 11:

Proposition 9.4. *Suppose that $Y_1, Y_2, \ldots, Y_n$, are independent and identically distributed, with each $Y_i \sim N(0,1)$. Let $\bar{Y}$ and S^2 be the usual sample mean and sample variance. Then*

$$\frac{\sqrt{n}(\bar{Y}-\mu)}{S} \sim t_{n-1}.$$

Proof. We know from Proposition 4.1 that $\bar{Y} \sim N(\mu, \sigma^2/n)$, so

$$A = \frac{\sqrt{n}(\bar{Y}-\mu)}{\sigma} \sim N(0,1).$$

We also know from Theorem 6.1 that

$$B = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

So, by Proposition 9.3,

$$\frac{A}{\sqrt{B/(n-1)}} \sim t_{n-1}.$$

Simplifying,

$$\begin{aligned} \frac{A}{\sqrt{B/(n-1)}} &= A\sqrt{n-1}\frac{1}{\sqrt{B}} \\ &= \frac{\sqrt{n}(\bar{Y}-\mu)}{\sigma}\sqrt{n-1}\frac{\sigma}{\sqrt{n-1}S} \\ &= \frac{\sqrt{n}(\bar{Y}-\mu)}{S}, \end{aligned}$$

so

$$\frac{\sqrt{n}(\bar{Y}-\mu)}{S} \sim t_{n-1}$$

as required. □

9.5 The F distribution

Definition 9.4. A random variable Y has F distribution with m and n degrees of freedom if it has pdf

$$f(y) = \frac{m^{\frac{n}{2}} n^{\frac{m}{2}}}{B\left(\frac{m}{2}, \frac{n}{2}\right)} \frac{y^{\frac{m}{2}-1}}{(n+my)^{\frac{m+n}{2}}}, \quad y > 0.$$

We write $Y \sim F_{m,n}$.

We may obtain the F distribution as the ratio of two independent chi-squared random variables, each divided by their degrees of freedom. As with the t distribution, this makes the F distribution important in statistics:

Proposition 9.5. *Suppose that $Y_1 \sim \chi^2_m$ and $Y_2 \sim \chi^2_n$, and that Y_1 and Y_2 are independent. Then*

$$U_1 = \frac{Y_1/m}{Y_2/n} \sim F_{m,n}.$$

Proof. Write $Y_1^* = \frac{Y_1}{m}$ and $Y_2^* = \frac{Y_2}{n}$, so $U_1 = Y_1^*/Y_2^*$. In order to use Theorem 9.1, we first need to define a second random variable U_2. We will choose $U_2 = Y_2^*$.

We have $Y_1 \sim \chi^2_m \equiv \text{gamma}(m/2, 1/2)$, so by Proposition 6.4 (with b $= 1/m$)

$$Y_1^* = \frac{Y_1}{m} \sim \text{gamma}(m/2, m/2).$$

Similarly, $Y_2^* \sim \text{gamma}(n/2, n/2)$. So the joint pdf of Y_1^* and Y_2^* is

$$f(y_1^*, y_2^*) = \frac{\left(\frac{m}{2}\right)^{m/2}}{\Gamma\left(\frac{m}{2}\right)} (y_1^*)^{\frac{m}{2}-1} e^{-\frac{m}{2}y_1^*} \cdot \frac{\left(\frac{n}{2}\right)^{n/2}}{\Gamma\left(\frac{n}{2}\right)} (y_2^*)^{\frac{n}{2}-1} e^{-\frac{n}{2}y_2^*}$$

for $y_1^* > 0$, $y_2^* > 0$.

The inverse transformation is $Y_1^* = U_1 U_2$, $Y_2^* = U_2$. The domain of U_1 and U_2 is $U_1 > 0$, $U_2 > 0$. The Jacobian is

$$\boldsymbol{J} = \begin{pmatrix} u_2 & * \\ 0 & 1 \end{pmatrix},$$

so $\det \boldsymbol{J} = u_2 > 0$. So the joint pdf of U_1 and U_2 is

$$\begin{aligned} g(u_1, u_2) &= \frac{\left(\frac{m}{2}\right)^{m/2}}{\Gamma\left(\frac{m}{2}\right)} (u_1 u_2)^{\frac{m}{2}-1} e^{-\frac{m}{2}u_1 u_2} \cdot \frac{\left(\frac{n}{2}\right)^{n/2}}{\Gamma\left(\frac{n}{2}\right)} u_2^{\frac{n}{2}-1} e^{-\frac{n}{2}u_2} \cdot u_2 \\ &= \frac{m^{\frac{m}{2}} n^{\frac{n}{2}}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right) 2^{(m+n)/2}} u_1^{\frac{m}{2}-1} u_2^{\frac{m+n}{2}-1} e^{-\frac{1}{2}(n+mu_1)u_2}, \end{aligned}$$

for $u_1 > 0$, $u_2 > 0$.

The marginal pdf of U_1 is

$$g_1(u_1) = \frac{m^{\frac{m}{2}} n^{\frac{n}{2}}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right) 2^{(m+n)/2}} u_1^{\frac{m}{2}-1} \int_0^\infty u_2^{\frac{m+n}{2}-1} e^{-\frac{1}{2}(n+mu_1)u_2} du_2.$$

The integrand is proportional to a gamma(α, β) pdf, with $\alpha = \frac{m+n}{2}$ and $\beta = \frac{1}{2}(n + mu_1)$:

$$h(u_2) = \frac{\left[\frac{1}{2}(n+mu_1)\right]^{\frac{m+n}{2}}}{\Gamma\left(\frac{m+n}{2}\right)} u_2^{\frac{m+n}{2}-1} e^{-\frac{1}{2}(n+mu_1)u_2}$$

so we have

$$\begin{aligned} g_1(u_1) &= \frac{m^{\frac{m}{2}} n^{\frac{n}{2}}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right) 2^{(m+n)/2}} u_1^{\frac{m}{2}-1} \frac{\Gamma\left(\frac{m+n}{2}\right)}{\left[\frac{1}{2}(n+mu_1)\right]^{\frac{m+n}{2}}} \int_0^\infty h(u_2) du_2 \\ &= \frac{m^{\frac{m}{2}} n^{\frac{n}{2}} \Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \frac{u_1^{\frac{m}{2}-1}}{(n+mu_1)^{\frac{m+n}{2}}} \\ &= \frac{m^{\frac{m}{2}} n^{\frac{n}{2}}}{B\left(\frac{m}{2}, \frac{n}{2}\right)} \frac{u_1^{\frac{m}{2}-1}}{(n+mu_1)^{\frac{m+n}{2}}}, \quad u_1 > 0, \end{aligned}$$

which is the pdf of a $F_{m,n}$ distribution, so $U_1 \sim F_{m,n}$. □

The importance of the F distribution in practice comes from the following proposition, which we will use in Chapter 11 to test whether two sets of normal random variables have the same variance:

Proposition 9.6. *Suppose $Y_1^{(1)}, \ldots, Y_m^{(1)}$ and $Y_1^{(2)}, \ldots, Y_n^{(2)}$ are two sets of independent random variables, with each $Y_i^{(1)} \sim N(\mu_1, \sigma^2)$ and each $Y_i^{(2)} \sim N(\mu_2, \sigma^2)$. Let S_1^2 be the sample variance of $Y_1^{(1)}, \ldots, Y_m^{(1)}$ and S_2^2 be the sample variance of $Y_1^{(2)}, \ldots, Y_n^{(2)}$. Then*

$$\frac{S_1^2}{S_2^2} \sim F_{m-1,n-1}.$$

Proof. We know from Theorem 6.1 that that $(m-1)S_1^2/\sigma^2 \sim \chi^2_{m-1}$ and $(n-1)S_2^2/\sigma^2 \sim \chi^2_{n-1}$, and S_1^2 and S_2^2 are independent. So by Proposition 9.5

$$\frac{S_1^2/\sigma^2}{S_2^2/\sigma^2} = \frac{S_1^2}{S_2^2} \sim F_{m-1,n-1}$$

as required. □

Part II

Statistical inference

Chapter 10

Parameter estimation

10.1 Estimators and estimates

Suppose that $Y_1, \ldots, Y_n$ are independent identically distributed random variables, each with a distribution depending on unknown parameter(s) θ. In the continuous case, we know the density function $f(y_i; \theta)$ up to the unknown θ, and in the discrete case we have probability function $p(y_i; \theta)$. We might be interested in estimating θ based on $Y_1, \ldots, Y_n$.

