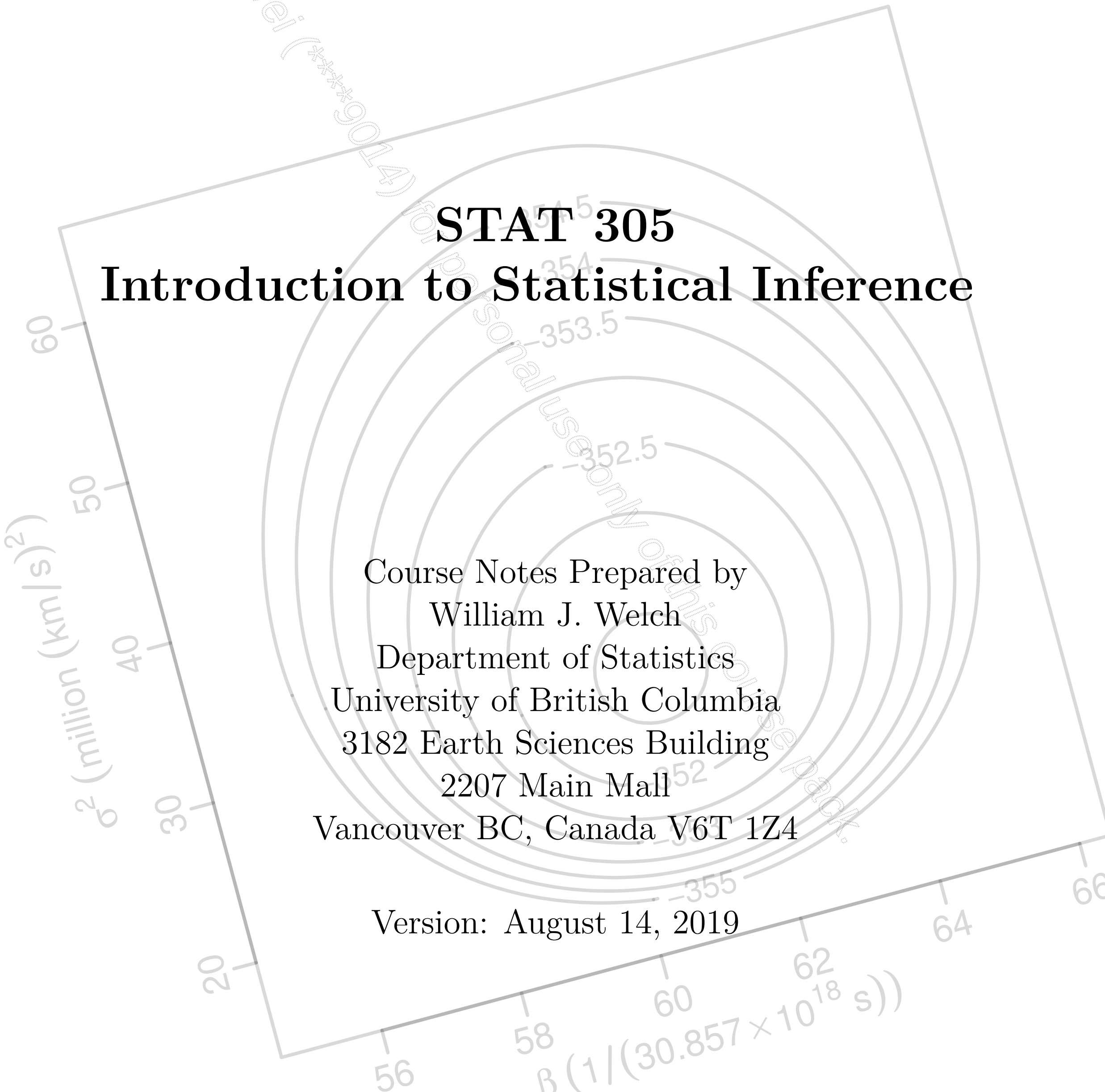
Statistical Inference: A Primer on Likelihood and Bayesian Methods



STAT 305
Introduction to Statistical Inference

Course Notes Prepared by
William J. Welch
Department of Statistics
University of British Columbia
3182 Earth Sciences Building
2207 Main Mall
Vancouver BC, Canada V6T 1Z4

Version: August 14, 2019

Contents

List of Tables

List of Figures

List of Definitions

List of Lemmas and Theorems

List of Examples

Abbreviations

Abbreviation	Description
CLT	Central limit theorem
GLR	Generalized likelihood ratio
IID	Independent and identically distributed
LR	Likelihood ratio
ML	Maximum likelihood
MSE	Mean squared error
PDF	Probability density function
PMF	Probability mass function

Greek Symbols

The following Greek letters are used in the book. Pronunciations by statisticians vary but are often close to those given here.

Case			Case		
Lower	Upper	Pronunciation	Lower	Upper	Pronunciation
α	A	al-fah	ν	N	new
β	B	bay-tah	ξ	Ξ	zie (rhymes with pie)
γ	Γ	gam-ah	o	O	oh-my-kron
δ	Δ	del-tah	π	Π	pie
ϵ	E	ep-si-lon	ρ	R	roe
ζ	Z	zay-tah	σ	Σ	sig-mah
η	H	ay-tah	τ	T	tow (rhymes with now)
θ	Θ	thay-tah	υ	Υ	up-sigh-lon
ι	I	eye-oh-tah	ϕ	Φ	fie (rhymes with pie)
κ	K	kap-ah	χ	X	kie (rhymes with pie)
λ	Λ	lam-dah	ψ	Ψ	sigh
μ	M	mew	ω	Ω	oh-me-gah

Chapter 1

Probability Tools

Statistical methods are strongly dependent on probability tools. Indeed, a statistical method typically starts and ends with probability models. The first step is to specify a probability model for the way the data were generated, and the last step often involves a calculation such as looking up a probability to compute a confidence interval or a Bayesian credible interval. In between, much of statistical inference is concerned with the unknown parameters of the probability model, which has possibly been refined along the way.

Thus, statistics and probability are intertwined, and this chapter reviews the probability tools we will need for statistical inference. It starts with one random variable and general properties like expectation and variance. The specific properties of some common probability models—those we will use frequently in later chapters—are collected together as a resource. Most statistical work involves samples of more than one observation, and hence we also need to review results for several random variables, including their joint distribution and properties of their sum or arithmetic mean. Finally, the chapter outlines the use of moment generating functions as a relatively simple tool for obtaining properties, particularly those of sums and linear functions of random variables, as needed for statistical work involving sample totals or sample means.

1.1 Discrete and Continuous Random Variables

In our journey through this book we will meet random variables that take either discrete values (e.g., integers) or continuous values (e.g., positive real numbers). In both instances we denote the random variable by an upper case letter like Y and its values by the corresponding lower case letter, y.

1.1.1 Probability mass function and probability density function

The distribution of Y over its possible values is denoted by $f_Y(y)$.

For a discrete random variable, $f_Y(y)$ can be interpreted as $\Pr(Y = y)$, the probability that Y takes the value y, and $f_Y(y)$ is called a probability mass function (PMF). The mass function is positive and sums to 1 over the possible y values.

Example 1.1 (Poisson PMF)
If Y has a Poisson distribution, it has possible values $y = 0, 1, \ldots, \infty$ and PMF

$$f_Y(y) = \frac{e^{-\mu}\mu^y}{y!}.$$

The Poisson distribution is actually a family of distributions depending on the value of the parameter $\mu > 0$, and we will use the notation $\mathsf{Pois}\,(\mu)$ to denote the family. (The properties of the Poisson and other commonly used distributions will be summarized in Sections 1.4 and 1.5.) In practice, the value of μ is usually unknown for a specific application, and much of our statistical work will be about how to estimate the values of parameters like μ from a sample of data.

The Poisson PMF sums to 1, as required:

$$\sum_{y=0}^{\infty} f_Y(y) = \sum_{y=0}^{\infty} \frac{e^{-\mu}\mu^y}{y!} = e^{-\mu} \sum_{y=0}^{\infty} \frac{\mu^y}{y!} = e^{-\mu}e^{\mu} = 1,$$

because of the series representation $1 + \mu + \mu^2/2! + \cdots$ for e^{μ}. ◊◊◊

(The end of an example is marked by a ◊◊◊ symbol.)

For a continuous random variable, $f_Y(y)$ is called a probability density function (PDF), and $f_Y(y)$ cannot be interpreted as a probability. It is, however, *proportional* to the probability that Y falls in a small interval around y (Exercise 1.1). The density function is positive and integrates to 1 over the range of possible y values.

Example 1.2 (Exponential distribution: PDF)
If Y has an exponential distribution, it has possible values $0 < y < \infty$ and PDF

$$f_Y(y) = \lambda e^{-\lambda y} \quad (0 < y < \infty; \lambda > 0).$$

The distribution, denoted $\mathsf{Expon}\,(\lambda)$, depends on the parameter $\lambda > 0$.

The exponential PDF integrates to 1, as required:

$$\int_0^{\infty} f_Y(y)\,dy = \int_0^{\infty} \lambda e^{-\lambda y}\,dy = -e^{-\lambda y}\Big|_{y=0}^{y=\infty} = 0 - (-1) = 1.$$ ◊◊◊

A distribution is symmetric if there exists μ such that its PMF or PDF can be written

$$f_Y(\mu - x) = f_Y(\mu + x),$$

for values x that generate all possible values y.

Example 1.3 (Normal distribution: symmetry)
A random variable with a normal distribution has PDF

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}$$

over possible values $-\infty < y < \infty$, for given constants μ and $\sigma^2 > 0$. The PDF satisfies

$$f_Y(\mu - x) = f_Y(\mu + x) \quad (0 \leq x < \infty)$$

and hence is symmetric around μ for any value of σ^2.

1.1.2 Cumulative distribution function

For either a continuous or a discrete random variable, Y, the cumulative distribution function (CDF) is defined as

$$F_Y(y) = \Pr(Y \leq y).$$

For a particular value y, the probability will be evaluated by summation (discrete Y) or integration (continuous Y) over values up to y. For a continuous random variable, it does not matter whether the CDF is defined as $\Pr(Y \leq y)$ or $\Pr(Y < y)$.

For a discrete random variable, there is usually little choice but to sum the PDF explicitly to compute the CDF. For instance, the Poisson CDF evaluated at, say, $y = 3$ is

$$F_Y(3) = \Pr(Y \leq 3) = \sum_{y=0}^{3} \frac{e^{-\mu}\mu^y}{y!},$$

and not much simplification is possible.

For some commonly met continuous distributions, however, simple expressions for the CDF are available by integrating the PDF. Conversely, the PDF is obtained by differentiating the CDF.

Example 1.4 (Exponential distribution: CDF)
Let Y have an $\mathsf{Expon}(\lambda)$ distribution. From the definition of the CDF,

$$F_Y(y) = \Pr(Y \leq y) = \int_0^y f_Y(t)\,dt = \int_0^y \lambda e^{-\lambda t}\,dt = -e^{-\lambda t}\Big|_{t=0}^{t=y} = -e^{-\lambda y} - (-1) = 1 - e^{-\lambda y}.$$

(Here, t is a dummy variable as we want to integrate over all values t of Y up to y.)

Similarly, we can go from the CDF to the PDF:

$$\frac{dF_Y(y)}{dy} = \frac{d}{dy}(1 - e^{-\lambda y}) = \lambda e^{-\lambda y} = f_Y(y). \qquad \diamond\diamond\diamond$$

Statisticians sometimes find it convenient to work in terms of the *survival* function or *survivor* function,

$$S_Y(y) = \Pr(Y > y) = 1 - F_Y(y).$$

It is just the complement of the CDF.

Grade on final (y)	$f_Y(y)$
60	0.2
90	0.8

Table 1.1: Probability mass function for the final-exam grade of a randomly chosen student in a given section of a statistics course

1.2 Mean, Median, and Mode

Much statistical analysis is concerned with estimating an average or typical value to represent a distribution of possible values. There are several definitions of "average".

1.2.1 Mean or expectation

The *mean* or *expected value* of a random variable Y is just a weighted average over the possible values, y, with the weights given by $f_Y(y)$.

Definition 1.1 (Expectation (mean))
The expected value or mean of a random variable Y is given by the sum

$$\mathrm{E}(Y) = \sum_y y f_Y(y)$$

if Y takes discrete values, or by the integral

$$\mathrm{E}(Y) = \int y f_Y(y)\, dy$$

if Y takes continuous values. The integral or sum is over all possible values y.

Example 1.5 (Final-exam grade: expectation)
As a simple illustration of expectation of a discrete random variable, let Y represent the grade on the final exam of a randomly chosen student from a given section of a statistics course. For simplicity, let us say Y can take only two values, 60% and 90%. The probability mass function, $f_Y(y)$, for Y is given in Table 1.1.

Definition 1.1 immediately gives

$$\mathrm{E}(Y) = 60(.2) + 90(.8) = 84,$$

i.e., the mean grade of students is 84%. This example shows that the so-called expected value does not have to be a possible value of the random variable.◇◇◇

Example 1.6 (Uniform distribution: expectation)
If Y has a uniform distribution, it has possible values $a < y < b$, for given constants a and b, and PDF

$$f_Y(y) = \frac{1}{b-a} \quad (a < y < b; a < b).$$

The distribution is denoted by $\mathsf{Unif}(a, b)$.

From Definition 1.1, the expectation or mean of Y is

$$\mathrm{E}(Y) = \int_a^b y f_Y(y)\,dy = \int_a^b y \frac{1}{b-a}\,dy = \frac{y^2}{2(b-a)}\Bigg|_{y=a}^{y=b} = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}.$$

◊◊◊

In later probability and statistical results we will often have a condition that a property, like expectation, of a random variable has to *exist*. The condition is just requiring that the expectation is defined, that is, if and only if $\sum_y |y| f_Y(y)$ (discrete) or $\int |y| f_Y(y)\,dy$ (continuous) is finite. To illustrate this technicality, consider the Poisson distribution,

$$f_Y(y) = \frac{e^{-\mu}\mu^y}{y!} \quad (y = 0, 1, \ldots, \infty;\ \mu > 0),$$

where μ is a parameter controlling the shape of the distribution. The expectation is

$$\mathrm{E}(Y) = \sum_{y=0}^{\infty} y \frac{e^{-\mu}\mu^y}{y!}.$$

It may look like this sum diverges, because the infinite sum averages y values tending to infinity. But the growth in y (and possibly μ^y) is dominated by $1/y!$, which decreases much more rapidly. Thus, the sum converges to a finite quantity, and the expectation is μ (Exercise 1.4). The notation μ is often used for the mean of a random variable in general.