An *estimator* $T = T(Y_1, ..., Y_n)$ is just some suitable function (a "statistic") of $Y_1, Y_2, \ldots, Y_n$, which is used to estimate a parameter. It must not contain any unknown parameters or any unobservable quantities. For example, the sample mean $\bar{Y}$ is an obvious estimator for the population mean.

Note that an estimator is a function of random variables and hence can itself be treated as a random variable. Once a set of data has been collected (i.e. the observed values $y_1, y_2, \ldots, y_n$ of $Y_1, Y_2, \ldots, Y_n$ are available), then $T(y_1, ..., y_n)$ is the corresponding *estimate*. When deriving properties of T, one should always work with the estimator.

You have already seen several desirable properties of estimators in MATH1024, which we now review.

10.2 Bias

Definition 10.1. The *bias* of an estimator $T = T(Y_1, \ldots, Y_n)$ of a parameter θ is:

$$B(T; \theta) = E(T) - \theta.$$

T is said to be an *unbiased* estimator of θ if

$$E(T) - \theta = 0$$

for all possible values of θ.

Example 10.1 (Bias of estimator of Bernoulli success probability)**.** Let $Y_1, Y_2, \ldots, Y_n$ be independent and identically distributed, where each $Y_i \sim \text{Bernoulli}(\theta)$.

We know that $E(Y_i) = \theta$, so a natural estimator of θ is $T = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, the sample mean.

Here T is an unbiased estimator of θ, as

$$E(T) = E(\bar{Y}) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n} n\theta = \theta.$$

Example 10.2 (Bias of estimator of exponential rate parameter)**.** Let $Y_1, Y_2, \ldots, Y_n$ be independent and identically distributed, where each $Y_i \sim \text{exponential}(\theta)$.

We know that $E(Y_i) = 1/\theta$, so one possible estimator for θ is $T = 1/\bar{Y}$.

Since $Y_i \sim \text{gamma}(1, \theta)$, by Propositions 6.4 and 6.5,

$$\bar{Y} \sim \text{gamma}(n, n\theta).$$

Recall from Proposition 6.3 that $E(X) = \alpha/\beta$ if $X \sim \text{gamma}(\alpha, \beta)$. So $\bar{Y}$ is an unbiased estimator of $1/\theta$, as

$$E(\bar{Y}) = \frac{n}{n\theta} = \frac{1}{\theta}.$$

Given this, an obvious choice for an estimator of θ is $T = 1/\bar{Y}$. Since $\bar{Y} \sim \text{gamma}(n, n\theta)$, we have

$$\begin{aligned}
E(1/\bar{Y}) &= \int_0^\infty \frac{1}{y} \frac{(n\theta)^n}{\Gamma(n)} y^{n-1} e^{-n\theta y} dy \\
&= \frac{(n\theta)^n}{\Gamma(n)} \int_0^\infty y^{n-2} e^{-n\theta y} dy \\
&= \frac{(n\theta)^n}{\Gamma(n)} \frac{\Gamma(n-1)}{(n\theta)^{n-1}} \int_0^\infty \frac{(n\theta)^{n-1}}{\Gamma(n-1)} y^{n-2} e^{-n\theta y} dy \\
&\quad \text{(integrating the gamma}(n-1, n\theta) \text{ pdf)} \\
&= \frac{\Gamma(n-1)}{\Gamma(n)} n\theta \cdot 1 \\
&= \frac{1}{n-1} n\theta,
\end{aligned}$$

since $\Gamma(n) = (n-1)\Gamma(n-1)$, by Proposition 6.1. So the bias of $1/\bar{Y}$ as an estimator of θ is

$$B(1/\bar{Y};\theta) = \frac{n}{n-1}\theta - \theta = \frac{1}{n-1}\theta \neq 0.$$

$1/\bar{Y}$ is not an unbiased estimator of θ, but the bias shrinks with n.

10.3 Consistency

Definition 10.2. An estimator $T = T(Y_1, \ldots, Y_n)$ of a parameter θ is said to be a *consistent* estimator of θ if:

1. it is *asymptotically unbiased*: $B(T;\theta) \to 0$ as $n \to \infty$; and
2. $\mathrm{Var}(T) \to 0$ as $n \to \infty$.

for all possible values of θ.

Note that a consistent estimator can be biased, and an unbiased estimator can be inconsistent (i.e. not consistent). So consistency and unbiasedness are different types of property.

Example 10.3 (Consistency of estimator of Bernoulli success probability)**.** Continue example 10.1, with $Y_i \sim$ Bernoulli(θ) and $T = \bar{Y}$. T is a consistent estimator of θ, as:

1. $E(T) = E(\bar{Y}) = \theta$, so $B(T;\theta) = 0$. Therefore $B(T;\theta) \to 0$ as $n \to \infty$ (any unbiased estimator is asymptotically unbiased).
2. The variance is

$$\begin{aligned}
\mathrm{Var}(T) = \mathrm{Var}(\bar{Y}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(Y_i) \quad \text{by independence} \\
&= \frac{1}{n^2} n\theta(1-\theta) \\
&\to 0 \quad \text{as } n \to \infty, \text{ for all } \theta \in [0,1].
\end{aligned}$$

Example 10.4 (Consistency of estimator of exponential rate parameter)**.** Continue example 10.2, with $Y_i \sim$ exponential(θ) and $T = \frac{1}{\bar{Y}}$. T is a consistent estimator of θ, as:

1. The bias of T is
$$B(T;\theta) = \frac{1}{n-1}\theta \to 0$$
as $n \to \infty$.

2. The variance of T is $\text{Var}(T) = E(T^2) - [E(T)]^2$. Recall $\bar{Y} \sim \text{gamma}(n, n\theta)$, so

$$\begin{aligned} E(T^2) = E(\bar{Y}^{-2}) &= \int_0^\infty \frac{1}{y^2} \frac{(n\theta)^n}{\Gamma(n)} y^{n-1} e^{-n\theta y} dy \\ &= \frac{(n\theta)^n}{\Gamma(n)} \int_0^\infty y^{n-3} e^{-n\theta y} dy \\ &= \frac{(n\theta)^n}{\Gamma(n)} \frac{\Gamma(n-2)}{(n\theta)^{n-2}} \int_0^\infty \frac{(n\theta)^{n-2}}{\Gamma(n-2)} y^{n-3} e^{-n\theta y} dy \\ &\quad \text{(integrating the gamma}(n-2, n\theta) \text{ pdf)} \\ &= \frac{\Gamma(n-2)}{\Gamma(n)} (n\theta)^2 \cdot 1 \\ &= \frac{1}{(n-1)(n-2)} (n\theta)^2, \end{aligned}$$

since $\Gamma(n) = (n-1)\Gamma(n-1) = (n-1)(n-2)\Gamma(n-2)$, by Proposition 6.1. So

$$\begin{aligned} \text{Var}(T) &= E(T^2) - [E(T)]^2 \\ &= \frac{1}{(n-1)(n-2)} (n\theta)^2 - \frac{1}{(n-1)^2} (n\theta)^2 \\ &= \frac{(n\theta)^2}{(n-1)^2} \left[\frac{n-1}{n-2} - 1 \right] \\ &\to 0 \quad \text{as } n \to \infty, \text{ for all } \theta > 0. \end{aligned}$$

10.4 Mean squared error

If we have two unbiased estimators, T_1 and T_2, of θ then typically we prefer the one with the smaller variance. However, having a small variance on its own is not sufficient if the estimator is biased.

Definition 10.3. The *mean squared error* (or *MSE*) of an estimator T of θ is

$$\text{MSE}(T; \theta) = E\{(T - \theta)^2\},$$

the mean of the squared distance of the T from its target value θ.

We can use the MSE to choose between competing estimators whether they are unbiased or not.

Proposition 10.1. *For any estimator T of θ,*

$$MSE(T; \theta) = Var(T) + [B(T; \theta)]^2 .$$

Proof. We have

$$\begin{aligned}
\mathrm{MSE}(T;\theta) &= E\{(T-\theta)^2\} \\
&= E\{(T-E(T)+E(T)-\theta)^2\} \\
&= E\left\{(T-E(T))^2 + 2(T-E(T))(E(T)-\theta) + (E(T)-\theta)^2\right\} \\
&= E\left\{(T-E(T))^2\right\} + 2(E(T)-\theta)E\left\{T-E(T)\right\} + (E(T)-\theta)^2 \\
&= \mathrm{Var}(T) + 2(E(T)-\theta)\cdot 0 + \left[B(T;\theta)\right]^2 \\
&= \mathrm{Var}(T) + \left[B(T;\theta)\right]^2,
\end{aligned}$$

as required. □

An immediate consequence of Proposition 10.1 is that if T is an unbiased estimator of θ, then $\mathrm{MSE}(T;\theta) = \mathrm{Var}(T)$.

Note that it is not possible to find a *uniformly* minimum MSE estimator, that is, an estimator which will have the lower MSE than any other estimator for all value of θ. For instance, consider the estimator $T^* = 1$, which takes the simple (but usually bad!) strategy of estimating θ as 1, irrespective of the observed data. This has **zero** mean squared error at $\theta = 1$, so will beat any other estimator at $\theta = 1$. However, it will be a very bad estimator for other values of θ.

It is sometimes possible to find a uniformly minimum MSE estimator within a class of "sensible" estimators.

Example 10.5 (Comparing estimators of the normal variance parameter)**.** Suppose that $Y_1, Y_2, \ldots, Y_n$ are independent, with each $Y_i \sim N(\mu, \sigma^2)$, and that we wish to estimate σ^2. Two possible choices are: $S^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y})^2$ and $\frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})^2$. You can think of these as special cases of estimators of the form

$$T_c = c\sum_{i=1}^n (Y_i - \bar{Y})^2,$$

where c is some positive constant.

Can we find a value of c such that $\mathrm{MSE}(T_c;\sigma^2)$ is minimised for all $\sigma^2 > 0$? From Theorem 6.1, $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$, so

$$\frac{T_c}{c\sigma^2} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{\sigma^2} \sim \chi^2_{n-1}.$$

We may use this to find $\mathrm{MSE}(T_c;\sigma^2)$ for any c, and then minimise it with respect to c.

By Proposition 10.1,

$$\mathrm{MSE}(T_c;\sigma^2) = \mathrm{Var}(T_c) + \left[B(T_c;\sigma^2)\right]^2.$$

Let $X = \frac{T_c}{c\sigma^2} \sim \chi^2_{n-1}$. Recall from Section 6.3 that $E(X) = n-1$ and $\text{Var}(X) = 2(n-1)$. We have

$$E(T_c) = E(c\sigma^2 X) = c\sigma^2 E(X) = c\sigma^2(n-1),$$

so

$$B(T_c; \sigma^2) = \sigma^2[c(n-1) - 1].$$

The variance of T_c is

$$\text{Var}(T_c) = \text{Var}(c\sigma^2 X) = c^2\sigma^4 \text{Var}(X) = c^2\sigma^4 2(n-1),$$

so the mean squared error is

$$\begin{aligned} \text{MSE}(T_c; \sigma^2) &= c^2\sigma^4 2(n-1) + \sigma^4[c(n-1) - 1]^2 \\ &= \sigma^4 \left[2c^2(n-1) + c^2(n-1)^2 - 2c(n-1) + 1\right]. \end{aligned}$$

To minimise this, for any σ^2, we need to find c to minimise

$$g(c) = 2c^2(n-1) + c^2(n-1)^2 - 2c(n-1) + 1.$$

Differentiating, we have

$$g'(c) = 4c(n-1) + 2c(n-1)^2 - 2(n-1),$$

so we want to find c^* such that $g'(c^*) = 0$, or

$$4c^*(n-1) + 2c^*(n-1)^2 - 2(n-1) = 0.$$

Dividing by $2(n-1)$ and rearranging gives $c^*(2 + n - 1) = 1$, so $c^* = \frac{1}{n+1}$. So

$$T_{c^*} = \frac{1}{n+1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

has minimum mean squared error of all estimators in this class, for any value of σ^2.

Once we have come up with an estimator, we can check whether it has good properties, such as consistency or low mean squared error. However, it is not yet clear yet how we should go about finding an estimator for any given model, and it would be useful to have some general methods that will produce estimators. We will consider two general recipes for finding estimators, the method of moments, and maximum likelihood estimation.

10.5 Method of moments estimation

This approach essentially formalises an argument we have met before – "if I have data from a population with mean μ, it is natural to estimate μ by the corresponding sample mean $\bar{Y}$".

Suppose we have a sample of independent and identically distributed random variables $Y_1, Y_2, \ldots, Y_n$, whose probability function or probability density function depends on k unknown parameters $(\theta_1, \theta_2, \ldots, \theta_k)$. Then we may write the rth population moment about the origin as a function of the parameters:

$$\mu'_r = \mu'_r(\theta_1, \theta_2, \ldots, \theta_k), \quad \text{for } r = 1, 2, 3, \ldots.$$

We write

$$m'_r = \frac{1}{n} \sum_{i=1}^{n} Y_i^r$$

for the rth sample moment, for $r = 1, 2, 3, \ldots$.

Then (assuming each of the first k moments involves at least one parameter) we find values $\tilde{\theta}_1, \ldots, \tilde{\theta}_k$ such that the first k sample moments match the first k population moments

$$m'_r = \mu'_r(\tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_k) \text{ for } r = 1, 2, \ldots, \text{k.}$$

We call $\tilde{\theta}_1, \ldots, \tilde{\theta}_k$ the *method of moments estimators* of of $\theta_1, \ldots, \theta_k$.

Sometimes one of the first k expressions does not involve any parameters, in which case we need to add an extra simultaneous equation

$$m'_{k+1} = \mu'_{k+1}(\tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_k)$$

in order to be able to solve the system of simultaneous equations for the k unknowns $\tilde{\theta}_1, \ldots, \tilde{\theta}_k$ We always need at least k simultaneous equations to solve for the k unknowns: if any of those simultaneous equations does not provide any new information, then we need to generate more equations by matching higher-order population and sample moments, until it is possible to find a solution to the system of simultaneous equations.

This will become clearer with some examples. We first begin with some simple one-parameter examples to illustrate the method.

Example 10.6 (Bernoulli)**.** Suppose $Y_i \sim$ Bernoulli(θ). We have $\mu'_1 = \mu'_1(\theta) = \theta$. and $m_1 = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}$. So the method of moments estimator is $\tilde{\theta} = \bar{Y}$.

Example 10.7 (Normal mean)**.** Suppose $Y_i \sim N(\theta, 1)$. We have $\mu'_1 = \mu'_1(\theta) = \theta$, and $m_1 = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}$. So the method of moments estimator is $\tilde{\theta} = \bar{Y}$.

Example 10.8 (Exponential)**.** Suppose $Y_i \sim$ exponential(θ). We have $\mu'_1 = \mu'_1(\theta) = 1/\theta$, and $m_1 = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}$. So the method of moments estimator is $\tilde{\theta} = 1/\bar{Y}$.

Example 10.9 (Normal variance)**.** Suppose $Y_i \sim N(0, \theta)$. We know that $\mu_1' = \mu_1'(\theta) = 0$, which does not involve the parameter of interest. But $\mu_2' = \mu_2'(\theta) = \theta$, and $m_2' = \frac{1}{n}\sum_{i=1}^n Y_i^2$, so $\tilde{\theta} = \frac{1}{n}\sum_{i=1}^n Y_i^2$.

We can also use method of moment estimation for models with more than one unknown parameter.

Example 10.10 (Normal mean and variance)**.** Suppose $Y_i \sim N(\theta_1, \theta_2)$ (i.e. $\theta_1 = \mu$, $\theta_2 = \sigma^2$ in the usual notation). The first two population moments are $\mu_1'(\theta_1, \theta_2) = \theta_1$ and

$$\mu_2'(\theta_1, \theta_2) = \mathrm{Var}(Y) + [E(Y)]^2 = \theta_2 + \theta_1^2.$$

The first two sample moments are $m_1' = \bar{Y}$ and $m_2' = \frac{1}{n}\sum_{i=1}^n Y_i^2$. So we choose $\tilde{\theta}_1$ and $\tilde{\theta}_2$ to solve $\mu_1'(\tilde{\theta}_1, \tilde{\theta}_2) = m_1'$ and $\mu_2'(\tilde{\theta}_1, \tilde{\theta}_2) = m_2'$. So

$$\tilde{\theta}_1 = \bar{Y} \quad \text{and} \quad \tilde{\theta}_2 + \tilde{\theta}_1^2 = \frac{1}{n}\sum_{i=1}^n Y_i^2,$$

so

$$\tilde{\theta}_2 = \frac{1}{n}\sum_{i=1}^n Y_i^2 - \bar{Y}^2.$$

Example 10.11. Suppose $Y_i \sim \text{gamma}(\theta_1, \theta_2)$ In this notation the shape parameter is $\alpha = \theta_1$ and the rate parameter is $\beta = \theta_2$.