In contrast, take the distribution

$$f_Y(y) = \frac{6}{\pi^2}\frac{1}{y^2} \quad \text{for } y = 1, 2, \ldots, \infty,$$

where $\pi \simeq 3.14159$ (not a parameter). This is a valid PMF, because its values are positive and sum to 1. If we try to calculate

$$\mathrm{E}(Y) = \sum_{y=1}^{\infty} y \frac{6}{\pi^2}\frac{1}{y^2},$$

however, the sum does not converge ($\sum_{y=1}^{\infty} 1/y$ is divergent). Here, the PMF does not decay fast enough to offset the growth in the value of y; the expectation is *infinite*. This simple illustration shows that not every PMF or PDF yields an expected value.

A constant a has expectation a. This is seen by applying Definition 1.1 to the degenerate discrete random variable A that takes value a with probability 1.

y	0	1	2	3
PMF	0.421875	0.421875	0.140625	0.015625
CDF	0.421875	0.843750	0.984375	1.000000

Table 1.2: Binomial PMF and CDF for $n = 3$ trials and probability of success $\pi = 1/4$

1.2.2 Median

The median m of a random variable Y essentially divides its range such that the total probability of 1 is divided equally left and right of m. Thus, from the CDF, m satisfies $F_Y(m) = \Pr(Y \leq m) = 1/2$ or $F_Y(m) \simeq 1/2$. The latter approximation arises because there may be no solution m exactly satisfying $F_Y(m) = 1/2$ when Y is discrete. The definition of the median has to accommodate such cases.

Definition 1.2 (Median of a distribution)
The median m of a random variable Y is the smallest possible value of Y such that $F_Y(m) = \Pr(Y \leq m) \geq 1/2$. For a continuous random variable with continuous CDF, m is the solution of $F_Y(m) = 1/2$.

The definition is thus straightforward for continuous distributions.

Example 1.7 (Exponential distribution: median)
As $F_Y(y) = 1 - e^{-\lambda y}$, the median m satisfies $1 - e^{-\lambda m} = 1/2$. Rearrangement gives $e^{-\lambda m} = 1/2$, then $-\lambda m = -\ln(2)$, and finally $m = \ln(2)/\lambda$. ♢♢♢

For a discrete random variable, there are rules for some special cases, but m is usually found by enumeration.

Example 1.8 (Binomial distribution: median)
The binomial distribution with n trials and probability of "success" π has PMF

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

Suppose $n = 3$ and $\pi = 1/4$, for which the PMF and CDF are given computed in Table 1.2. It is seen that the $y = 1$ is the smallest value such that $F_Y(y) \geq 1/2$, and the median is $m = 1$. Also note that $F_Y(1) = \Pr(Y \leq 1) \geq 1/2$ and $\Pr(Y \geq 1) = 1 - 0.421875 \geq 1/2$, and in this sense $m = 1$ divides the total probability of 1 into 2 halves. ♢♢♢

1.2.3 Mode

The mode of a distribution is a value maximizing the PMF or PDF. It may not be unique.

Example 1.9 (Binomial distribution: mode)
The binomial distribution with $n = 3$ trials and probability of success $\pi = 1/4$ has the PMF computed in Table 1.2. We see that the PMF is maximized by both $y = 0$ and $y = 1$. Hence, there are two modal values. ♢♢♢

1.3 Variance

1.3.1 Computation

The variance of Y is the expected (mean) of the squared deviation of Y around its mean.

Definition 1.3 (Variance)
The variance of Y is

$$Var(Y) = \mathrm{E}\left(Y - \mathrm{E}(Y)\right)^2,$$

where the expectation on the right is with respect to the distribution of Y. An equivalent definition, often used, is

$$Var(Y) = \mathrm{E}(Y^2) - (\mathrm{E}(Y))^2.$$

The definition of variance requires computation of expectations, which are handled by referring back to Definition 1.1.

For a discrete random variable, expectation and hence variance are computed by summation over all the possible values y:

$$\mathrm{Var}(Y) = \mathrm{E}\left(Y - \mathrm{E}(Y)\right)^2 = \mathrm{E}\left(Y - \mu\right)^2 = \sum_y (y - \mu)^2 f_Y(y)$$

or, equivalently,

$$\mathrm{Var}(Y) = \mathrm{E}(Y^2) - (\mathrm{E}(Y))^2 = \mathrm{E}(Y^2) - \mu^2 = \sum_y y^2 f_Y(y) - \mu^2,$$

where $\mu = \mathrm{E}(Y)$.

Example 1.10 (Final-exam grade: variance)
For the distribution $f_Y(y)$ of final grades in Table 1.1, we already computed $\mathrm{E}(Y) = \mu = 84$. Hence,

$$\mathrm{Var}(Y) = \mathrm{E}\left(Y - \mu\right)^2 = (60 - 84)^2(.2) + (90 - 84)^2(.8) = 144.$$

Alternatively, using the second definition of variance,

$$\mathrm{Var}(Y) = \mathrm{E}(Y^2) - \mu^2 = (60)^2(.2) + (90)^2(.8) - (84)^2 = 144.$$

To see the equivalence of the definition in general for discrete random variables, we just expand the square in the first definition and rearrange the sum for the expectation:

$$\begin{aligned}\mathrm{E}(Y-\mu)^2 &= \sum_y (y-\mu)^2 f_Y(y) = \sum_y (y^2 - 2\mu y + \mu^2) f_Y(y) \\ &= \sum_y y^2 f_Y(y) - 2\mu \sum_y y f_Y(y) + \mu^2 = \mathrm{E}(Y^2) - 2\mu\mathrm{E}(Y) + \mu^2 \\ &= \mathrm{E}(Y^2) - 2\mu\mu + \mu^2 = \mathrm{E}(Y^2) - \mu^2.\end{aligned}$$

For a continuous random variable, summation is again replaced by integration, and

$$\mathrm{Var}(Y) = \mathrm{E}\left(Y-\mu\right)^2 = \int (y-\mu)^2 f_Y(y)\, dy$$

or, equivalently,

$$\mathrm{Var}(Y) = \mathrm{E}(Y^2) - (\mathrm{E}(Y))^2 = \int y^2 f_Y(y)\, dy - \mu^2.$$

The equivalence is shown in the same way as for a discrete random variable.

Example 1.11 (Uniform distribution: variance)
Let Y have a $\mathsf{Unif}\,(a,b)$ distribution, i.e., it has PDF

$$f_Y(y) = \frac{1}{b-a} \quad (a < y < b; a < b).$$

From Example 1.6 we already know that $\mathrm{E}(Y) = (a+b)/2$.

To use the second expression for the variance in Definition 1.3, we also need

$$\mathrm{E}(Y^2) = \int_a^b y^2 f_Y(y)\, dy = \int_a^b y^2 \frac{1}{b-a}\, dy = \left.\frac{y^3}{3(b-a)}\right|_{y=a}^{y=b} = \frac{b^3-a^3}{3(b-a)} = \frac{a^2+b^2+ab}{3}.$$

Hence,

$$\mathrm{Var}(Y) = \mathrm{E}(Y^2) - (\mathrm{E}(Y))^2 = \frac{a^2+b^2+ab}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}. \qquad \diamondsuit\diamondsuit\diamondsuit$$

The variance exists only if the sum or integral converges. The expectation must exist for the variance to exist.

A constant a has variance 0. This is seen by applying Definition 1.3 to the degenerate discrete random variable A that takes value a with probability 1:

$$\mathrm{Var}(A) = \mathrm{E}(A - \mathrm{E}(A))^2 = (a-a)^2 \times 1 = 0.$$

Often, σ^2 is used as notation for a variance.

1.3.2 Standard deviation

The standard deviation, often denoted by σ, is

$$\text{sd}(Y) = \sqrt{\text{Var}(Y)}.$$

As the variance and standard deviation of a random variable are trivially related, we can use either. For mathematical manipulation, it is often easier to work with variances. For example, variances, not standard deviations, add for independent random variables (Section 1.7.7). On the other hand, when reporting results the standard deviation is easier to interpret because it has the same physical units as Y and not the square of the original units. For instance, the variance of the exam grade in Example 1.10 is $144\%^2$ and having units of $\%^2$ is bizarre. We could also say that the standard deviation of Y is

$$\text{sd}(Y) = \sqrt{\text{Var}(Y)} = 12\%,$$

which is much easier to interpret. Hence, we will switch back and forth between variance and standard deviation.

1.3.3 Chebyshev's inequality

Chebyshev's inequality uses the variance to bound how far a random variable, Y, can deviate from its mean in the following probabilistic sense.

Theorem 1.1 (Chebyshev's inequality)
Let the random variable Y have a distribution such that the mean and variance, μ and σ^2, exist. Then

$$\Pr(|Y - \mu| > t) \leq \frac{\sigma^2}{t^2},$$

for any $t > 0$.

The result holds for *any* distribution for Y, and hence the probability bound on the right can be weak. Nonetheless, if Y has a small enough variance then there is only a small probability that Y is more than an arbitrary distance from its mean, an argument used to prove the law of large numbers in Theorem 3.1, for instance.

1.4 Commonly Used Discrete Distributions

Table 1.3 summarizes some commonly used discrete distributions, along with their expectations and variances. It also gives their moment generating functions (to be developed in Section 1.8). In the table, parameters of the distributions (e.g., the parameter π of the Bernouilli distribution) are denoted by lower-case Greek letters if they are usually unknown in practice (and hence estimated in statistical inference)

Distribution and notation	PMF, $f_Y(y)$	E(Y)	Var(Y)	MGF, $M_Y(t)$
Bernoulli $\mathsf{Bern}\,(\pi)$	$f_Y(0) = 1-\pi$, $f_Y(1) = \pi$ $(y = 0, 1; 0 < \pi < 1)$	π	$\pi(1-\pi)$	$1-\pi+\pi e^t$ $(-\infty < t < \infty)$
Binomial $\mathsf{Bin}\,(n, \pi)$	$\binom{n}{y}\pi^y(1-\pi)^{n-y}$ $(y = 0, 1, \ldots, n;$ $n = 1, 2, \ldots;$ $0 < \pi < 1)$	$n\pi$	$n\pi(1-\pi)$	$(1-\pi+\pi e^t)^n$ $(-\infty < t < \infty)$
Geometric $\mathsf{Geom0}\,(\pi)$	$(1-\pi)^y\pi$ $(y = 0, 1, \ldots, \infty;$ $0 < \pi < 1)$	$\frac{1-\pi}{\pi}$	$\frac{1-\pi}{\pi^2}$	$\frac{\pi}{1-(1-\pi)e^t}$ $(-\infty < t < -\ln(1-\pi))$
Geometric $\mathsf{Geom1}\,(\pi)$	$(1-\pi)^{y-1}\pi$ $(y = 1, 2, \ldots, \infty;$ $0 < \pi < 1)$	$\frac{1}{\pi}$	$\frac{1-\pi}{\pi^2}$	$\frac{e^t\pi}{1-(1-\pi)e^t}$ $(-\infty < t < -\ln(1-\pi))$
Negative binomial $\mathsf{NegBin}\,(n, \pi)$	$\binom{y-1}{n-1}(1-\pi)^{y-n}\pi^n$ $(y = n, n+1, \ldots, \infty;$ $n = 1, 2, \ldots, \infty;$ $0 < \pi < 1)$	$\frac{n}{\pi}$	$\frac{n(1-\pi)}{\pi^2}$	$\left(\frac{e^t\pi}{1-(1-\pi)e^t}\right)^n$ $(-\infty < t < -\ln(1-\pi))$
Poisson $\mathsf{Pois}\,(\mu)$	$\frac{e^{-\mu}\mu^y}{y!}$ $(y =$ $0, 1, \ldots, \infty; \mu > 0)$	μ	μ	$e^{\mu(e^t-1)}$ $(-\infty < t < \infty)$

Table 1.3: Some commonly used discrete distributions, along with their expectations, variances, and moment generating functions (MGFs)

or by Roman lower-case letters if they are usually known quantities. (The Greek alphabet, with pronunciations, is given on page xxi.)

The distributions in Table 1.3 are now briefly described.

1.4.1 Bernoulli distribution

A Bernoulli random variable has only two possible outcomes, coded as 0 ("no", "absent", "failure", etc.) or 1 ("yes", "present", "success", etc.), with probabilities $1-\pi$ and π, respectively. Thus, the PMF can be represented as

$$f_B(b) = \pi^b(1-\pi)^{1-b} \quad (b = 0, 1; 0 < \pi < 1).$$

The Bernoulli distribution $\mathsf{Bern}\,(\pi)$ is the building block for the remaining discrete distributions, which can all be thought of as counting the number of "successes" ($B = 1$) observed from independent Bernoulli events. (We will refer to the event $B = 1$ generically as a "success" when outlining the remaining distributions.)

1.4.2 Binomial distribution

The binomial distribution counts the number of "successes" among a fixed number, n, of independent and identically distributed (IID) Bernoulli trials, each of which is a success or not. Thus, $Y \sim \mathsf{Bin}\,(n, \pi)$ is $\sum_{i=1}^{n} B_i$, where the B_i are independent $\mathsf{Bern}\,(\pi)$. The binomial distribution is perhaps the most important discrete distribution, because Y/n is the sample proportion, of interest in numerous applications. For instance, the efficacy of an experimental drug might be assessed by the proportion of patients in a clinical trial of n patients who respond positively to the drug.

1.4.3 Geometric distribution

A random variable with a geometric distribution arises from a sequence of IID Bernoulli trials. It counts the number of trials until one success is observed.

There are two equivalent versions of the geometric distribution; the one used is just a matter of convenience for the application. The difference is whether the terminating trial with a success outcome is counted. A $\mathsf{Geom0}\,(\pi)$ random variable does *not* count the terminating successful trial, and hence takes values $0, 1, 2, \ldots$ for the number of failures observed. The $\mathsf{Geom1}\,(\pi)$ version *does* count the terminating trial, so there must be at least one trial, and the random variable takes values $1, 2, \ldots$. Thus, the $\mathsf{Geom0}\,(\pi)$ versus $\mathsf{Geom1}\,(\pi)$ notation indicates whether the support of the random variable starts at 0 or 1.

1.4.4 Negative-binomial distribution

A negative-binomial random variable $Y \sim \mathsf{NegBin}(n, \pi)$ arises as the sum of n independent $\mathsf{Geom1}(\pi)$ random variables. Thus it counts the number of Bernoulli trials until n successes have occurred. The $\mathsf{Geom1}(\pi)$ distribution is the special case of the $\mathsf{NegBin}(n, \pi)$ distribution with $n = 1$.