By Proposition 6.3, we have

$$\mu_1'(\theta_1, \theta_2) = \frac{\theta_1}{\theta_2}$$

and

$$\begin{aligned}\mu_2'(\theta_1, \theta_2) &= \mathrm{Var}(Y) + [E(Y)]^2 \\ &= \frac{\theta_1}{\theta_2^2} + \frac{\theta_1^2}{\theta_2^2} \\ &= \frac{\theta_1(1+\theta_1)}{\theta_2^2}.\end{aligned}$$

The first two sample moments are $m_1' = \bar{Y}$ and $m_2' = \frac{1}{n}\sum_{i=1}^n Y_i^2$.

So we choose $\tilde{\theta}_1$ and $\tilde{\theta}_2$ to solve $\mu_1'(\tilde{\theta}_1, \tilde{\theta}_2) = m_1'$ and $\mu_2'(\tilde{\theta}_1, \tilde{\theta}_2) = m_2'$. So

$$\frac{\tilde{\theta}_1}{\tilde{\theta}_2} = \bar{Y} \quad \text{and} \quad \frac{\tilde{\theta}_1(1+\tilde{\theta}_1)}{\tilde{\theta}_2^2} = \frac{1}{n}\sum_{i=1}^n Y_i^2.$$

So $\tilde{\theta}_1 = \bar{Y}\tilde{\theta}_2$, and

$$\frac{\bar{Y}\tilde{\theta}_2}{(1+\bar{Y}\tilde{\theta}_2)}\tilde{\theta}_2^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2,$$

or

$$\bar{Y}\tilde{\theta}_2^{-1} + \bar{Y}^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2.$$

Rearranging, we get

$$\tilde{\theta}_2 = \frac{\bar{Y}}{\frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \bar{Y}^2},$$

so

$$\tilde{\theta}_1 = \bar{Y}\tilde{\theta}_2 = \frac{\bar{Y}^2}{\frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \bar{Y}^2}.$$

10.6 Maximum likelihood estimation

Maximum likelihood estimation is a versatile method – it is the standard method in modern (frequentist) statistics – and it can be shown (see MATH3044) that maximum likelihood estimators (MLEs) have some nice optimality properties.

Maximum likelihood estimation can be applied in complex situations, but in this module we will stick to the situation where $Y_1, Y_2, \ldots, Y_n$ are independent and identically distributed random variables,

The likelihood function is a function of the unknown parameters θ, which gives the joint probability (or joint probability density) of seeing the observed data $y_1, \ldots y_n$, assuming the data were generating from the model with parameter value θ:

Definition 10.4. Suppose $Y_1, Y_2, \ldots, Y_n$ are independent and identically distributed random variables, with distribution depending on a vector of unknown parameters θ. The *likelihood function* is

$$L(\theta; y_1, \ldots, y_n) = f(y_1;\theta) \times f(y_2;\theta) \times \ldots \times f(y_n;\theta),$$

if each Y_i has continuous distribution, with probability density function $f(y;\theta)$, or

$$L(\theta; y_1, \ldots, y_n) = p(y_1;\theta) \times p(y_2;\theta) \times \ldots \times p(y_n;\theta).$$

if each Y_i has discrete distribution, with probability function $p(y;\theta)$.

The values of the observations $y_1, \ldots, y_n$ have been observed, so these are known, and $L(\cdot)$ may be regarded as a function of the unknown θ.

If $L(\theta_1; y_1, \ldots, y_n) > L(\theta_2; y_1, \ldots, y_n)$, we should prefer θ_1 to θ_2 because there is a greater likelihood that the observed data would occur with this value of

θ_1 than with θ_2. It follows that we should try to maximise the likelihood to find the value of θ which is most likely to have given rise to the data that were actually observed.

The maximum likelihood estimate $\hat{\theta} = \hat{\theta}(y_1, \ldots, y_n)$ of θ. is the value of θ which maximises the likelihood:

$$\hat{\theta} = \arg\max_{\theta} L(\theta; y_1, \ldots, y_n)$$

The corresponding maximum likelihood estimator is $\hat{\theta} = \hat{\theta}(Y_1, \ldots, Y_n)$.

Since we assume that the Y_i are independent and identically distributed, the likelihood is just a product of pdf or probability function terms. Maximising a sum is usually easier than maximising a product, so we often work with the log-likelihood

$$\log L(\theta; y_1, \ldots, y_n) = \sum_{i=1}^{n} \log f(y_i; \theta).$$

The value of θ which maximises $L(\cdot)$ is the same as that which maximises $\log L(\cdot)$, as $L(\theta_1) > L(\theta_2)$ if and only if $\log L(\theta_1) > \log L(\theta_2)$. So we may find the MLE as by maximising $\log L(\cdot)$:

$$\hat{\theta} = \arg\max_{\theta} \left\{ \log L(\theta; y_1, \ldots, y_n) \right\}.$$

Example 10.12 (Bernoulli)**.** Suppose $Y_i \sim \text{Bernoulli}(\theta)$. The likelihood is

$$L(\theta; y_1, \ldots, y_n) = \prod_{i=1}^{n} p(y_i; \theta) = \prod_{i=1}^{n} \theta^{y_i} (1-\theta)^{1-y_i}.$$

Hence the log-likelihood is

$$\begin{aligned}
\log L(\theta; y_1, \ldots, y_n) &= \log \left(\prod_{i=1}^{n} \theta^{y_i} (1-\theta)^{1-y_i} \right) \\
&= \sum_{i=1}^{n} \log \left(\theta^{y_i} (1-\theta)^{1-y_i} \right) \\
&= \sum_{i=1}^{n} \log \theta^{y_i} + \log(1-\theta)^{1-y_i} \\
&= \sum_{i=1}^{n} y_i \log \theta + (1-y_i) \log(1-\theta) \\
&= \left(\sum_{i=1}^{n} y_i \right) \log \theta + \left(n - \sum_{i=1}^{n} y_i \right) \log(1-\theta).
\end{aligned}$$

We now maximise $\log L(\cdot)$ as a function of θ. We attempt to find a stationary point of $\log L(\cdot)$ by differentiating and setting to zero. We have

$$\frac{d}{d\theta}\log L(\theta; y_1, \ldots, y_n) = \frac{\sum_{i=1}^n y_i}{\theta} - \frac{n - \sum_{i=1}^n y_i}{1-\theta},$$

and we want to find $\hat{\theta}$ such that

$$\frac{d}{d\theta}\log L(\theta; y_1, \ldots, y_n)|_{\theta=\hat{\theta}} = 0.$$

We have

$$\frac{\sum_{i=1}^n y_i}{\hat{\theta}} - \frac{n - \sum_{i=1}^n y_i}{1-\hat{\theta})} = 0,$$

and rearranging gives that

$$\hat{\theta} = \frac{\sum_{i=1}^n y_i}{n} = \bar{y}$$

is a stationary point of the log-likelihood, provided that $\sum_{i=1}^n y_i > 0$ and $\sum_{i=1}^n y_i < n$. If $\sum_{i=1}^n y_i = 0$, then $\log L(\theta) = n\log(1-\theta)$, which is decreasing with θ, so $\hat{\theta} = 0$ is the MLE. Similarly, if $\sum_{i=1}^n y_i = n$, then $\log L(\theta) = n\log\theta$, which is increasing with θ, so $\hat{\theta} = 1$ is the MLE.

If $0 < \sum_{i=1}^n y_i < n$, we have shown that $\hat{\theta} = \bar{y}$ is a stationary point of the the log-likelihood, but we still need to show that it is a maximum. To do this, we need to show that

$$\frac{d^2}{d\theta^2}\log L(\theta; y_1, \ldots, y_n)|_{\theta=\hat{\theta}} < 0.$$

We have

$$\frac{d^2}{d\theta^2} = -\frac{\sum_{i=1}^n y_i}{\theta^2} - \frac{n - \sum_{i=1}^n y_i}{(1-\theta)^2},$$

so

$$\frac{d^2}{d\theta^2}\log L(\theta; y_1, \ldots, y_n)|_{\theta=\hat{\theta}} = -\frac{\sum_{i=1}^n y_i}{\hat{\theta}^2} - \frac{n - \sum_{i=1}^n y_i}{(1-\hat{\theta})^2}.$$

We know $\hat{\theta} = \bar{y}$, so $0 < \hat{\theta} < 1$, as we are considering the case $0 < \sum_{i=1}^n y_i < n$. So $\hat{\theta} > 0$ and $1 - \hat{\theta} > 0$, so

$$\frac{d^2}{d\theta^2}\log L(\theta; y_1, \ldots, y_n)|_{\theta=\hat{\theta}} < 0,$$

and $\hat{\theta} = \bar{Y}$ is the maximum likelihood estimate.

In this case, the MLE $\hat{\theta}$ is the same as the method of moments estimate $\tilde{\theta}$ we found in Example 10.6. We have already studied the properties of this estimator, so know that $\hat{\theta}$ is an unbiased (see Example 10.1) and consistent (see Example 10.3) estimator of θ.

Example 10.13 (Exponential)**.** Suppose the $Y_i \sim$ exponential(θ), so that each Y_i has pdf $f(y;\theta) = \theta e^{-\theta y}$ for $y > 0$.

The likelihood is

$$L(\theta; y_1, \ldots, y_n) = \prod_{i=1}^{n} f(y_i;\theta) = \prod_{i=1}^{n} \theta e^{-\theta y_i},$$

since we know $y_i > 0$ for all i.

The log-likelihood is

$$\begin{aligned}
\log L(\theta; y_1, \ldots, y_n) &= \sum_{i=1}^{n} \log\left(\theta e^{-\theta y_i}\right) \\
&= \sum_{i=1}^{n} \left\{\log\theta + \log\left(e^{-\theta y_i}\right)\right\} \\
&= n\log\theta - \theta\sum_{i=1}^{n} y_i.
\end{aligned}$$

Differentiating, we have

$$\frac{d}{d\theta}\log L(\theta; y_1, \ldots, y_n) = \frac{n}{\theta} - \sum_{i=1}^{n} y_i,$$

so a stationary point of the log-likelihood $\hat{\theta}$ satisfies

$$\frac{n}{\hat{\theta}} - \sum_{i=1}^{n} y_i = 0.$$

Rearranging this, we get

$$\hat{\theta} = \frac{n}{\sum_{i=1}^{n} y_i} = \frac{1}{\bar{y}}.$$

Differentiating again, we have

$$\frac{d^2}{d\theta^2}\log L(\theta; y_1, \ldots, y_n) = -\frac{n}{\hat{\theta}} < 0 \quad \text{for all } \theta,$$

so $\hat{\theta}$ is a maximum of $\log L(\cdot)$, and hence $\hat{\theta}$ is the MLE.

In this case, the MLE $\hat{\theta}$ is the same as the method of moments estimate $\tilde{\theta}$ we found in Example 10.8.

Example 10.14 (Gamma)**.** Suppose $Y_i \sim \text{gamma}(\theta, 1)$.

To find the method of moments estimate of θ, we have $\mu_1' = E(Y_i) = \theta$ and $m_1' = \bar{Y}$, so $\tilde{\theta} = \bar{Y}$.

To find the maximum likelihood estimate, we have

$$f(y;\theta) = \frac{\Gamma(\theta)}{y}^{\theta-1} e^{-y}, \quad y > 0,$$

so the likelihood is

$$L(\theta; y_1, \ldots, y_n) = \prod_{i=1}^{n} \frac{1}{\Gamma(\theta)} \frac{\Gamma(\theta)}{y}_i^{\theta-1} e^{-y_i},$$

since all $y_i > 0$.

The log-likelihood is

$$\begin{aligned}\log L(\theta; y_1, \ldots, y_n) &= \sum_{i=1}^{n} \log\left(\frac{1}{\Gamma(\theta)} y_i^{\theta-1} e^{-y_i}\right) \\ &= \sum_{i=1}^{n} \log\left(\frac{1}{\Gamma(\theta)}\right) + (\theta - 1)\sum_{i=1}^{n} \log y_i - \sum_{i=1}^{n} y_i.\end{aligned}$$

Differentiating, we have

$$\frac{d}{d\theta} \log L(\theta; y_1, \ldots, y_n) = -n\frac{d}{d\theta} \log \Gamma(\theta) + \sum_{i=1}^{n} \log y_i,$$

where $\psi(\theta) = \frac{d}{d\theta} \log \Gamma(\theta)$ is called the *digamma function*.

So $\hat{\theta}$ satisfies

$$-n\psi(\hat{\theta}) + \sum_{i=1}^{n} \log y_i = 0,$$

or

$$\psi(\hat{\theta}) = \frac{\sum_{i=1}^{n} \log y_i}{n} \tag{10.1}$$

which has no closed-form solution for $\hat{\theta}$. We could use numerical methods to find a root of (10.1), or to directly maximise the log-likelihood function: see MATH3044 (Statistical Inference) for details. It is very common in practice that it is not possible to write down a closed-form expression for the MLE, so we must use numerical methods.

Example 10.15 (Normal mean and variance)**.** Suppose $Y_i \sim N(\mu, \sigma^2)$, where μ and σ^2 are both unknown parameters. We have

$$f(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\},$$

so the likelihood is

$$L(\mu, \sigma^2; y_1, \ldots, y_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i-\mu)^2\right\}$$

and the log-likelihood is

$$\begin{aligned}\log L(\mu, \sigma^2; y_1, \ldots, y_n) &= \sum_{i=1}^{n} \log\left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i-\mu)^2\right\}\right] \\ &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2.\end{aligned}$$

Differentiating, we obtain

$$\frac{\partial}{\partial\mu}\log L(\mu, \sigma^2; y_1, \ldots, y_n) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n} -2(y_i-\mu) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\mu)$$

and

$$\frac{\partial}{\partial\sigma^2}\log L(\mu, \sigma^2; y_1, \ldots, y_n) = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2)}\sum_{i=1}^{n}(y_i-\mu)^2.$$

So stationary points $\hat{\mu}$ and $\hat{\sigma}^2$ solve

$$\frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}(y_i-\hat{\mu}) = 0$$

and

$$-\frac{n}{2\hat{\sigma}^2} + \frac{1}{2(\hat{\sigma}^2)^2)}\sum_{i=1}^{n}(y_i-\hat{\mu})^2 = 0.$$

From the first equation, we obtain $\hat{\mu} = \bar{y}$, so

$$-\frac{n}{2\hat{\sigma}^2} + \frac{1}{2(\hat{\sigma}^2)^2)}\sum_{i=1}^{n}(y_i-\bar{y})^2 = 0,$$

and multiplying through by $2\hat{\sigma}^2$ gives

$$-n\hat{\sigma}^2 + \sum_{i=1}^{n}(y_i-\bar{y})^2 = 0,$$

so $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2$. In the exercises, you can check that $(\hat{\mu}, \hat{\sigma}^2)$ is a maximum of the log-likelihood, so is the MLE.

Chapter 11

Confidence intervals and hypothesis testing

11.1 Expressing uncertainty in parameter estimates

Suppose that $Y_1, \ldots, Y_n$ are independent identically distributed random variables, each with a distribution depending on unknown parameter θ, and let $y_1, \ldots, y_n$ be corresponding observed values.

In Chapter 10, we have seen how to find an estimate of θ given $y_1, \ldots, y_n$. However, an important part of statistical inference is to express our level of uncertainty in this estimate. This could be done by writing down an interval containing a range of values of θ which could plausibly have generated the data. Alternatively, we might be interested in testing if some particular value of θ could plausibly have generated the observed data.

11.2 Confidence intervals

Write $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ and $\boldsymbol{y} = (y_1, \ldots, y_n)^T$.

Definition 11.1. A $100(1-\alpha)\%$ *confidence interval* for θ is an interval $[L(\boldsymbol{y}), U(\boldsymbol{y})]$ such that

$$P\left(L(\boldsymbol{Y}) \leq \theta \leq U(\boldsymbol{Y})\right) = 1 - \alpha.$$

Often, we take $\alpha = 0.05$, and obtain a 95% confidence interval. This means that if we were to generate a large number of datasets from the model with some fixed

value of θ, and find a 95% confidence interval for each dataset, approximately 95% of those intervals would contain the true value of θ.

Example 11.1 (Normal mean, known variance)**.** Suppose that $Y_1, \ldots, Y_n$ are independent and identically distributed, with each $Y_i \sim N(\mu, \sigma^2)$, where μ is an unknown parameter, but the value of σ^2 is known. Suppose that we require a 95% confidence interval for μ.

We know (by Proposition 4.1) that $\bar{Y} \sim N(\mu, \sigma^2/n)$, so $\frac{\sqrt{n}(\bar{Y}-\mu)}{\sigma} \sim N(0,1)$, so

$$P\left(z_{0.025} \leq \frac{\sqrt{n}(\bar{Y}-\mu)}{\sigma} \leq z_{0.975}\right) = 0.95.$$

where z_p is the p-quantile of the standard normal distribution, so that $P(X \leq z_p) = p$ if $X \sim N(0,1)$. We can find these quantiles in R:

```
qnorm(0.025)
```

```
## [1] -1.959964
```

```
qnorm(0.975)
```

```
## [1] 1.959964
```

Notice that $z_p = -z_{1-p}$, because the standard normal distribution is symmetric about zero.