The negative-binomial distribution is also related to the binomial in the following sense. If $Y \sim \mathsf{NegBin}(n, \pi)$, then Y represents a random sample size to achieve a fixed number, n, of successes. The binomial switches what is random and what is fixed: $Y \sim \mathsf{Bin}(n, \pi)$ represents a random number of successes for a fixed sample size, n.

1.4.5 Poisson distribution

A Poisson random variable can be thought of as a limiting case of the binomial. If $Y \sim \mathsf{Bin}(n, \pi)$, and we take the limits $n \to \infty$ and $\pi \to 0$ such that $\mu = n\pi$ tends to a constant, then $Y \sim \mathsf{Pois}(\mu)$ is the limiting distribution. Thus, the Poisson distribution is called the law of rare events: the probability of a success is vanishingly small, but a proper distribution arises because there are many such potential events.

The parameter μ is the mean *and* variance, which can be restrictive for applications. Often empirical data suggest that the variance is larger than the mean, so-called "over-dispersion". Thus, even when first principles suggest a Poisson probability model, a distribution with more flexibility in the mean-variance relationship, such as the negative-binomial, might be substituted.

1.5 Commonly Used Continuous Distributions

Table 1.4 summarizes some commonly used continuous distributions, which we now describe briefly.

1.5.1 Beta distribution

The beta distribution takes values on $(0, 1)$ and hence is useful for modelling quantities that must lie on that interval. It finds particular utility in Chapter 6 on Bayesian inference, where uncertainty about a parameter representing a probability is often treated as a beta random variable. In that chapter we shall see that the parameters a and b of the $\mathsf{Beta}(a, b)$ distribution make the shape of its PDF fairly flexible.

The beta function,

$$B(a, b) = \int_0^1 y^{a-1}(1 - y)^{b-1}\, dy, \tag{1.1}$$

is the normalizing factor of the beta distribution.

Distribution and notation	PDF, $f_Y(y)$	E(Y)	Var(Y)	MGF, $M_Y(t)$
Beta $\mathsf{Beta}\,(a,b)$	$\frac{1}{B(a,b)}y^{a-1}(1-y)^{b-1}$ $(0<y<1; a>0; b>0)$	$\frac{a}{a+b}$	$\frac{ab}{(a+b)^2(a+b+1)}$	Not useful
Chi-squared χ^2_d	$\frac{1}{2^{d/2}\Gamma(d/2)}y^{d/2-1}e^{-y/2}$ $(y>0;\ d=1,2,\ldots)$	d	$2d$	$\frac{1}{(1-2t)^{d/2}}$ $(-\infty<t<\frac{1}{2})$
Exponential $\mathsf{Expon}\,(\lambda)$	$\lambda e^{-\lambda y}$ $(y>0; \lambda>0)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{\lambda}{\lambda-t}$ $(-\infty<t<\lambda)$
Fisher's F F_{d_1,d_2}	$\frac{(d_1/d_2)^{d_1/2}y^{d_1/2-1}}{B\left(\frac{d_1}{2},\frac{d_2}{2}\right)\left(1+\frac{d_1}{d_2}y\right)^{\frac{d_1+d_2}{2}}}$ $(y>0;\ d_1,d_2=1,2,\ldots)$	$\frac{d_2}{d_2-2}$ $(d_2>2)$	$\frac{2d_2^2(d_1+d_2-2)}{d_1(d_2-2)^2(d_2-4)}$ $(d_2>4)$	Does not exist
Gamma $\mathsf{Gamma}\,(\nu,\lambda)$	$\frac{1}{\Gamma(\nu)}\lambda(\lambda y)^{\nu-1}e^{-\lambda y}$ $(y>0; \nu>0; \lambda>0)$	$\frac{\nu}{\lambda}$	$\frac{\nu}{\lambda^2}$	$\left(\frac{\lambda}{\lambda-t}\right)^{\nu}$ $(-\infty<t<\lambda)$
Laplace $\mathsf{Lap}\,(\mu,\phi)$	$\frac{1}{2\phi}e^{-\frac{\lvert y-\mu\rvert}{\phi}}$ $(-\infty<y<\infty; -\infty<\mu<\infty; \phi>0)$	μ	$2\phi^2$	$\frac{e^{\mu t}}{1-\phi^2t^2}$ $(\lvert t\rvert<1/\phi)$
Log-normal $\mathsf{logN}\,(\mu,\sigma^2)$	$\frac{1}{\sqrt{2\pi}\sigma y}e^{-\frac{1}{2\sigma^2}(\ln(y)-\mu)^2}$ $(y>0;\ \mu>0;\ \sigma^2>0)$	$e^{\mu+\sigma^2/2}$	$(e^{\sigma^2}-1)e^{2\mu+\sigma^2}$	Does not exist at $t=0$
Normal $\mathsf{N}\,(\mu,\sigma^2)$	$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}(y-\mu)^2}$ $(-\infty<y<\infty;$ $-\infty<\mu<\infty;\ \sigma^2>0)$	μ	σ^2	$e^{\mu t+\frac{1}{2}\sigma^2t^2}$ $(-\infty<t<\infty)$
Student's t t_d	$\frac{1}{B\left(\frac{1}{2},\frac{d}{2}\right)\sqrt{d}}\left(1+\frac{y^2}{d}\right)^{-\frac{d+1}{2}}$ $(-\infty<y<\infty;$ $d=1,2,\ldots)$	0 $(d>1)$	$\frac{d}{d-2}$ $(d>2)$	Does not exist
Uniform (rectangular) $\mathsf{Unif}\,(a,b)$	$\frac{1}{b-a}$ $(a<y<b; a<b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{bt}-e^{at}}{(b-a)t}$ $(-\infty<t<\infty)$

Table 1.4: Some commonly used continuous distributions, along with their expectations, variances, and moment generating functions (MGFs)

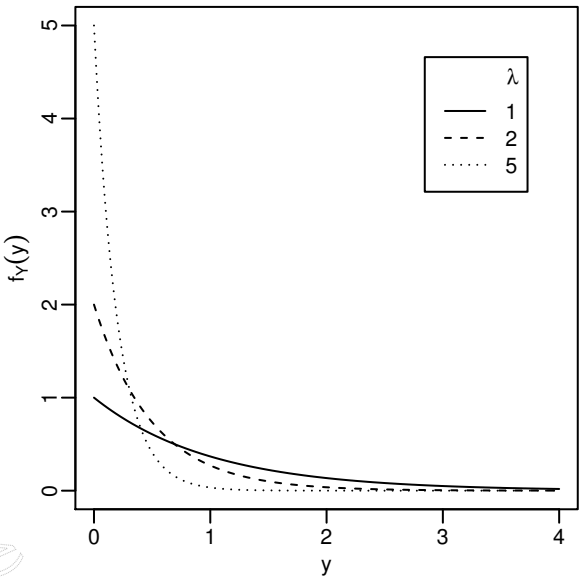


Figure 1.1: PDF of the exponential distribution with rate parameter λ taking values 1, 2, or 5

1.5.2 Exponential distribution

The PDF of the exponential distribution has a parameter λ called the *rate,* and we denote the PDF by $\mathsf{Expon}\,(\lambda)$. As its name suggests, the PDF is exponentially decreasing, as illustrated in Figure 1.1. As the rate increases, the distribution has more mass to the left. For instance, if Y is an exponential distribution representing the time to occurrence of an event, then a larger value of λ says that the rate at which the event occurs is faster and Y tends to take smaller values. Mathematically, $\mathrm{E}(Y) = 1/\lambda$ for the exponential distribution, i.e., the mean decreases with increasing rate (Exercise 1.7), which is summarized in Table 1.4 along with other properties. Hence, if Y is measured in days say, $\mathrm{E}(Y)$ also has units of days, and $\lambda = 1/\mathrm{E}(Y)$ has units 1/day, a rate per day. That is why λ is often called the "rate" parameter.

1.5.3 Gamma distribution

The gamma PDF is a generalization of the exponential PDF: putting $\nu = 1$ in $\mathsf{Gamma}\,(\nu, \lambda)$ gives $\mathsf{Expon}\,(\lambda)$. As with the exponential distribution, λ is interpreted as a rate, but the extra parameter ν controls the shape of the distribution. Figure 1.2 shows that the gamma PDF is skewed like the exponential but approaches a symmetric, bell-shape as ν increases.

A further connection between the exponential and gamma distributions is that a sum of ν IID $\mathsf{Expon}\,(\lambda)$ random variables has a $\mathsf{Gamma}\,(\nu, \lambda)$ distribution, a result demonstrated in Example 1.31.

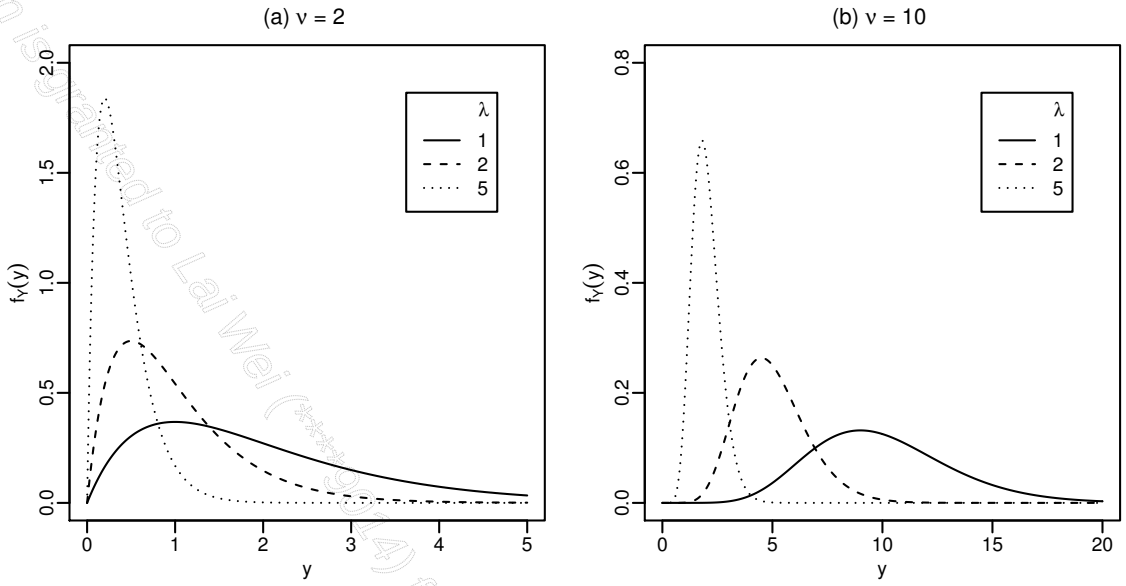


Figure 1.2: PDF of the gamma distribution with rate parameter λ taking values 1, 2, or 5: (a) shape parameter $\nu = 2$ and (b) shape parameter $\nu = 10$

The normalizing factor in the denominator of the gamma PDF is the gamma function,

$$\Gamma(\nu) = \int_0^\infty y^{\nu-1} e^{-y}\, dy \quad (\nu > 0). \tag{1.2}$$

It has the following properties.

- $\Gamma(1) = 1$ and $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ (Exercise 1.8).
- $\Gamma(\nu + 1) = \nu\Gamma(\nu)$ (by integration by parts).
- For integer $\nu > 0$, from the previous result and $\Gamma(1) = 1$, we have $\Gamma(\nu+1) = \nu!$.

The gamma and beta functions are related via $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a + b)$.

If Y has a $\mathsf{Gamma}\,(\nu, \lambda)$ distribution, then $Z = 1/Y$ has an inverse-gamma distribution with PDF

$$f_Z(z) = \frac{1}{\Gamma(\nu)} \frac{1}{z} \left(\frac{\lambda}{z}\right)^\nu e^{-\lambda/z} \quad (0 < z < \infty; \nu > 0; \lambda > 0)$$

(Exercise 1.9). Here ν is the shape parameter of the gamma distribution, but λ is now called the *scale* (not rate).

The gamma and inverse-gamma distributions are much used in Bayesian estimation of the parameters of other distributions (Chapter 6).

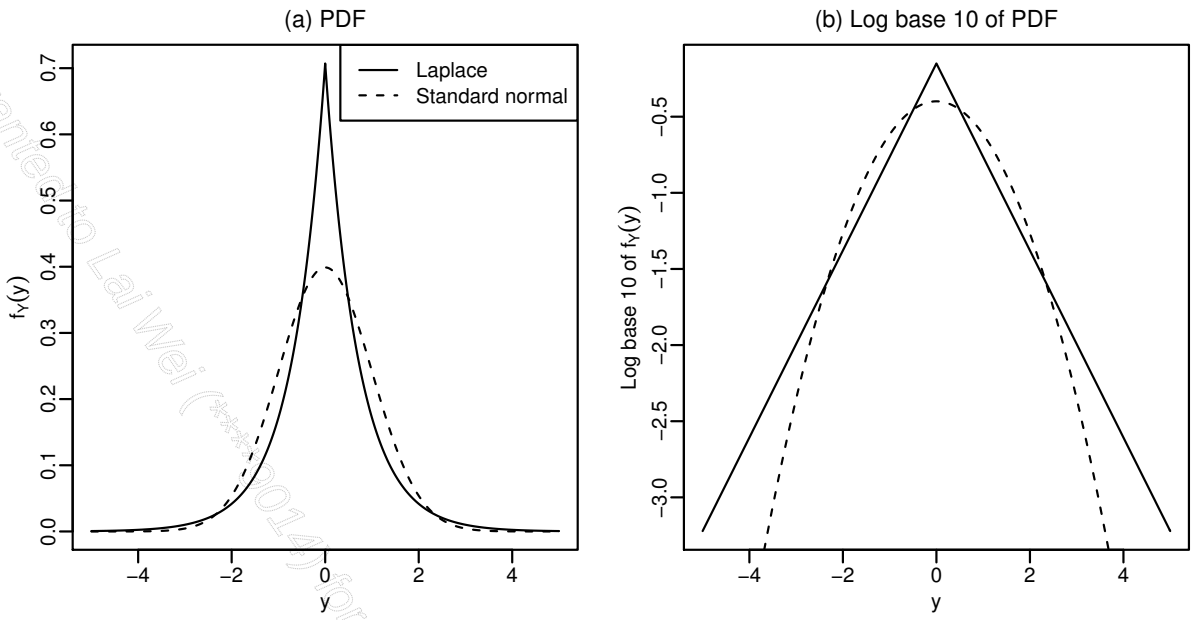


Figure 1.3: PDFs of the Laplace and normal distributions: $\mathsf{Lap}\left(\mu = 0, \phi = 1/\sqrt{2}\right)$ (solid line) versus $\mathsf{N}\left(\mu = 0, \sigma^2 = 1\right)$ (dashed line). The two distributions have the same variance: $2\phi^2 = \sigma^2 = 1$. (a) PDF and (b) log base 10 of the PDF

1.5.4 Laplace distribution

The Laplace distribution is also known as the double-exponential, because the PDF decays exponentially on the left and on the right of the location parameter μ. Note that the decay, $\exp(-|y - \mu|/\phi)$, is a function of the absolute distance from μ, in contrast to the normal distribution's decay with squared distance, $\exp(-\frac{1}{2\sigma^2}(y-\mu)^2)$. Thus, even if the two distributions have the same variance ($2\phi^2 = \sigma^2$), the Laplace distribution has fatter tails, as illustrated in Figure 1.3. The use of log scale for the PDFs in Figure 1.3(b) emphasizes that the Laplace PDF is relatively much larger in the tails. The Laplace distribution is therefore useful in statistics for modelling data with outlying observations.

1.5.5 Normal distribution

The normal distribution has great importance in statistical work, and Chapter 2 is devoted to it.

1.5.6 Log-normal distribution

By definition, Y has a log-normal distribution denoted $\mathsf{logN}\left(\mu, \sigma^2\right)$ distribution if $Z = \ln(Y)$ has a $\mathsf{N}\left(\mu, \sigma^2\right)$ distribution. Note that in the definition, μ and σ^2 are

the mean and variance *after* applying the log transformation, and not the mean and variance of Y.

Having positive support, the log-normal distribution is useful for modelling quantities such as losses in actuarial science, precipitation over a period of time, etc.

1.5.7 χ^2, F, and t distributions

These distributions arise from IID normal random variables, particularly through their sample mean and sample variance. Properties of the χ^2, F, and t distributions are developed in Chapter 2.

1.5.8 Uniform distribution

The uniform distribution is a special case of the beta distribution: a $\mathsf{Beta}\,(0,0)$ random variable has a $\mathsf{Unif}\,(0,1)$ distribution, and the latter can easily be rescaled to have range (a,b). Like the beta, the uniform finds most utility in this book for Bayesian inference (Chapter 6), where a uniform distribution on a parameter is often taken to represent no prior information about the value of the parameter.

1.6 Function of a Random Variable

1.6.1 PDF of a function of a continuous random variable

Suppose the PDF of a random variable Y is known, but we are interested in the function $g(Y)$. It is easy to write down the PDF of the new random variable $g(Y)$, if $g(\cdot)$ is a monotonic function.

Example 1.12 (PDF of a scaled exponential random variable)
Let $Y \sim \mathsf{Expon}\,(\lambda)$, where the notation "$\sim$" stands for "is distributed as". For instance, suppose Y is the time in years between earthquakes in a region. Then, as described in Section 1.5, λ is interpreted as the rate of occurrences per year. If we change the time scale to months, then Y becomes $12Y$, a simple function of Y. What is the PDF of $Z = 12Y$?

From the definition of the CDF,

$$F_Z(z) = \Pr(Z < z) = \Pr(Z/12 < z/12) = \Pr(Y < z/12) = F_Y(z/12) = 1 - e^{-\lambda z/12},$$

where the last equality is obtained by evaluating the exponential CDF for Y at $y = z/12$. Then differentiating,

$$f_Z(z) = \frac{dF_Z(z)}{dz} = \frac{d}{dz}(1 - e^{-\lambda z/12}) = (\lambda/12)e^{-(\lambda/12)z}.$$

This is seen to be the PDF of an exponential random variable, except that the original rate of occurrence λ per year becomes $\lambda/12$ per month, which is intuitive.
◇◇◇

The derivation in Example 1.12 used an explicit expression for the exponential CDF, and it may be easier to see the argument that way, but closer inspection shows that the CDF could be used implicitly. Writing $Z = g(Y)$, where $g(Y) = 12Y$, and $Y = g^{-1}(Z)$, where $g^{-1}(Z) = Z/12$, we see that the argument boils down to differentiating the CDF of Y as a function of z and applying the chain rule:

$$f_Z(z) = \frac{dF_Z(z)}{dz} = \frac{dF_Y(g^{-1}(z))}{dz} = \frac{dg^{-1}(z)}{dz} f_Y(g^{-1}(z)).$$

The CDF of Y is not explicitly required, as differentiating the CDF returns to the PDF, a handy feature for distributions like the normal with no closed form for the CDF.

This type of computation can be done in general for transformations $Z = g(Y)$, where $g(\cdot)$ is a differentiable, monotonic function:

$$f_Z(z) = \left| \frac{dg^{-1}(z)}{dz} \right| f_Y(g^{-1}(z)), \tag{1.3}$$

where the absolute value takes care of monotonic decreasing functions.

Example 1.13 (PDF of the log-normal distribution from the normal) By definition, Z has a log-normal distribution denoted $\mathsf{logN}\,(\mu, \sigma^2)$ if $Y = \ln(Z)$ has a $\mathsf{N}\,(\mu, \sigma^2)$ distribution.

Thus, Y and Z are related by the monotonic functions $Z = g(Y) = \exp(Y)$ and $Y = g^{-1}(Z) = \ln(Z)$, and the log-normal PDF of Z may be obtained via (1.3) from the normal PDF of Y in Table 1.4:

$$f_Z(z) = \left| \frac{dg^{-1}(z)}{dz} \right| f_Y(g^{-1}(z)) = \left| \frac{d\ln(z)}{dz} \right| f_Y(\ln(z)) = \frac{1}{z} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(\ln(z)-\mu)^2}.$$

◇◇◇

1.6.2 Expectation of a function of a random variable

Suppose the random variable Y is transformed to

$$Z = g(Y),$$

where $g(\cdot)$ is a known function. Because Y is random, so is Z, and the distribution of Z has properties such as expectation. An extension of Definition 1.1 gives $\mathrm{E}(g(Y))$.

Definition 1.4 (Expectation of a function of random variable) *The expected value of $g(Y)$ is given by the sum*

$$\mathrm{E}(g(Y)) = \sum_y g(y) f_Y(y)$$

if Y takes discrete values, or by the integral

$$\mathrm{E}(g(Y)) = \int g(y) f_Y(y)\, dy$$

if Y takes continuous values. The integral or sum is over all possible values y.

Again, the expectation is defined only if $\sum_y |g(y)| f_Y(y)$ or $\int |g(y)| f_Y(y)\, dy$, respectively, is finite. Thus, depending on the function $g(\cdot)$, Z could have a well-defined (finite) expectation whether Y does or not.

Example 1.14 (Expectation of log uniform)
A chemist represents her uncertainty about the concentration of a chemical species by thinking of it as a random variable, Z. Chemists often work on log scales for concentrations, and she thinks $Y = \ln(Z)$ has a continuous uniform distribution,

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

where a and b are known bounds. (Log base 10 would be used in practice by chemists, but it's easier mathematically to work with natural logs.) But what is the expected value of the *unlogged* concentration, $Z = g(Y) = e^Y$? We have

$$\mathrm{E}(Z) = \int_a^b g(y) f_Y(y)\, dy = \int_a^b e^y \frac{1}{b-a}\, dy = \frac{1}{b-a} \int_a^b e^y\, dy = \frac{1}{b-a}(e^b - e^a).$$

Suppose $a = \ln(10^{-4})$ and $b = \ln(10^{-2})$, for example, which correspond to unlogged concentrations from $10^{-4}M$ to $10^{-2}M$ (M = "mole"). Then the expected concentration is

$$\mathrm{E}(Z) = \frac{10^{-2} - 10^{-4}}{\ln(10^{-2}) - \ln(10^{-4})} = \frac{.0099}{4.605} = 0.0021M.$$

Note there is no need to compute the PDF of Z to obtain its mean here.

Alternatively, if we do the work to find the PDF of Z first, applying the result (1.3) with $g^{-1}(Z) = \ln(Z)$ we have

$$f_Z(z) = \left| \frac{d \ln(z)}{dz} \right| f_Y(\ln(z)) = \frac{1}{z} \frac{1}{b-a} \quad (e^a < z < e^b).$$

Then we can find $\mathrm{E}(Z)$ from its PDF:

$$\mathrm{E}(Z) = \int_{e^a}^{e^b} z f_Z(z) d\, z = \int_{e^a}^{e^b} z \frac{1}{z} \frac{1}{b-a} d\, z = \frac{1}{b-a} \int_{e^a}^{e^b} d\, z = \frac{1}{b-a}(e^b - e^a).$$

This is the same result as before. ♢♢♢

A well-known use of the expectation of a function of a random variable is computing the random variable's variance. From Definition 1.3, we can write

$$\mathrm{Var}(Y) = \mathrm{E}(Y^2) - (\mathrm{E}(Y))^2.$$

The term $\mathrm{E}(Y^2)$ is the expectation of the function $g(Y) = Y^2$, which is computed just as in Definition 1.4: see Example 1.11 for instance.

1.7 Several Variables

1.7.1 Joint and marginal distributions

Data for statistical models are usually multiple observations, which are considered realizations of random variables for deriving statistical properties of quantities such as sample means and proportions. Thus, probability results for several random variables are of interest. For simplicity, we concentrate here mainly on properties of two random variables; extensions to more than two are fairly immediate.

Suppose the two random variables are Y and Z. If both are discrete, the joint PMF, written $f_{Y,Z}(y, z)$, is the probability that Y takes the value y and Z takes the value z. We could also write

$$f_{Y,Z}(y, z) = \Pr(Y = y \cap Z = z),$$

where the intersection symbol "$\cap$" denotes "and". The joint PMF sums to 1 over all possible y and z values. For continuous random random variables, the joint PDF $f_{Y,Z}(y, z)$ integrates to 1 over the possible y and z values. The following rules also apply to one discrete and one continuous random variable with appropriate summation or integration.

The marginal distribution of one of the variables is given by summing or integrating over the other variable. For example, the marginal distribution of Y is

$$f_Y(y) = \begin{cases} \sum_z f_{Y,Z}(y, z) & (Z \text{ is discrete}) \\ \int f_{Y,Z}(y, z)\, dz & (Z \text{ is continuous}), \end{cases} \tag{1.4}$$

where the summation or integration is over all possible values z. The result here is just a version of the law of total probability, which is discussed further in Section 1.7.2. As one and only value of Z must occur, the marginal PMF or PDF of Y for any value y is obtained by totalling the joint distribution over all possible values of z. Similarly, $f_Z(z)$ is obtained by summing or integrating over the y values.

Example 1.15 (HIV vaccination trial: joint and marginal probabilities) In recent years, studies like the one examined here have suggested that the search for an effective vaccine against HIV will eventually pay off. A trial in Thailand

Treatment (x)	HIV infected? No ($y = 0$)	Yes ($y = 1$)	Total
Placebo ($x = 0$)	8124	74	8198
Vaccine ($x = 1$)	8146	51	8197
	16 270	125	16 395

Table 1.5: HIV vaccine: two-way frequency table by treatment and HIV-infection status

reported by Rerks-Ngarm et al. (2009) compared vaccination with ALVAC and AIDSVAX against a placebo (no vaccination) in a double-blind, randomized clinical trial involving about 16 000 volunteers.

We will focus on the "modified intention to treat" data presented by the authors and summarized in Table 1.5. The data are arranged by two variables: whether a subject received a placebo (no treatment) or the vaccine, coded by $x = 0, 1$, respectively, and whether the subject is HIV positive at the end of the trial, coded by $y = 0, 1$ for no and yes, respectively.

Thus, random variables X and Y take the values $x = 0, 1$ and $y = 0, 1$, respectively. As this chapter reviews probability, we think of the 16 395 subjects in Table 1.5 as the population of interest, from which the probabilities of various events involving X and Y can be calculated. Of course, the real problem, the statistical problem from Chapter 2 on, is to estimate such probabilities, regarding the data as a random sample from a bigger population.

Considering the 16 395 subjects in the trial as the population of interest, suppose a subject is sampled at random from the 16 395. From the observed frequencies in Table 1.5, we can compute, for instance, the joint probability that X takes the value 0 *and* Y takes the value 1:

$$f_{X,Y}(0, 1) = \Pr(X = 0 \cap Y = 1) = \frac{74}{16\,395}.$$

The probabilities for the other values of X and Y are analogous.

Marginal probabilities, i.e., probabilities relating to only one variable, can be calculated directly or via (1.4). For a randomly chosen subject, for instance,

$$f_Y(1) = \Pr(Y = 1) = \frac{125}{16\,395}$$

or equivalently by summing joint probabilities over the two possible values of X,

$$f_Y(1) = f_{X,Y}(0, 1) + f_{X,Y}(1, 1) = \frac{74}{16\,395} + \frac{51}{16\,395} = \frac{125}{16\,395}.$$ ♢♢♢

1.7.2 Conditional distributions

Conditioning allows the PMF or PDF of one or more random variables to depend on the value(s) of one or more other variables. For simplicity, we will again consider just two random variables, Y and Z say. Their joint PMF or PDF is related to their marginal and conditional distributions via

$$f_{Y,Z}(y,z) = f_Y(y) f_{Z|Y}(z \mid y) = f_Z(z) f_{Y|Z}(y \mid z) \tag{1.5}$$

for all y such that $f_Y(y) > 0$ and all z such that $f_Z(z) > 0$. The symbol "$|$" is read as "given" or "conditional on". This result builds the joint distribution in two steps: the distribution of Y, then the conditional distribution of Z given the value of Y; or conversely the distribution of Z, then the conditional distribution of Y given the value of Z. The requirement "for all y such that $f_Y(y) > 0$" is there as conditioning on the value y implies that the value has occurred, which in turn implies the value is possible, and similarly the condition $f_Z(z) > 0$.

Hence, if Y and Z are continuous or discrete random variables with joint PMF or PDF $f_{Y,Z}(y,z)$, by simple rearrangement the conditional distribution of Z given the value of Y is

$$f_{Z|Y}(z \mid y) = \frac{f_{Y,Z}(y,z)}{f_Y(y)}. \tag{1.6}$$

Here we assume $f_Y(y) > 0$, which is computed as in (1.4). The other conditional distribution, $f_{Y|Z}(y \mid z)$, is analogous.