By "making μ the subject of the inequality" we can rearrange this probability statement to give

$$P\left(\bar{Y} - \frac{1.96\sigma}{\sqrt{n}} \leq \mu \leq \bar{Y} + \frac{1.96\sigma}{\sqrt{n}}\right) = 0.95.$$

If we replace the end-points of the inequality with their sample equivalents, we get a 95% confidence interval

$$\left[\bar{y} - \frac{1.96\sigma}{\sqrt{n}}, \bar{y} + \frac{1.96\sigma}{\sqrt{n}}\right]$$

for μ.

Example 11.2 (Normal mean, unknown variance)**.** Suppose that $Y_1, \ldots, Y_n$ are independent and identically distributed, with each $Y_i \sim N(\mu, \sigma^2)$, where μ and σ^2 are both unknown parameters. Suppose that we require a 95% confidence interval for μ.

We estimate σ^2 by $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ and then proceed as in Example 11.1, taking care to replace the normal quantile with a corresponding quantile from the relevant t distribution, as we know (by Proposition 9.4) that

$$\frac{\sqrt{n}(\bar{Y}-\mu)}{S} \sim t_{n-1}.$$

We have

$$P(t_{n-1,0.025} \leq \frac{\sqrt{n}(\bar{Y}-\mu)}{\sigma} \leq t_{n-1,0.975}) = 0.95.$$

where $t_{n-1,p}$ is the p-quantile of the t_{n-1} distribution, so that $P(X \leq t_{n-1,p}) = p$ if $X \sim t_{n-1}$. We can find these quantiles in R, e.g. if $n = 10$:

```
qt(0.025, df = 9)
```

```
## [1] -2.262157
```

```
qt(0.975, df = 9)
```

```
## [1] 2.262157
```

Again, $t_{p,n-1} = -t_{1-p,n-1}$, because the t distribution is symmetric about zero. So we have

$$P\left(-t_{n-1,0.975} \leq \frac{\sqrt{n}(\bar{Y}-\mu)}{\sigma} \leq t_{n-1,0.975}\right) = 0.95.$$

and rearranging this to "make μ the subject" gives

$$P\left(\bar{Y} - \frac{t_{n-1,0.975}S}{\sqrt{n}} \leq \mu \leq \bar{Y} + \frac{t_{n-1,0.975}S}{\sqrt{n}}\right) = 0.95.$$

Replacing the end points with their sample versions

$$\left[\bar{y} - \frac{t_{n-1,0.975}s}{\sqrt{n}}, \bar{y} + \frac{t_{n-1,0.975}s}{\sqrt{n}}\right]$$

is a 95% confidence interval for μ.

We always have $t_{n-1,0.975} > z_{0.975} = 1.96$, so the confidence interval when σ^2 is unknown will be wider than it would be if σ^2 was known: this makes sense, as we have to account for some additional uncertainty. For large n, $t_{n-1,0.975} \approx 1.96$, as the t_{n-1} approaches a standard normal distribution as n increases.

Example 11.3 (Normal variance)**.** Suppose that $Y_1, \ldots, Y_n$ are independent and identically distributed, with each $Y_i \sim N(\mu, \sigma^2)$, where μ and σ^2 are both unknown parameters. Suppose that we require a 95% confidence interval for σ^2.

We may estimate σ^2 by S^2. By Theorem 6.1, we know

$$\frac{n-1}{\sigma^2}S^2 \sim \chi^2_{n-1},$$

so

$$P\left(c_{n-1,0.025} \leq \frac{n-1}{\sigma^2}S^2 \leq c_{n-1,0.975}\right) = 0.95,$$

where $c_{n-1,p}$ is the p-quantile of the χ^2_{n-1} distribution, so that $P(X \leq c_{n-1,p}) = p$ if $X \sim \chi^2_{n-1}$. We can find these quantiles in R, e.g. if $n = 10$:

```
qchisq(0.025, df = 9)
```

```
## [1] 2.700389
```

```
qchisq(0.975, df = 9)
```

```
## [1] 19.02277
```

Since chi-squared distribution has positive domain, all quantiles are positive, and $c_{p,n-1} \neq -c_{1-p,n-1}$. Rearranging to "make σ^2 the subject" gives

$$P\left(\frac{(n-1)S^2}{c_{n-1,0.975}} \leq \sigma^2 \leq \frac{(n-1)S^2}{c_{n-1,0.025}}\right) = 0.95$$

so

$$\left[\frac{(n-1)s^2}{c_{n-1,0.975}}, \frac{(n-1)s^2}{c_{n-1,0.025}}\right]$$

is a 95% confidence interval for σ^2.

From this we can obtain a corresponding confidence interval for σ if we prefer, as

$$P\left(\frac{\sqrt{n-1}S}{\sqrt{c_{n-1,0.975}}} \leq \sigma \leq \frac{\sqrt{n-1}S}{\sqrt{c_{n-1,0.025}}}\right) = 0.95$$

so

$$\left[\frac{\sqrt{n-1}s}{\sqrt{c_{n-1,0.975}}}, \frac{\sqrt{n-1}s}{\sqrt{c_{n-1,0.025}}}\right]$$

is a 95% confidence interval for σ.

All of our examples of confidence intervals have been for unknown parameters of a normal distribution. We have been able to find these confidence intervals because of the results we have proved earlier about the distributions of $\bar{Y}$ and S^2, which are natural estimators of μ and σ^2. For other distributions, the distribution of an estimator of θ (such as the maximum likelihood estimator $\hat{\theta}$) may be more complicated, which makes constructing confidence intervals more challenging. It turns out that for large n, the distribution of the maximum likelihood estimator is close to a normal distribution (see MATH3044), which is very useful in practice to construct confidence intervals for a wide range of models.

11.3 Hypothesis testing

In classical hypothesis testing we aim to choose between two competing hypotheses about a parameter of interest. The hypotheses are called the null hypothesis (H_0) and the alternative hypothesis (H_1). The null and alternative hypotheses are regarded somewhat differently – the null hypothesis will be rejected in favour of the alternative hypothesis only if there is strong evidence against it.

For now we will only consider a **simple** null hypothesis, which is one which specifies a single value for the parameter of interest.

We either reject H_0 in favour of H_1, or we do not reject H_0. There are two types of error we can make:

- We reject H_0 when H_0 is true (Type I error)
- We do not reject H_0 when H_1 is true (Type II error)

In classical hypothesis testing we choose the Type I error probability we work at (the **significance level**) and design our test so that the Type II error probability is suitably small (i.e. choose a large enough sample size n.)

In any given situation we need a **test statistic** – a quantity whose distribution is known when H_0 is true. We reject H_0 if the value of the test statistic is "extreme" relative to the distribution of the test statistic under H_0, otherwise we do not reject H_0. The threshold for what counts as "extreme" depends on the specified significance level of the test.

Example 11.4 (Normal mean, known variance)**.** As in Example 11.1, suppose that $Y_1, \ldots, Y_n$ are independent and identically distributed, with each $Y_i \sim N(\mu, \sigma^2)$, where μ is an unknown parameter, but the value of σ^2 is known.

Suppose that we wish to test that null hypothesis $H_0 : \mu = \mu_0$, where μ_0 is a fixed value chosen in advance of collecting the data. We take H_1 to be the complement of H_0, so $H_1 : \mu \neq \mu_0$.

We construct the test statistic

$$Z = \frac{\sqrt{n}(\bar{Y} - \mu_0)}{\sigma}.$$

If H_0 is true then $Z \sim N(0, 1)$. For our observed data the observed value of Z is

$$z = \frac{\sqrt{n}(\bar{y} - \mu_0)}{\sigma}.$$

To construct a hypothesis test at significance level $\alpha = 0.05$ we should only reject H_0 if $|z| > z_{0.975} = 1.96$. So if $z > 1.96$ or $z < -1.96$, we reject H_0. Otherwise, we do not reject H_0.

Example 11.5 (Normal mean, unknown variance)**.** As in Example 11.2, suppose that $Y_1, \ldots, Y_n$ are independent and identically distributed, with each $Y_i \sim N(\mu, \sigma^2)$, where μ and σ^2 are both unknown parameters. Suppose we are interested in testing the null $H_0 : \mu = \mu_0$ against the alternative $H_1 : \mu \neq \mu_0$.

We estimate σ^2 by S^2, and construct the test statistic

$$T = \frac{\sqrt{n}(\bar{Y} - \mu_0)}{S}.$$

If H_0 is true then $T \sim t_{n-1}$. For our observed data the observed value of T is

$$t = \frac{\sqrt{n}(\bar{y} - \mu_0)}{s}.$$

To construct a hypothesis test at significance level $\alpha = 0.05$ we should only reject H_0 if $|t| > t_{n-1,0.975} = 1.96$. So if $t > t_{n-1,0.975}$ or $z < -t_{n-1,0.975}$, we reject H_0. Otherwise, we do not reject H_0.