Example 1.16 (HIV vaccination trial: conditional probability)
In the context of Example 1.15, the random variable of interest is Y, the HIV-infection status, particularly how it depends on X. Technically, we will consider the probability that $Y = 1$ (HIV positive) *conditional on* or *given* the value of X. For instance, "Is the probability of HIV infection smaller for the vaccine treatment?"

We can compute directly from Table 1.5, for example, the probability that Y takes the value 1 conditional on X taking the value 0 (no treatment):

$$f_{Y|X}(1 \mid 0) = \Pr(Y = 1 \mid X = 0) = \frac{74}{8198} \simeq 0.0090.$$

Alternatively, to demonstrate the result in (1.6),

$$f_{Y|X}(1 \mid 0) = \frac{f_{Y,Z}(y,z)}{f_Y(y)} = \frac{\Pr(X = 0 \cap Y = 1)}{\Pr(X = 0)} = \frac{74/16\,395}{8198/16\,395} = \frac{74}{8198}.$$

A similar calculation shows that $f_{Y|X}(1 \mid 1) = \Pr(Y = 1 \mid X = 1) \simeq 0.0062$. So, based on these calculations, the treatment reduces the probability of being HIV positive for these 16 395 subjects. The statistical question to be addressed in later chapters is whether the apparent efficacy of the vaccine can be explained by chance variation. ◇◇◇

As already noted, the rule in (1.4) for obtaining the marginal distribution of one random variable from its joint distribution with another is a version of the law of total probability. Another version follows by rewriting $f_{Y,Z}(y,z)$ according to marginal and conditional distributions, as in (1.5). Thus, we have two ways of writing the law of total probability for two random variables. (The law is often written in terms of probabilities of events, which carries over immediately to discrete random variables. For continuous random variables, "probability" is interpreted as a PDF.)

Theorem 1.2 (Law of total probability)
Let $f_{Y,Z}(y,z)$ be the joint PMF or PDF of the random variables Y and Z with values y and z, respectively. The marginal distribution of Y is given by

$$f_Y(y) = \begin{cases} \sum_z f_{Y,Z}(y,z) & (Z \text{ is discrete}) \\ \int f_{Y,Z}(y,z)\,dz & (Z \text{ is continuous}), \end{cases}$$

or equivalently by

$$f_Y(y) = \begin{cases} \sum_z f_Z(z) f_{Y|Z}(y \mid z) & (Z \text{ is discrete}) \\ \int f_Z(z) f_{Y|Z}(y \mid z)\,dz & (Z \text{ is continuous}), \end{cases}$$

where the summation or integration is over all values z with $f_Z(z) > 0$.

1.7.3 Statistical independence

Independence of random variables is an assumption we will often, indeed nearly always, be making for the statistical models in later chapters.

Definition 1.5 (Statistical independence)
Two random variables Y and Z with joint PDF or PMF $f_{Y,Z}(y,z)$ are statistically independent if and only if the following equivalent conditions hold.

1. *The joint distribution factorizes as the product of the two marginal distributions:*

$$f_{Y,Z}(y,z) = f_Y(y) f_Z(z) \quad (\text{for all } y, z).$$

2. *The conditional distribution of Y does not depend on the value of Z:*

$$f_{Y|Z}(y \mid z) = f_Y(y) \quad (\text{for all } y \text{ and } z \text{ such that } f_Z(z) > 0).$$

3. *The conditional distribution of Z does not depend on the value of Y:*

$$f_{Z|Y}(z \mid y) = f_Z(z) \quad (\text{for all } y \text{ and } z \text{ such that } f_Y(y) > 0).$$

As the conditions are equivalent, to demonstrate independence it is sufficient to verify just one of them; note it has to hold for all possible values y and z. To show that two variables are not independent, i.e., *dependent,* it is sufficient to find one counter-example pair of y, z values in one condition.

The conditions could also be equivalently expressed in terms of CDFs. For example, the first condition becomes

$$F_{Y,Z}(y,z) = F_Y(y) F_Z(z) \quad (\text{for all } y, z).$$

In later chapters, however, we work more with PMFs and PDFs, hence the use of them in the above definition.

With more than two random variables, they are *pairwise* independent if all pairs of them satisfy the above conditions. They are *mutually* independent if their joint distribution factorizes as a product of all their marginal distributions, with similar definitions for the other equivalent conditions. When authors say just "independent", then "mutually independent" is usually assumed by default.

1.7.4 Random sample

Mutual independence is also usually implied for a "random sample" of size n from some distribution. The sample of size n comprises n random variables over possible samples, and the random variables are independent in the sense that the distribution of the second draw from the distribution does not depend on the value of the first draw, etc. As the random variables are drawn from the same distribution, we also have the "identically distributed" part of "IID". Hence, "random sample from a distribution" and "IID random variables" are usually taken as synonymous.

In contrast, random sampling from a finite population *without replacement* would give at best approximate independence: the first draw changes the membership of

the finite population, affecting the population available for the second draw, and so on.

1.7.5 Covariance

Two random variables have a covariance, and several random variables have pairwise covariances. As well as being useful in their own right, covariances are sometimes necessary to compute the variance of a linear combination of random variables.

In general, the covariance between two random variables Y and Z—discrete, continuous, or a mixture thereof—is defined as

$$\mathrm{Cov}(Y, Z) = \mathrm{E}((Y - \mu_Y)(Z - \mu_Z)) = \mathrm{E}(YZ) - \mu_Y \mu_Z. \tag{1.7}$$

where μ_Y and μ_Z are $\mathrm{E}(Y)$ and $\mathrm{E}(Z)$, respectively. The equivalence of the definitions is easily verified by multiplying out the product and applying expectation of a linear combination. The computation of $\mathrm{E}(YZ)$ is again via a weighted average of possible values, with the weights now given by the joint distribution $f_{Y,Z}(y, z)$. If Y and Z both take discrete values, for example, then

$$\mathrm{E}(YZ) = \sum_y \sum_z yz f_{Y,Z}(y, z).$$

Here the double sum is over the possible combinations of y and z values. If one or both of the random variables are continuous, then one or both of the sums becomes an integral.

From the definition of covariance, we immediately have

$$\mathrm{Cov}(Y, Z) = \mathrm{Cov}(Z, Y)$$

and

$$\mathrm{Cov}(Y, Y) = \mathrm{Var}(Y).$$

These identities are much used in mathematical manipulations.

Often, the correlation between Y and Z,

$$\rho(Y, Z) = \frac{\mathrm{Cov}(Y, Z)}{\sqrt{\mathrm{Var}(Y)}\sqrt{\mathrm{Var}(Z)}}, \tag{1.8}$$

is easier to interpret. It is on the scale $-1 \le \rho(Y, Z) \le 1$, and measures the strength of the linear relationship (negative or positive) between Y and Z.

Example 1.17 (Final-exam and quiz grades: covariance)
For the joint distribution in Table 1.1, we have already computed $\mu_Y = 84$, $\mu_Z = 71$, and $\mathrm{Var}(Y) = 144$. Similarly, $\mathrm{Var}(Z) = 189$. Using (1.7) we find that

$$\mathrm{Cov}(Y, Z) = (60 - 84)(50 - 71)(0.1) + \cdots = 36.$$

Hence, from (1.8),

$$\rho(Y, Z) = \frac{36}{\sqrt{144}\sqrt{189}} = 0.218.$$

(What features of Table 1.6 lead to a mildly positive correlation here?) ♢♢♢

If Y and Z are independent, then $\mathrm{E}(YZ) = \mathrm{E}(Y)\mathrm{E}(Z)$ and $\mathrm{Cov}(Y, Z) = 0$. The converse, that covariance of zero implies independence, is *not true* in general.

Example 1.18 (Covariance of zero does not imply independence) As a simple counter-example to a claim that zero covariance always implies independence, let $Y \sim \mathsf{Unif}\,(-1, 1)$ and $Z = Y^2$. Clearly, the distribution of Z depends on the value taken by Y, and from Definition 1.5 they are not independent. Their covariance is

$$\mathrm{Cov}(Y, Z) = \mathrm{E}(YZ) - \mathrm{E}(Y)\mathrm{E}(Z) = \mathrm{E}(Y^3) - \mathrm{E}(Y)\mathrm{E}(Z) = 0 - 0 \times \mathrm{E}(Z) = 0,$$

where $\mathrm{E}(Y) = 0$ and $\mathrm{E}(Y^3) = 0$ follow because Y is symmetric around 0. Thus, the covariance between Y and Z is zero but they are not independent. ♢♢♢

A covariance of zero *does* imply independence in the special case of the bivariate normal distribution. If Y and Z are bivariate normal with $\mathrm{Cov}(Y, Z) = 0$, then Y and Z are independent normal random variables, a result shown in Section 1.7.9.

1.7.6 Expectation of a linear combination of random variables

Linear combinations of random variables arise very frequently throughout statistical science. In general, suppose we have a linear combination,

$$a_0 + \sum_{i=1}^{n} a_i Y_i,$$

of n random variables, $Y_1, \ldots, Y_n$. Here, $a_0, \ldots, a_n$ are constants (not random). Then

$$\mathrm{E}\left(a_0 + \sum_{i=1}^{n} a_i Y_i\right) = a_0 + \sum_{i=1}^{n} a_i \mathrm{E}(Y_i). \tag{1.9}$$

All expectations must exist. Other than that requirement, there are no further conditions. In particular, the result holds whether the Y_i are independent *or not*.

Important special cases of the general result (1.9) include the expectation of a linear function of a single random variable,

$$\mathrm{E}(a_0 + a_1 Y) = a_0 + a_1 \mathrm{E}(Y)$$

Grade on final (y)	Grade from quizzes (z)		
	50	80	
60	0.1	0.1	0.2
90	0.2	0.6	0.8
	0.3	0.7	1

Table 1.6: Joint probability function, $f_{Y,Z}(y, z)$, for the final exam grade (Y) and the grade from the quizzes (Z) of a randomly chosen student in a given STAT 305 section

and the expectation of a sum of random variables,

$$\mathrm{E}(Y_1 + Y_2) = \mathrm{E}(Y_1) + \mathrm{E}(Y_2).$$

These special cases are proved as Exercise 1.14

Note also that other definitions of "average" or "location of a distribution" such as the median or mode do not obey such rules. Expectation is a sum or integral and is therefore a *linear operator*. If expectation is combined with another linear operator, like a linear combination, the order of operations can be interchanged.

When using this rule in deriving another result, it is helpful to include a statement such as "the expectation of a linear combination of random variables is the linear combination of their expectations" to explain the step in the argument.

Example 1.19 (Course grade: expectation of a linear combination)
Let Y be the final-grade random variable defined in Table 1.1. We now also have the random variable Z, the grade from the quizzes, taking the value z, which is either 50% or 80%. A randomly chosen student will have a value for Y and a value for Z.

Table 1.6 gives the joint distribution, $f_{Y,Z}(y, z)$, for this situation. For example, the probability that Y takes the value 60 and Z takes the value 50 is 0.1. By summing across the rows of the table we obtain the probability mass function $f_Y(y)$ in Table 1.1. Similarly, by summing down the columns we obtain $f_Z(z)$. (We should really write $f_Y(y)$ and $f_Z(z)$, respectively, to distinguish the two functions. Often, the arguments are used to distinguish the functions, even though this is sloppy mathematically.) The probability distributions $f_Y(y)$ and $f_Z(z)$ are called *marginal* distributions, because they are derived from the margins of the $f_{Y,Z}(y, z)$ table. It is easily verified that $\mathrm{E}(Z) = 71$.

The overall course grade, G, is important. To simplify, let us ignore the lab component and say

$$G = 0.55Y + 0.45Z.$$

Note that the weights in the linear combination are non-random.

From first principles, we could compute $\mathrm{E}(G)$ from the PMF, $f_G(g)$, for G. Using the joint distribution in Table 1.6, $f_G(g)$ is given in Table 1.7. For example, G

Grade (g)	$f_G(g)$
55.5	0.1
69.0	0.1
72.0	0.2
85.5	0.6

Table 1.7: Probability mass function for the course grade of a randomly chosen student in a given STAT 305 section

takes the value 55.5 if Y is 60 and Z is 50, with joint probability 0.1 from Table 1.6. From $f_G(g)$ it is easily verified that $\mathrm{E}(G) = 78.15$. Computing the expectation of G by first computing $f_G(g)$ would be a very tedious numerical problem, however, if many random variables were being combined to form G.

Alternatively, applying (1.9) which says that the expectation of a linear combination of random variables is the linear combination of their expectations, we have

$$\mathrm{E}(G) = \mathrm{E}(0.55Y + 0.45Z) = 0.55\mathrm{E}(Y) + 0.45\mathrm{E}(Z) = 78.15.$$

This calculation requires only the expectations of the marginal distributions; it does not require any other properties of the joint distribution. ◇◇◇

1.7.7 Variance of a linear combination of random variables

Providing all variances exist,

$$\mathrm{Var}\left(a_0 + \sum_{i=1}^{n} a_i Y_i\right) = \sum_{i=1}^{n} a_i^2 \mathrm{Var}(Y_i) + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} a_i a_j \mathrm{Cov}(Y_i, Y_j). \quad (1.10)$$

A simpler version of this result with $n = 2$ random variables is proved in Exercise 1.16.

Special cases of the result include:

$$\begin{aligned} \mathrm{Var}(Y+Z) &= \mathrm{Var}(Y) + \mathrm{Var}(Z) + 2\mathrm{Cov}(Y,Z) \\ \mathrm{Var}(Y-Z) &= \mathrm{Var}(Y) + \mathrm{Var}(Z) - 2\mathrm{Cov}(Y,Z) \\ \mathrm{Var}(a+bY) &= b^2\mathrm{Var}(Y). \end{aligned}$$

When using the general rule or special cases in deriving another result, it is helpful to explain the step in the argument by a statement such as "using the result on the variance of a linear combination of random variables".

Example 1.20 (Course grade: variance of a linear combination)
The distribution of grades in Table 1.7 gives $\mathrm{Var}(G) = 99.65$. Alternatively, from (1.10) we find

$$\mathrm{Var}(G) = (.55)^2\mathrm{Var}(Y) + (.45)^2\mathrm{Var}(Z) + 2(.55)(.45)\mathrm{Cov}(Y,Z) = 99.65.$$ ◇◇◇

Example 1.21 (Gamma distribution: mean and variance)
The gamma distribution (see Table 1.4) has PDF

$$f_Y(y) = \frac{1}{\Gamma(\nu)}\lambda(\lambda y)^{\nu-1}e^{-\lambda y} \quad (0 < y < \infty;\, \nu > 0;\, \lambda > 0),$$

which we write as $\mathsf{Gamma}\,(\nu, \lambda)$. It has a similar form to the exponential distribution, for which we have already found the mean and variance, and a similar approach could be used again.

With a slight loss of generality we can find the gamma distribution's mean and variance rather more simply, however. If the parameter ν is an integer greater than or equal to 1 (this is the loss of generality), then a $\mathsf{Gamma}\,(\nu, \lambda)$ random variable Y can be generated by:

$$Y = Y_1 + \cdots + Y_\nu,$$

where the Y_i are independent $\mathsf{Expon}\,(\lambda)$ random variables. (This result will be proved in Example 1.31.) We know Y_i has mean $1/\lambda$ and variance $1/\lambda^2$. Hence it immediately follows that a $\mathsf{Gamma}\,(\nu, \lambda)$ random variable has mean ν/λ (from (1.9)) and variance ν/λ^2 (from (1.10)). Note that because the Y_i are independent, all covariance terms in the variance calculation are zero. This result actually holds for general $\nu > 0$. ♢♢♢

Independence can be an important assumption in formal statistical models and derivations. For instance, the result (1.10) on the variance of a linear combination of random variables is applied to derive the variance of a sample mean or sample proportion from independent observations $Y_1, \ldots, Y_n$ (e.g., Exercise 1.17). But simple results are only obtained when all distinct pairs of observations are independent and hence all $\mathrm{Cov}(Y_i, Y_j)$ terms for $i \neq j$ are all zero. If the assumption of independence is false, the claimed variance of the sample mean or proportion could be highly misleading.

Furthermore, the assumption of independence is usually made out of necessity. To take account of covariance terms between any two observations in the calculation of the variance of a linear combination, one needs some insight into the structure of the covariance, insight which is often lacking. In practice, appealing to the way the data were collected—as a random sample or via randomization in an experiment—is the only feasible justification of an independence assumption.

1.7.8 Covariance between linear functions or combinations of random variables

There are analogous results for the covariance between linear functions of random variables:

$$\mathrm{Cov}(a + bY, c + dZ) = bd\mathrm{Cov}(Y, Z) \tag{1.11}$$

This is proved as Exercise 1.15.

Similarly, for linear combinations of random variables,

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

Note that in general the two linear combinations can have different numbers of random variables, n and m, respectively. When they involve the same random variables, i.e., with $m = n$ and $Y_i = Z_i$ for $i = 1, \ldots, n$, we have

$$\mathrm{Cov}\left(\sum_{i=1}^{n} a_i Y_i, \sum_{j=1}^{n} b_j Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i b_j \mathrm{Cov}(Y_i, Y_j).$$

1.7.9 Bivariate normal distribution

Two continuous random variables Y_1 and Y_2 with a bivariate normal distribution have joint PDF given by

$$f_{Y_1,Y_2}(y_1, y_2) = \frac{1}{2\pi \det^{\frac{1}{2}}(\boldsymbol{\Sigma})} \exp\left(-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})\right),$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

μ_1 and μ_2 are the means of Y_1 and Y_2, respectively, $\sigma_1 > 0$ and $\sigma_2 > 0$ are the standard deviations of Y_1 and Y_2, respectively, $-1 < \rho < 1$ is the correlation between Y_1 and Y_2, and $\det(\boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma}^{-1}$ denote matrix determinant and inverse of $\boldsymbol{\Sigma}$, respectively.

The off-diagonal element $\rho\sigma_1\sigma_2$ in the covariance matrix $\boldsymbol{\Sigma}$ is the covariance between Y_1 and Y_2. It is zero if Y_1 and Y_2 are uncorrelated, i.e., if $\rho = 0$.

The bivariate normal has the special property that a covariance of zero between the two random variables implies they are independent.

> **Lemma 1.1 (Bivariate normal: covariance of 0 implies independence)**
> *If Y_1 and Y_2 have a joint bivariate normal distribution and their covariance (correlation) is zero, then Y_1 and Y_2 are independent normal random variables.*

To show the result, assume $\rho = 0$. Independence will follow by showing that the joint distribution factorizes. First, we have

$$\det(\boldsymbol{\Sigma}) = \det\begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} = \sigma_1^2\sigma_2^2,$$

and hence $\det^{\frac{1}{2}}(\boldsymbol{\Sigma}) = \sigma_1\sigma_2$. Second,

$$\begin{aligned}(\mathbf{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu}) &= (y_1 - \mu_1, y_2 - \mu_2)\begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}^{-1}\begin{pmatrix} y_1 - \mu_1 \\ y_2 - \mu_2 \end{pmatrix} \\ &= \frac{(y_1 - \mu_1)^2}{\sigma_1^2} + \frac{(y_2 - \mu_2)^2}{\sigma_2^2}.\end{aligned}$$

Substituting these two results into the joint distribution gives

$$f_{Y_1,Y_2}(y_1,y_2) = \frac{1}{2\pi\sigma_1\sigma_2}\exp\left(-\frac{1}{2}\left(\frac{(y_1-\mu_1)^2}{\sigma_1^2}+\frac{(y_2-\mu_2)^2}{\sigma_2^2}\right)\right)$$
$$= \frac{1}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{1}{2\sigma_1^2}(y_1-\mu_1)^2\right)\times\frac{1}{\sqrt{2\pi}\sigma_2}\exp\left(-\frac{1}{2\sigma_2^2}(y_2-\mu_2)^2\right),$$

which is the product of a $\mathsf{N}(\mu_1,\sigma_1^2)$ PDF for Y_1 and a $\mathsf{N}(\mu_2,\sigma_2^2)$ PDF for Y_2. Hence Y_1 and Y_2 are independent by Definition 1.5.

Note that independence implies covariance of zero for any two random variables (see Section 1.7.5), including the bivariate normal. But the result that covariance of zero implies independence does not hold in general.

The same arguments apply to the multivariate normal with n variables: if all pairwise covariances are zero, the n random variables are mutually independent and normal.

1.8 Moment Generating Functions

1.8.1 Uses of moment generating functions

The moment generating function (MGF) is a powerful tool for proving probability results essential for statistical methods.

- We can sometimes find the distribution of a sum of IID random variables from the MGF of the underlying distribution. This is clearly useful for statistical properties of sample totals or sample means, which are sums. For instance:
  - Example 1.31 establishes that the sum of IID exponential random variables has a gamma distribution, a result used for statistical hypothesis testing in Example 7.3.
  - Exercise 1.20 shows that a sum of independent Poisson random variables has a Poisson distribution. This is again used in hypothesis testing, in Exercise 7.2.
  - The sum of IID geometric random variables has a negative-binomial distribution (Example 1.33).
- If we know the MGF of a random variable, it is easy to write down the MGF of any linear function of it. This provides an easy proof that a linear function of a normal random variable also has a normal distribution (Example 1.30), an important property.
- Using the MGF is a relatively easy way of establishing approximate normality of a sample mean or sample total under certain conditions (the central limit theorem of Theorem 2.2) and special cases like the approximation of a binomial

distribution by a normal distribution (Example 2.2). Normal approximations are widely used in statistical inference.

- The properties of the χ^2 distribution and hence the sample variance when sampling from a normal distribution are readily shown using MGFs (in Section 2.4.2).

1.8.2 Definition of the moment generating function

As its name suggests, the moment generating function generates the moments of a distribution or random variable.

Definition 1.6 (Moments of a random variable)
Let Y be a random variable. Its kth moment for $k = 1, 2, \dots$ is $\mathrm{E}(Y^k)$, which exists if the expectation is finite.

Thus, the first moment with $k = 1$ is simply $\mathrm{E}(Y)$. The first two moments, $\mathrm{E}(Y^1)$ and $\mathrm{E}(Y^2)$, give the variance from $\mathrm{Var}(Y) = E(Y^2) - (E(Y))^2$. The MGF, once found, can generate all the moments of a random variable, including these two.

The MGF is found by computing an expectation.

Definition 1.7 (Moment generating function)
Let Y be a random variable. The moment generating function (MGF) for Y is defined as

$$M_Y(t) = \mathrm{E}_Y(e^{tY}),$$

if it exists for t in a neighbourhood of 0, i.e., for t in the open interval $-T < t < T$, where $T > 0$.

Note that the expectation is with respect to the distribution of Y, and is just the expectation of a function of Y, namely e^{tY}. The parameter t is a dummy variable. The MGF has to exist in an interval around $t = 0$ because manipulations of it will involve the derivatives at $t = 0$ and Taylor series approximation at $t = 0$.

Example 1.22 (Exponential distribution: MGF)
Let Y be distributed $\mathsf{Expon}(\lambda)$. As this is a continuous random variable, we compute the expectation in the MGF via integration:

$$M_Y(t) = \mathrm{E}(e^{tY}) = \int_0^\infty e^{ty} f_Y(y)\, dy = \int_0^\infty e^{ty} \lambda e^{-\lambda y}\, dy = \int_0^\infty \lambda e^{-(\lambda - t)y}\, dy.$$

The integrand converges and the MGF exists if $\lambda - t > 0$.

Carrying out the integration is straightforward here, but this simple example is an opportunity to show a method that avoids explicit integration in more difficult cases. With the condition $\lambda - t > 0$, we can rewrite the integral as

$$\frac{\lambda}{\lambda - t} \int_0^\infty (\lambda - t) e^{-(\lambda - t)y}\, dy.$$

The integrand is now the PDF of an exponential random variable with parameter $\lambda - t$, and like any PDF it must integrate to 1. Thus, the MGF of the $\mathsf{Expon}\,(\lambda)$ distribution is

$$M_Y(t) = \frac{\lambda}{\lambda - t}, \tag{1.12}$$

which exists for $t < \lambda$. We also note that $\lambda > 0$ (see Table 1.4), so the interval $t < \lambda$ includes an open interval around $t = 0$, as required by Definition 1.7. ♢♢♢

Example 1.23 (Gamma distribution: MGF)
From Table 1.4, the PDF of $Y \sim \mathsf{Gamma}\,(\nu, \lambda)$ is

$$f_Y(y) = \frac{1}{\Gamma(\nu)} \lambda (\lambda y)^{\nu - 1} e^{-\lambda y}.$$

First, we apply the definition of the MGF and simplify a little:

$$\begin{aligned} M_Y(t) &= \mathrm{E}(e^{tY}) = \int_0^\infty e^{ty} f_Y(y)\, dy = \int_0^\infty e^{ty} \frac{1}{\Gamma(\nu)} \lambda (\lambda y)^{\nu - 1} e^{-\lambda y}\, dy \\ &= \int_0^\infty \frac{1}{\Gamma(\nu)} \lambda (\lambda y)^{\nu - 1} e^{-(\lambda - t) y}\, dy, \end{aligned}$$

which exists if $t < \lambda$. The integrand is of the form $y^a e^{by}$, and we note that a is not necessarily an integer. (From Table 1.4, the parameter ν takes values $\nu > 0$ and hence $a = \nu - 1 > -1$.) There are many ways to proceed:

- Use standard methods of calculus.
- Look up a table of integrals.
- Use software such as Mathematica or Maple.
- Note that the integrand is very similar to the form of the original gamma PDF and again use the fact that a PDF integrates to 1.

The last route turns out to be easy. All we need to do is take out a factor:

$$\int_0^\infty \frac{1}{\Gamma(\nu)} \lambda (\lambda y)^{\nu - 1} e^{-(\lambda - t) y}\, dy = \left(\frac{\lambda}{\lambda - t} \right)^\nu \int_0^\infty \frac{1}{\Gamma(\nu)} (\lambda - t) ((\lambda - t) y)^{\nu - 1} e^{-(\lambda - t) y}\, dy.$$

The integrand is now the PDF of a $\mathsf{Gamma}\,(\nu, \lambda - t)$ random variable, i.e., with the parameter λ replaced by $\lambda - t$ everywhere, and the integral is 1. We are left with just the factor in front of the integral, and the MGF of a $\mathsf{Gamma}\,(\nu, \lambda)$ random variable is

$$M_Y(t) = \left(\frac{\lambda}{\lambda - t} \right)^\nu. \tag{1.13}$$

It exists for $t < \lambda$ and hence in an open interval around $t = 0$, because again $\lambda > 0$. ♢♢♢

Example 1.24 (Standard normal distribution: MGF)
Let $Z \sim \mathsf{N}(0,1)$, i.e., the standard normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$. Substituting these parameter values into the general normal PDF in Table 1.4 gives

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}.$$

Thus,

$$\begin{aligned} M_Z(t) &= \int_{-\infty}^{\infty} e^{tz} f_Z(z)\,dz = \int_{-\infty}^{\infty} e^{tz} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}\,dz = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z^2-2tz)}\,dz \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}((z-t)^2-t^2)}\,dz = e^{\frac{1}{2}t^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z-t)^2}\,dz \\ &= e^{\frac{1}{2}t^2}. \end{aligned}$$

The last integral is 1, because we see that the integrand is a normal PDF (with $\mu = t$ and $\sigma^2 = 1$). We also note that $e^{\frac{1}{2}t^2}$ and hence $M_Z(t)$ exist for $-\infty < t < \infty$. ◊◊◊

The method used in Example 1.24 can also be applied to find the MGF of $Y \sim \mathsf{N}(\mu, \sigma^2)$, i.e., a normal random variable with arbitrary mean and variance. The MGF of Y is

$$M_Y(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}. \tag{1.14}$$

The details are left to Exercise 1.23.

Example 1.25 (Binomial distribution: MGF)
The binomial distribution has PMF

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

Because it takes discrete values, the MGF is found by summation:

$$M_Y(t) = \mathrm{E}(e^{tY}) = \sum_{y=0}^{n} e^{ty} f_Y(y) = \sum_{y=0}^{n} e^{ty} \binom{n}{y} \pi^y (1-\pi)^{n-y}.$$

Minor simplification is possible by collecting together the e^{ty} and π^y terms, and the task is to evaluate

$$M_Y(t) = \sum_{y=0}^{n} \binom{n}{y} (\pi e^t)^y (1-\pi)^{n-y}. \tag{1.15}$$

For this discrete problem we next try to turn the sum into a sum of a PMF over its possible values. The expression to evaluate in (1.15) looks like the binomial PMF, but a binomial PMF involves complementary probabilities, π and $1 - \pi$,

which sum to 1. In (1.15) πe^t and $1-\pi$ do not sum to 1 for $t \neq 0$, but this is easily fixed by dividing them by their sum to create

$$\dot{\pi} = \frac{\pi e^t}{1-\pi+\pi e^t} \quad \text{and} \quad 1-\dot{\pi} = \frac{1-\pi}{1-\pi+\pi e^t}.$$

The divisor is cancelled by a factor outside the sum when we rewrite (1.15) as

$$M_Y(t) = (1-\pi+\pi e^t)^n \sum_{y=0}^{n} \binom{n}{y} \dot{\pi}^y (1-\dot{\pi})^{n-y}.$$

Here the sum is over the possible values of a $\mathsf{Bin}\,(n, \dot{\pi})$ random variable, and the sum must be 1. Thus, the MGF of the binomial distribution is

$$M_Y(t) = (1-\pi+\pi e^t)^n.$$

It exists for $-\infty < t < \infty$. ♢♢♢

1.8.3 Finding moments from the MGF

As its name suggests, the MGF for a distribution generates the moments, $\mathrm{E}(Y^k)$. The first moment is $\mathrm{E}(Y)$, the mean of Y. The second moment is $\mathrm{E}(Y^2)$; from it and the first moment, we can compute the variance, $\mathrm{Var}(Y) = \mathrm{E}(Y^2) - (\mathrm{E}(Y))^2$. Similarly, skewness, etc., can be computed from higher-order moments.