Example 11.6 (Normal variance)**.** As in Example 11.3, suppose that $Y_1, \ldots, Y_n$ are independent and identically distributed, with each $Y_i \sim N(\mu, \sigma^2)$, where μ and σ^2 are both unknown parameters. Suppose we are interested in testing the null $H_0 : \sigma^2 = \sigma_0^2$ against the alternative $H_1 : \sigma^2 \neq \sigma_0^2$.

We estimate σ^2 by S^2, and construct the test statistic

$$C = \frac{(n-1)S^2}{\sigma_0^2}.$$

If H_0 is true then $C \sim \chi^2_{n-1}$. For our observed data the observed value of C is

$$c = \frac{(n-1)s^2}{\sigma_0^2}.$$

To construct a hypothesis test at significance level $\alpha = 0.05$ we should only reject H_0 if $c < c_{n-1,0.025}$ or $c > c_{n-1,0.975}$. Otherwise, we do not reject H_0.

11.4 Two-sample hypothesis testing

In many practical situations, such as clinical trials, we have two independent groups of subjects under study and wish to understand the relative effects of two "treatments" on some response of interest. For instance, in a classical two-armed clinical trial, there are two treatments of interest, such as an active treatment and a placebo, or old and new treatments, and we wish to know whether there is a difference in responses between the two treatment groups.

Example 11.7 (Classical two-sample t-test)**.** Suppose we have two sets of samples

$$X_1, \ldots, X_m \sim N(\mu_1, \sigma^2) \quad \text{and} \quad Y_1, \ldots, Y_n \sim N(\mu_2, \sigma^2),$$

where all the random variables are independent.

Write $\delta = \mu_1 - \mu_2$. We would like to test the null $H_0 : \delta = \delta_0$, for some prespecified value of the difference between treatments, against the alternative $H_1 : \delta \neq \delta_0$). Often $\delta_0 = 0$, in which case we are testing if there is any difference in distribution of the response between the two treatment groups.

We know that $\bar{X} \sim N(\mu_1, \sigma^2/m)$ and $\bar{Y} \sim N(\mu_2, \sigma^2/n)$, and $\bar{X}$ and $\bar{Y}$ are independent, so

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma^2}{m} + \frac{\sigma^2}{n}\right).$$

Under H_0, $\bar{X} - \bar{Y} \sim N(\delta_0, \sigma^2/m + \sigma^2/n)$, or

$$\frac{\bar{X} - \bar{Y} - \delta_0}{\sigma^2/m + \sigma^2/n} \sim N(0,1), \tag{11.1}$$

but we cannot use this as a test statistic, because it depends on the unknown σ^2.

We estimate σ^2 based on all the samples combined, by

$$S_c^2 = \frac{\sum_{i=1}^m (X_i - \bar{X})^2 + \sum_{i=1}^n (Y_i - \bar{Y})^2}{m+n-2}.$$

We choose the denominator $m+n-2$ to make this an unbiased estimator of σ^2: since $\sum_{i=1}^m (X_i - \bar{X})^2 \sim \sigma^2 \chi^2_{m-1}$ and $\sum_{i=1}^n (Y_i - \bar{Y})^2 \sim \sigma^2 \chi^2_{n-1}$, and these two quantities are independent, their sum has $\sigma^2 \chi^2_{m+n-2}$ distribution. So

$$\frac{(m+n-2)S_c^2}{\sigma^2} \sim \chi^2_{m+n-2}.$$

Replacing σ^2 by S_c^2 in Equation (11.1), we get

$$T = \frac{\bar{X} - \bar{Y} - \delta_0}{S_c^2/m + S_c^2/n} \sim t_{m+n-2}$$

under H_0. For our observed data the observed value of T is

$$t = \frac{\bar{x} - \bar{y} - \delta_0}{s_c^2/m + s_c^2/n}.$$

To construct a hypothesis test at significance level $\alpha = 0.05$ we should only reject H_0 if $|t| > t_{n-1,0.975} = 1.96$. So if $t > t_{n-1,0.975}$ or $z < -t_{n-1,0.975}$, we reject H_0. Otherwise, we do not reject H_0.

Example 11.8 (F-test for equality of variances)**.** In the classical two-sample t-test (Example 11.7), an assumption is made that the (population) variances of the two groups are equal. We might want to test whether this assumption appears to be reasonable, given the data. To do this, we now assume that each $X_i \sim N(\mu_1, \sigma_1^2)$ and each $Y_i \sim N(\mu_2, \sigma_2^2)$, and test the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ against the alternative $H_1 : \sigma_1^2 \neq \sigma_2^2$.

We estimate σ_1^2 by

$$S_1^2 = \frac{1}{m-1} \sum_{i=1}^m (X_i - \bar{X})^2$$

and σ_2^2 by

$$S_2^2 = \frac{1}{m-1} \sum_{i=1}^m (Y_i - \bar{Y})^2.$$

Let

$$F = \frac{S_1^2}{S_2^2}$$

Under H_0, by Proposition 9.6, we know that $F \sim F_{m-1,n-1}$. For our observed data the observed value of F is

$$f = \frac{s_1^2}{s_2^2}.$$

To construct a hypothesis test at significance level $\alpha = 0.05$ we should only reject H_0 if $f < f_{m-1,n-1,0.025}$ or $f > f_{m-1,n-1,0.975}$, where $f_{m-1,n-1,p}$ is the p-quantile of an $F_{m-1,n-1}$ distribution. Otherwise, we do not reject H_0.

Chapter 12

Bayesian inference

12.1 Frequentist and Bayesian inference

In frequentist inference, uncertainty about a parameter value is usually expressed through a confidence interval for that parameter: an interval $[L(\boldsymbol{y}), U(\boldsymbol{y})]$ such that

$$P(L(\boldsymbol{Y}) \leq \theta \leq U(\boldsymbol{Y})) = 1 - \alpha.$$

We treat θ as fixed (but unknown), and the probabilities are in terms of the random variables $\boldsymbol{Y} = (Y_1, \ldots Y_n)^T$. For instance, if we find $(1.1, 2.3)$ is a 95% confidence interval for θ, this **does not** mean that $P(1.1 \leq \theta \leq 2.3) = 0.95$, as we treat θ as a fixed value.

By contrast, in Bayesian inference, we treat θ as a random variable, and construct a probability distribution which summarises our belief about the likely value of a parameter θ. Our belief about which values of θ are likely (the "posterior" distribution for θ) is influenced by two factors: how likely the observed data $y_1, \ldots, y_n$ were to be generated using that value of θ (the likelihood $L(\theta; y_1, \ldots, y_n)$); and how likely we thought each value θ was before conducting the experiment (the "prior" distribution for θ).

We will give a very brief overview of Bayesian inference: see MATH3044 for more details.

12.2 Prior and posterior distributions

The first step of Bayesian inference is to express our beliefs about θ before conducting the experiment. We specify these beliefs through a probability distribution, which is called the *prior distribution*. Typically θ is a continuous

random variable, so we specify the distribution through a density $\pi(\theta)$. This idea will become clearer later on, when we consider an example.

The *posterior distribution* is the probability distribution for a parameter θ, conditional on the event $\{Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n\}$.

To find this distribution, we will use Bayes' Theorem, which you have already seen in MATH1024:

Theorem 12.1 (Bayes' Theorem)**.** *For two events A and B, such that $P(B) > 0$,*

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

Proof. By definition,

$$P(B|A) = \frac{P(A \cap B)}{P(A)}, \quad \text{if } P(A) > 0,$$

so $P(A \cap B) = P(B|A)P(A)$, which holds even if $P(A) = 0$.

So

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)}, \quad \text{since } P(B) > 0 \\ &= \frac{P(B|A)P(A)}{P(B)} \end{aligned}$$

as required. □

We use a continuous version of Bayes' Theorem to construct the probability distribution for the parameters θ, given $Y_1 = y_1, \ldots, Y_n = y_n$.