We will prove the result relating the MGF to the moments using a Taylor series expansion around $t = 0$. This explains the mysterious condition in the definition of the MGF that it needs to exist for t in an open interval around 0.

Suppose, $M_Y(t)$ exists in a neighbourhood of $t = 0$. Then,

$$M_Y^{(k)}(0) = \mathrm{E}(Y^k), \tag{1.16}$$

where $M_Y^{(k)}(0)$ is $M_Y(t)$ differentiated k times and evaluated at $t = 0$. To show this we make a Taylor series expansion of e^{tY} in the definition of the MGF:

$$\begin{aligned} M_Y(t) &= \mathrm{E}(e^{tY}) = \mathrm{E}\left(1 + tY + \frac{t^2Y^2}{2!} + \frac{t^3Y^3}{3!} + \cdots\right) && (1.17) \\ &= 1 + t\mathrm{E}(Y) + \frac{t^2}{2!}\mathrm{E}(Y^2) + \frac{t^3}{3!}\mathrm{E}(Y^3) + \cdots . && (1.18) \end{aligned}$$

Differentiating once with respect to t gives

$$M_Y^{(1)}(t) = \mathrm{E}(Y) + \frac{2t}{2!}\mathrm{E}(Y^2) + \frac{3t^2}{3!}\mathrm{E}(Y^3) + \cdots ,$$

and evaluating at $t = 0$ gives $M^{(1)}(0) = \mathrm{E}(Y)$. Similarly, differentiating twice gives

$$M_Y^{(2)}(t) = \mathrm{E}(Y^2) + \frac{6tY^3}{3!} + \cdots ,$$

and $M_Y^{(2)}(0) = \mathrm{E}(Y^2)$. The general result in (1.16) for $\mathrm{E}(Y^k)$ is just a continuation of this process. The proof given here makes it obvious how the Taylor-series expansion of e^{tY} generates powers of Y and hence the moments after taking expectation.

Example 1.26 (Exponential distribution: mean and variance via the MGF) The first and second derivatives of the exponential distribution's MGF in (1.12) are

$$M_Y^{(1)}(t) = \frac{1}{\lambda}\left(\frac{\lambda}{\lambda - t}\right)^2$$

and

$$M_Y^{(2)}(t) = \frac{2}{\lambda^2}\left(\frac{\lambda}{\lambda - t}\right)^3.$$

Putting $t = 0$ in the first expression gives $\mathrm{E}(Y) = M_Y^{(1)}(0) = 1/\lambda$. Similarly, $t = 0$ in the second expression gives $\mathrm{E}(Y^2) = M_Y^{(2)}(0) = 2/\lambda^2$, and hence $\mathrm{Var}(Y) = \mathrm{E}(Y^2) - (\mathrm{E}(Y))^2 = 1/\lambda^2$. ◊◊◊

We can compute the mean and variance of the exponential distribution directly (Exercise 1.7), so what is gained by use of the MGF in Example 1.26? The direct attack on the mean and variance involves moderately complicated integrals. In contrast the integration to find the MGF of the exponential distribution was straightforward. After some minor algebra, we recognized the integral of a PDF, which we know must be 1. Differentiating the MGF to get the moments was also easy. So we replaced nontrivial integrations with algebra and differentiation. (For a discrete random variable, potentially nontrivial summations are similarly avoided.)

Example 1.27 (Gamma distribution: mean and variance via the MGF) First, rewrite the MGF of the gamma distribution in (1.13) as

$$M_Y(t) = \left(\frac{\lambda}{\lambda - t}\right)^\nu = \left(\frac{1}{1 - t/\lambda}\right)^\nu.$$

(Application of the chain rule is then a little easier.)

The first derivative of $M_Y(t)$ is

$$M_Y^{(1)}(t) = -\nu\left(\frac{1}{1 - t/\lambda}\right)^{\nu+1}\left(-\frac{1}{\lambda}\right) = \frac{\nu}{\lambda}\left(\frac{1}{1 - t/\lambda}\right)^{\nu+1},$$

which equals ν/λ at $t = 0$. Therefore,

$$\mathrm{E}(Y) = \frac{\nu}{\lambda}.$$

The second derivative is

$$M_Y^{(2)}(t) = \frac{\nu}{\lambda}(-(\nu+1))\left(\frac{1}{1 - t/\lambda}\right)^{\nu+2}\left(-\frac{1}{\lambda}\right) = \frac{\nu(\nu+1)}{\lambda^2}\left(\frac{1}{1 - t/\lambda}\right)^{\nu+2},$$

which evaluates to $\nu(\nu+1)/\lambda^2$ at $t=0$. This is $\mathrm{E}(Y^2)$. Therefore,

$$\mathrm{Var}(Y) = \mathrm{E}(Y^2) - (\mathrm{E}(Y))^2 = \frac{\nu(\nu+1)}{\lambda^2} - \left(\frac{\nu}{\lambda}\right)^2 = \frac{\nu}{\lambda^2}.$$

The same expectation and variance were found in Example 1.21, but the result here holds for any $\nu > 0$ and not just for positive integer values of ν. ♢♢♢

Example 1.28 (Binomial distribution: mean and variance via the MGF) From Example 1.25 the MGF of the binomial distribution is

$$M_Y(t) = (1-\pi+\pi e^t)^n.$$

The first two derivatives of the MGF are

$$M_Y^{(1)}(t) = n\pi e^t(1-\pi+\pi e^t)^{n-1}$$

and

$$M_Y^{(2)}(t) = n\pi e^t(1-\pi+\pi e^t)^{n-1} + n\pi e^t(n-1)\pi e^t(1-\pi+\pi e^t)^{n-2}.$$

Evaluating these derivatives at $t=0$ gives

$$\mathrm{E}(Y) = n\pi$$

and

$$\mathrm{E}(Y^2) = n\pi + n(n-1)\pi^2.$$

Therefore,

$$\mathrm{Var}(Y) = \mathrm{E}(Y^2) - (\mathrm{E}(Y))^2 = n\pi + n(n-1)\pi^2 - (n\pi)^2 = n\pi(1-\pi). \quad ♢♢♢$$

The argument to obtain the moments from the MGF hinges on the Taylor series expansion (1.18) at $t=0$. Thus, the MGF has to exist at $t=0$, which was straightforward to demonstrate for the examples up to here. The next example is a little more subtle.

Example 1.29 (Uniform distribution: existence of the MGF) If $Y \sim \mathsf{Unif}\,(a,b)$, Table 1.4 says its MGF is

$$M_Y(t) = \frac{e^{bt}-e^{at}}{(b-a)t} \quad (-\infty < t < \infty),$$

i.e., the table claims the MGF exists for all t including $t=0$.

The appearance of t in the denominator may create some doubt about the existence, but note that the numerator is also 0 at $t=0$. Expanding the exponential functions, however, shows that all is well, however:

$$\begin{aligned} M_Y(t) =& \frac{1+bt+(bt)^2/(2!)+(bt)^3/(3!)+\cdots-(1+at+(at)^2/(2!)+(at)^3/(3!)+\cdots)}{(b-a)t} \\ =& \frac{(b-a)t+(b^2-a^2)t^2/2+(b^3-a^3)t^3/6+\cdots}{(b-a)t} \\ =& 1+(a+b)t/2+(a^2+b^2+ab)t^2/6+\cdots, \end{aligned}$$

which equals 1 at $t = 0$. (Simplifying the t^3 term uses $(b^3 - a^3) = (b - a)(a^2 + b^2 + ab)$.)

Hence, the first two moments give the expectation and variance of Y in Table 1.4 (Exercise 1.21). ♢♢♢

1.8.4 MGF of a linear function or a sum

Lemma 1.2 (MGF of a linear function of a random variable)
If the MGF of Y is $M_Y(t)$, then $Z = a + bY$ has MGF

$$M_Z(t) = e^{at} M_Y(bt).$$

The result follows from

$$M_Z(t) = \mathrm{E}_Z(e^{tZ}) = \mathrm{E}_Y(e^{t(a+bY)}) = e^{at}\mathrm{E}_Y(e^{btY}) = e^{at} M_Y(bt).$$

We next derive a result useful for a sum of independent random variables.

Lemma 1.3 (MGF of a sum of independent random variables)
Suppose $Y_1, \ldots, Y_n$ are independent random variables, and Y_i has MGF $M_{Y_i}(t)$ (which must exist). Then the MGF of $X = Y_1 + \cdots + Y_n$ is

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

The result follows from

$$\begin{aligned} M_X(t) &= \mathrm{E}_X(e^{tX}) = \mathrm{E}_{Y_1,\ldots,Y_n}(e^{t\sum_{i=1}^n Y_i}) = \mathrm{E}_{Y_1,\ldots,Y_n}\left(\prod_{i=1}^{n} e^{tY_i}\right) = \prod_{i=1}^{n} \mathrm{E}_{Y_i}(e^{tY_i}) \\ &= \prod_{i=1}^{n} M_{Y_i}(t). \end{aligned}$$

Here, we are using the result that the expectation of a product of independent random variables is the product of expectations.

1.8.5 The MGF identifies a distribution

An important property of the MGF is that it identifies a distribution uniquely.

Theorem 1.3 (The MGF identifies a distribution)
Let Y and Z be two random variables with MGFs $M_Y(t)$ and $M_Z(t)$, respectively. If $M_Y(t) = M_Z(t)$ for all t in an open interval of 0, then $\Pr(Y \le y) = \Pr(Z \le y)$, i.e., Y and Z have the same CDF.

In other words, if we can find the MGF of a random variable, and this MGF is on a list of MGFs we have computed (as in Tables 1.3 and 1.4), we can use the MGF to tell us the random variable's distribution.

The MGF is thus a mathematical "fingerprint" for a distribution. A human fingerprint can identify a person on a list of suspects already fingerprinted. However, if the fingerprint belongs to a person not on the list of suspects, identification is not possible. It is similarly difficult to work backwards from an MGF to its distribution without a list of suspect distributions with MGFs like those in Tables 1.3 and 1.4. Fortunately, those tables suffice for most of our needs in this book.

Billingsley (2012, Sections 9 and 30) gives several proofs of Theorem 1.3, for discrete random variables and more generally. The essence of the general argument is that, if the open-interval condition of the lemma is satisfied, a probability distribution is determined by its moments, which are in turn determined by the MGF.

We use Theorem 1.3 in the following examples to find various distributions of random variables derived from a linear function of a random variable or a sum of independent random variables.

Example 1.30 (Normal distribution: linear function)
Let Y have a $\mathsf{N}(\mu, \sigma^2)$ distribution, and let $Z = a + bY$. From the MGF of Y in (1.14), the MGF of Z is

$$M_Z(t) = e^{at} M_Y(bt) = e^{at} e^{\mu(bt) + \frac{1}{2}\sigma^2 (bt)^2} = e^{(a+b\mu)t + \frac{1}{2} b^2 \sigma^2 t^2}.$$

This has the form of the MGF in (1.14) for a normal random variable, except that the mean μ is replaced by $a + b\mu$, and the variance σ^2 is replaced by $b^2\sigma^2$. Thus, Z is identified to have a $\mathsf{N}(a + b\mu, b^2\sigma^2)$ distribution. ♢♢♢

Example 1.31 (Exponential distribution: sum of IID random variables)
Let $Y_1, \ldots, Y_\nu$ be ν IID $\mathsf{Expon}(\lambda)$ random variables. Note they have the same value of the parameter λ. Thus, from (1.12) each Y_i has MGF $\lambda/(\lambda - t)$. Their sum, Y, has MGF

$$M_Y(t) = \prod_{i=1}^{\nu} \frac{\lambda}{\lambda - t} = \left(\frac{\lambda}{\lambda - t} \right)^{\nu}.$$

This was shown to be the MGF of a $\mathsf{Gamma}(\nu, \lambda)$ random variable in Example 1.23. Thus, the sum of independent exponential random variables with the same value of λ follows a gamma distribution. ♢♢♢

Example 1.32 (Normal distribution: sum of independent random variables)
Let Y_1 and Y_2 be independent normal random variables. Y_1 has mean μ_1 and variance σ_1^2; similarly, Y_2 has mean μ_2 and variance σ_2^2. The MGFs of Y_1 and Y_2 are available from (1.14). The sum $X = Y_1 + Y_2$ has MGF

$$M_X(t) = M_{Y_1}(t) M_{Y_2}(t) = e^{\mu_1 t + \frac{1}{2}\sigma_1^2 t^2} e^{\mu_2 t + \frac{1}{2}\sigma_2^2 t^2} = e^{(\mu_1 + \mu_2)t + \frac{1}{2}(\sigma_1^2 + \sigma_2^2) t^2}.$$

This is seen to be the MGF of a normal random variable with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$. Thus, we have shown that the sum of two independent normal

random variables is also normal. By repeating the argument, the obvious extension of the result holds for the sum of n independent normal random variables.

(The result also holds more generally for n random variables following a multivariate normal distribution, a generalization of the bivariate normal in Section 1.7.9, even if they are not independent. Their sum or a linear combination of them has a normal distribution. When computing the variance of the sum or linear combination via (1.10), the covariance terms will have to be included.) ♢♢♢

Example 1.33 (Casualty insurance: sum of IID geometric random variables) An actuary models the distribution of the number of months to the first claim for drivers insured in a particular high-risk rating class. She uses a geometric distribution, i.e., $\mathsf{Geom1}\,(\pi)$, where π is the probability of at least one claim in a month. A $\mathsf{Geom1}\,(\pi)$ random variable, Y, has probability mass function

$$f_Y(y) = (1-\pi)^{y-1}\pi \quad (y = 1, 2, \ldots, \infty; 0 < \pi < 1),$$

expectation $1/\pi$, and variance $(1-\pi)/\pi^2$. Its MGF is

$$M_Y(t) = \frac{e^t\pi}{1-(1-\pi)e^t}$$

(see Exercise 1.22 or Table 1.3).