If $Y_1, \ldots, Y_n$ have discrete distribution, the probability density for θ, given the event $B = \{Y_1 = y_1, \ldots, Y_n = y_n\}$ is

$$\begin{aligned} \pi(\theta|y_1, \ldots, y_n) &= \frac{P(Y_1 = y_1, \ldots, Y_n = y_n|\theta)\pi(\theta)}{P(Y_1 = y_1, \ldots, Y_n = y_n)} \\ &= \frac{L(\theta; y_1, \ldots, y_n)\pi(\theta)}{P(Y_1 = y_1, \ldots, Y_n = y_n)} \end{aligned}$$

where

$$P(Y_1 = y_1, \ldots, Y_n = y_1) = \int L(\theta; y_1, \ldots, y_n)\pi(\theta)d\theta,$$

because

$$\int \pi(\theta|y_1, \ldots, y_n)d\theta = \frac{\int L(\theta; y_1, \ldots, y_n)\pi(\theta)d\theta}{P(Y_1 = y_1, \ldots, Y_n = y_1)} = 1,$$

as $\pi(\theta|y_1, \ldots, y_n)$ is a probability density function.

The denominator $P(Y_1 = y_1, \ldots Y_n = y_n)$ does not depend on θ, so usually we write

$$\pi(\theta|y_1, \ldots, y_n) \propto L(\theta; y_1, \ldots, y_n)\pi(\theta),$$

and if necessary find the constant of proportionality to make sure that $\pi(\theta|y_1, \ldots y_n)$ integrates to 1. Sometimes we recognise the pdf of a known distribution, and do not need to compute the constant.

If $Y_1, \ldots Y_n$ have continuous distribution, with p.d.f. $f(y; \theta)$, the posterior density has the same form

$$\pi(\theta|y_1, \ldots, y_n) \propto L(\theta; y_1, \ldots, y_n)\pi(\theta).$$

Example 12.1 (Bernoulli)**.** Suppose that $Y_1, \ldots Y_n$ are independent and identically distributed, with each $Y_i \sim$ Bernoulli(θ) where θ is an unknown parameter.

To be able to conduct Bayesian inference, we need to write down a prior distribution for θ. In this case, it is convenient to choose a beta distribution $\theta \sim$ beta(m_0, n_0) for the prior, so that

$$\pi(\theta) \propto \theta^{m_0-1}(1-\theta)^{n_0-1}.$$

We will see that if we make this choice, then the posterior distribution will also be a beta distribution.

The likelihood function for θ is

$$L(\theta; y_1, \ldots, y_n) \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} = \theta^{\sum_{i=1}^n y_i}(1-\theta)^{n-\sum_{i=1}^n y_i}.$$

The posterior distribution is

$$\begin{aligned}
\pi(\theta|y_1, \ldots, y_n) &\propto L(\theta; y_1, \ldots, y_n)\pi(\theta) \\
&\propto \theta^{\sum_{i=1}^n y_i}(1-\theta)^{n-\sum_{i=1}^n y_i}\theta^{m_0-1}(1-\theta)^{n_0-1} \\
&= \theta^{\sum_{i=1}^n y_i+n_0-1}(1-\theta)^{n-\sum_{i=1}^n y_i+n_0-1},
\end{aligned}$$

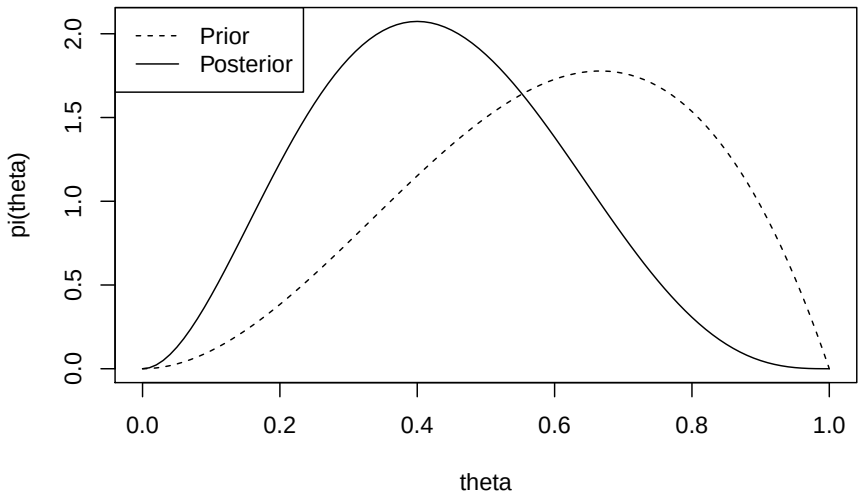
which is proportional to the pdf of a beta $(\sum_{i=1}^n y_i + n_0, n - \sum_{i=1}^n y_i + n_0)$ distribution, so this is the posterior distribution. Writing $s = \sum_{i=1}^n y_i$ for the number of observed "successes", and $f = n - \sum_{i=1}^n y_i$ for the number of observed "failures", we have

$$\theta|y_1, \ldots, y_n \sim \text{beta}(s + n_0, f + m_0).$$

This means that we may interpret the prior distribution as equivalent to the information we would gain by seeing m_0 successes and n_0 failures.

In reality, in order to choose a sensible prior we need to know more information about the type of process we are modelling. For instance, suppose that our data are the outcomes of n tennis matches between two friends, Alex and Bob, where $Y_i = 1$ denotes a victory for Alex, and $Y_i = 0$ a victory for Bob. Suppose that Alex is 25 and healthy, whereas Bob is 52 and slightly overweight. Before the matches are played, you have some prior belief about θ, the probability that Alex will win a game of tennis against Bob. The prior distribution reflects your personal beliefs: your prior may well look quite different to another person's prior. In this example, we might suppose $\theta \sim \text{beta}(3, 2)$. If we then observe two matches, both won by Bob ($s = 0$, $f = 2$), the posterior distribution will be $\theta|\boldsymbol{Y} \sim \text{beta}(3, 4)$. In this case, the smaller values of θ are given a higher probability in the posterior than in the prior, because of the what we have learnt from the data:

```
curve(dbeta(x, 3, 4), 0, 1, ylab = "pi(theta)", xlab = "theta")
curve(dbeta(x, 3, 2), 0, 1, lty = 2, add = TRUE)
legend("topleft", lty = c(2, 1), c("Prior", "Posterior"))
```



12.3 The posterior predictive distribution

Suppose we wanted to predict the outcome of a new random variables Y_{n+1}, assumed to have the same distribution as $Y_1, \ldots, Y_n$.

If the Y_i are discrete random variables, the posterior predictive distribution is

$$P(Y_{n+1} = y|y_1, \ldots, y_n) = \int_\theta p(y;\theta)\pi(\theta|y_1, \ldots, y_n)d\theta.$$

Example 12.2 (Bernoulli)**.** Continuing Example 12.1, with $Y_i \sim$ Bernoulli(θ) and prior $\theta \sim$ beta(m_0, n_0), recall that $\theta|y_1, \ldots, y_n \sim$ beta($s + m_0, f + n_0$), where $s = \sum_{i=1}^n y_i$ and $f = n - \sum_{i=1}^n y_i$. We have

$$\pi(\theta|y_1, \ldots, y_n) = \frac{1}{B(s + m_0, f + n_0)}\theta^{s+m_0-1}(1-\theta)^{f+n_0-1}, \quad 0 < \theta < 1,$$

and

$$p(y;\theta) = \theta^y(1-\theta)^{1-y}, \quad y \in \{0, 1\}.$$

We have

$$\begin{aligned}
P(Y_{n+1} = 1|y_1 \ldots y_n) &= \int_0^1 \theta \frac{1}{B(s + m_0, f + n_0)}\theta^{s+m_0-1}(1-\theta)^{f+n_0-1}d\theta \\
&= \frac{1}{B(s + m_0, f + n_0)}\int_0^1 \theta^{s+m_0}(1-\theta)^{f+n_0-1}d\theta \\
&= \frac{B(s + m_0 + 1, f + n_0)}{B(s + m_0, f + n_0)}\int_0^1 h(\theta)d\theta \\
&\qquad \text{where } h(\theta) \text{ the beta}(s + m_0 + 1, f + n_0) \text{ pdf} \\
&= \frac{\Gamma(s + m_0 + 1)\Gamma(f + n_0)}{\Gamma(s + m_0 + 1 + f + n_0)} \cdot \frac{\Gamma(s + m_0 + f + n_0)}{\Gamma(s + m_0)\Gamma(f + n_0)} \cdot 1 \\
&= \frac{\Gamma(n + m_0 + n_0)}{\Gamma(n + m_0 + n_0 + 1)}\frac{\Gamma(s + m_0 + 1)}{\Gamma(s + m_0)} \\
&= \frac{s + m_0}{n + m_0 + n_0}.
\end{aligned}$$

In our tennis match example, with $m_0 = 3$, $n_0 = 2$, after observing two matches both won by Bob ($s = 0$, $f = 2$), our predicted probability that Alex will win the next match is

$$P(Y_3 = 1|y_1, y_2) = \frac{0 + 3}{2 + 3 + 2} = \frac{3}{7}.$$