Suppose the actuary is interested in the distribution of the number of months to the *second claim*. She assumes that the distribution arises as $X = Y_1 + Y_2$, where Y_1 and Y_2 are independent $\mathsf{Geom1}\,(\pi)$ random variables. (This ignores the possibility of more than one claim in a month.) What is the distribution of X?

From Lemma 1.3,

$$M_X(t) = M_{Y_1}(t)M_{Y_2}(t) = \left(\frac{e^t\pi}{1-(1-\pi)e^t}\right)^2.$$

This is the MGF of a negative-binomial random variable (again see Exercise 1.22 or Table 1.3) with parameters $n = 2$ and π, i.e., $X \sim \mathsf{NegBin}\,(2, \pi)$.

Table 1.3 gives the PMF of the negative-binomial distribution in general. With $n = 2$ and, say, $\pi = 0.1$, the first few numerical values of the PMF are given in Table 1.8.

1.9 Getting It Done in R

In later chapters of this book we have to compute PDFs, PMFs, and CDFs for a variety of distributions, such as those in Tables 1.3 and 1.4. R can compute these quantities for all commonly used distributions.

x	$f_X(x) = \binom{x-1}{2-1}(1-\pi)^{x-2}\pi^2$
2	$(2-1)(1-0.1)^0(0.1)^2 = 0.01$
3	$(3-1)(1-0.1)^1(0.1)^2 = 0.018$
4	$(4-1)(1-0.1)^2(0.1)^2 = 0.0243$
etc.	

Table 1.8: Probability mass function of a negative-binomial random variable with $n = 2$ and $\pi = 0.1$

Distribution	R function with arguments and defaults	Translation into our notation
Discrete distributions		
Binomial	`dbinom(x, size, prob)`	$\mathsf{Bin}(n = \texttt{size}, \pi = \texttt{prob})$
Geometric	`dgeom(x, prob)`	$\mathsf{Geom0}(\pi = \texttt{prob})$
Negative binomial	`dnbinom(x, size, prob)`	See text
Poisson	`dpois(x, lambda)`	$\mathsf{Pois}(\mu = \texttt{lambda})$
Continuous distributions		
Beta	`dbeta(x, shape1, shape2)`	$\mathsf{Beta}(a = \texttt{shape1}, b = \texttt{shape2})$
χ^2	`dchisq(x, df)`	$\chi^2_{d=\texttt{df}}$
Exponential	`dexp(x, rate = 1)`	$\mathsf{Expon}(\lambda = \texttt{rate})$
Fisher's F	`df(x, df1, df2)`	$F_{d_1=\texttt{df1},d_2=\texttt{df2}}$
Gamma	`dgamma(x, shape, rate = 1)`	$\mathsf{Gamma}(\nu = \texttt{shape}, \lambda = \texttt{rate})$
Normal	`dnorm(x, mean = 1, sd = 1)`	$\mathsf{N}(\mu = \texttt{mean}, \sigma^2 = \texttt{sd}^2)$
Student's t	`dt(x, df)`	$t_{d=\texttt{df}}$
Uniform	`dunif(x, min = 0, max = 1)`	$\mathsf{Unif}(a = \texttt{min}, b = \texttt{max})$

Table 1.9: R functions for the PMF or PDF of some common distributions

Table 1.9 lists the R functions to compute the PMF or PDF of the distributions in Tables 1.3 and 1.4. In all cases, the argument `x` is the value of a random variable X. The table also translates the R arguments into our notation for a particular distribution and its parameter(s). In most cases, R's syntax and our notation agree well, and X in Table 1.9 has the same distribution as the random variable Y in Table 1.3 or 1.4. There are a few exceptions, however.

1. R has no function corresponding to the $\mathsf{Geom1}(\pi)$ distribution, the version of the geometric distribution with range $y = 1, 2, \ldots, \infty$. If X is a $\mathsf{Geom0}(\pi)$ random variable, then $Y = X + 1$ is $\mathsf{Geom1}(\pi)$ random variable, however, and hence the PMF of Y is given by `dgeom(x = y - 1, prob)`, which could be called for values $y \geq 1$.

2. R's negative-binomial distribution for X is defined by the PMF

$$f_X(x) = \binom{x+n-1}{x}(1-\pi)^x \pi^n \quad (x = 0, 1, \ldots, \infty),$$

whereas Y in Table 1.3 has PMF

$$f_Y(y) = \binom{y-1}{n-1}(1-\pi)^{y-n} \pi^n \quad (y = n, n+1, \ldots, \infty).$$

Again, there is a simple relationship, and Y and $n + X$ have the same distribution. Thus, `dnbinom(x = y - n, size, prob)` for $y \geq n$ gives our negative-binomial PMF for Y.

In addition, there is no R function for the Bernoulli distribution; it is a special case of the binomial with $n = 1$ (or `size` $= 1$).

The functions in Table 1.9 are useful for sketching a PMF or PDF. For instance, the following R code plots $f_Y(y)$ against y when Y has the t distribution with $d = 10$ degrees of freedom, to produce Figure 1.4.

```
# Values of y at which to plot the PDF, namely -5, -4.99, ..., 5
y <- seq(-5, 5, by = 0.01)

# PDF for t distribution with 10 degrees of freedom
fy <- dt(y, df = 10)

# Plot fy against y using type  = "l" to join the points with lines
plot(y, fy, xlab = "y", ylab = expression(f[Y](y)), type = "l")
```

The last line is perhaps more complicated than necessary, with `expression` and `[Y]` generating the subscript Y in the y-axis label. The simpler syntax

```
plot(y, fy, xlab = "y", ylab = "f(y)", type = "l")
```

would often suffice. (Type `demo(plotmath)` to demonstrate R's capabilities to format mathematical expressions in text of plots.)

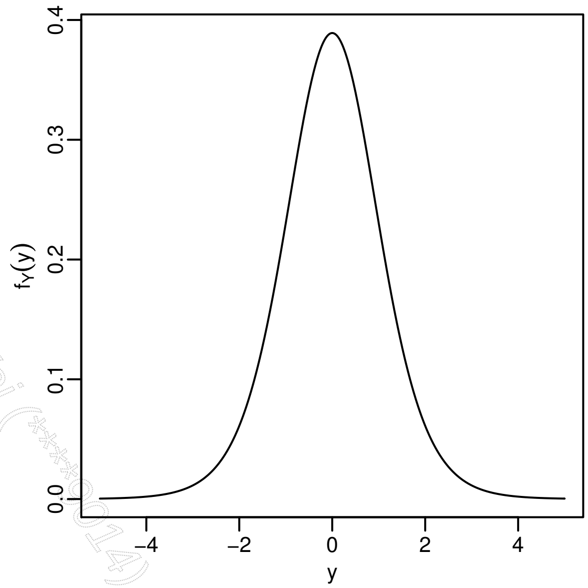


Figure 1.4: PDF of the t distribution with 10 degrees of freedom

The functions in Table 1.9 give a PMF or PDF, $f_Y(y)$, but statistical calculations often require the CDF, $F_Y(y) = \Pr(Y \le y)$. For each PMF or PDF function, which has a name starting with `d`, there is a corresponding CDF function, which has a name starting with `p`. For instance, `ppois(y, lambda)` computes the CDF, $\Pr(Y \le y)$, of the Poisson distribution.

```
> ppois(1, lambda = 2)
[1] 0.4060058
```

Here, `>` is the `R` prompt, and recall that `lambda` is μ in the $\mathsf{Pois}(\mu)$ distribution of Table 1.3. The calculation is easily verified:

$$\begin{aligned}\Pr(Y \le 1) &= \Pr(Y = 0) + \Pr(Y = 1) = \frac{e^{-\mu}\mu^0}{0!} + \frac{e^{-\mu}\mu^1}{1!} \\ &= e^{-\mu}(1 + \mu) = e^{-2}(1 + 2) = 0.4060058.\end{aligned}$$

Similarly, analogous to `dnorm(x, mean, sd)`, there is `pnorm(x, mean, sd)`, etc.

1.10 Learning Outcomes

On completion of this chapter you should be able to demonstrate the following skills. They relate to the probability mass function (PMF) or probability density function (PDF), $f_Y(y)$, of a random variable, Y.

1. Understand the relationship between a PMF or PDF and its CDF, including the interpretation of a PDF $f_Y(y)$ as proportional to the probability that Y is in a small interval around y.

2. From a given PMF or PDF of the random variable Y, write down the definition of $\mathrm{E}(Y)$ and $\mathrm{Var}(Y)$. Simplify to a closed-form expression when readily available.

3. From the PMF or PDF of Y, write down the definition of the expectation of $g(Y)$. Simplify to a closed-form expression when readily available.

4. From the mean and variance of Y, apply Chebyshev's inequality to bound how far Y can deviate from its mean in a probabilistic sense.

5. From the PDF of a random variable Y, find the PDF of a monotonic function (transformation) of Y.

6. From the PMF or PDF of a random variable Y, find the expectation of a function $g(Y)$.

7. From the joint distribution of two random variables, find their marginal and conditional distributions.

8. Check whether two random variables are independent or not from their joint distribution.

9. Find the covariance and correlation between two random variables. Interpret the correlation.

10. Understand the relationship between covariance and independence.

11. Find the expectation of a linear combination of several random variables from their individual expectations. As a special case, find the expectation of a linear function of a random variable from its expectation.

12. Find the variance of a linear combination of several random variables from their individual variances and their pair-wise covariances. As a special case, find the variance of a linear function of a random variable from its variance.

13. From the PMF or PDF of Y, find its moment generating function (MGF).

14. Check whether the MGF exists.

15. From the MGF of a random variable, find the mean and variance.

16. From the MGF of a random variable, Y, find the MGF of $a + bY$.

17. From the (common) MGF of several independent random variables, find the MGF of their sum.

18. Use the MGF of a random variable to identify its distribution.

19. Explain your reasoning. When using a result such as the expectation or variance of a linear function of a random variable or a linear combination of random variables as part of a longer derivation, briefly state the result you are using. If the result depends on an assumption such as statistical independence of random variables, remind the reader that you are using the assumption.

Perhaps just as important in practice is a list of the tasks that would *not* be tested on the quizzes and final exam, or at least will receive less emphasis in grading.

1. There is no doubt that facility with calculus, summation, and algebra is helpful for the mathematical manipulations in STAT 305. Nonetheless, STAT 305 is a course in probability and statistical inference, not mathematical manipulation. Thus, the *formulation* of, say, an expectation, is at least as important as the subsequent calculations to implement the final answer.

2. Long proofs involving many steps will not be tested. On the other hand, simple derivations that can be done in a few lines may appear. Again, it is the demonstration of the use of appropriate probability and statistics tools to carry out the derivation that is most important.

3. You are not expected to memorize specific PMFs, PDFs, and MGFs. You may put them on your formula sheet. If a PMF, PDF, or MGF is required to compute properties stemming from it, it will be given in the question. Similarly, you will not be asked, e.g., "find the mean of an exponential random variable" if the mean is required for subsequent calculations. You will be asked "to show that the mean is $1/\lambda$".

1.11 Exercises

Exercise 1.1
Let Y be a normal random variable with mean 100 and variance 5^2 (i.e., standard deviation 5).

1. Use `pnorm` in `R` to compute the following probabilities: $\Pr(90.0 < Y < 90.1)$, $\Pr(95.0 < Y < 95.1)$, and $\Pr(100.0 < Y < 100.1)$.

2. Use `dnorm` in `R` to compute the PDF of Y at the following values: $y = 90.05$, $y = 95.05$, and $y = 100.05$. (Note that these values are the midpoints of the intervals in part 1.)

3. Multiply each of the PDF values in part 2 by the width of the intervals used in part 1. What do you get? Hence, draw a picture illustrating the assertion, "For a given value of y, the probability that Y falls in a small interval around y is approximately the PDF computed at y multiplied by the width of the interval."

Exercise 1.2

Let Y be an *indicator* random variable, i.e., with possible values 0 and 1. Show $\mathrm{E}(Y) = \Pr(Y = 1)$. (Interchanging expectation and probability like this is a frequently used trick for indicator variables.)

Exercise 1.3

In Section 1.3, two definitions were given for the variance of a random variable Y:

$$\mathrm{Var}(Y) = \mathrm{E}(Y - \mathrm{E}(Y))^2$$

and

$$\mathrm{Var}(Y) = \mathrm{E}(Y^2) - (\mathrm{E}(Y))^2.$$

Show that these two expressions are equivalent.

Exercise 1.4

Let $Y \sim \mathsf{Pois}\,(\mu)$.

1. Show that $\mathrm{E}(Y) = \mu$. Hint: You can rewrite the sum in the expectation to include the factor

   $$\sum_{y=0}^{\infty} \frac{\mu^y}{y!} = e^{\mu}.$$

   This factor will cancel.

2. Show that $\mathrm{Var}(Y) = \mu$.

Exercise 1.5

Let $Y \sim \mathsf{Geom1}\,(\pi)$.

1. From the definition of expectation, show that $\mathrm{E}(Y) = 1/\pi$.
2. From the definition of variance, show that $\mathrm{Var}(Y) = (1 - \pi)/\pi^2$.
3. Show that the CDF of Y is

   $$\Pr(Y \leq y) = 1 - (1 - \pi)^y.$$

4. Hence, find the survival function, $\Pr(Y > y)$, i.e., the probability that Y exceeds or "survives" y trials.
5. Suppose it is given that Y exceeds a value y_0. Show that

   $$\Pr(Y > y_0 + y \mid Y > y_0) = \Pr(Y > y),$$

   i.e., the probability of surviving at least another y trials does not depend on the number of trials already survived. (Such a random variable is said to have a "memoryless" property.)

Exercise 1.6
Let $Y \sim \mathsf{Geom0}\,(\pi)$.

1. Find $\Pr(Y \geq 1)$.
2. Suppose it is given that $Y \geq 1$. Show that
$$\Pr(Y = y \mid Y \geq 1) = (1-\pi)^{y-1}\pi \quad (y = 1, 2, \ldots).$$
3. What is the distribution of Y conditional on $Y \geq 1$?

Exercise 1.7
Let Y have an $\mathsf{Expon}\,(\lambda)$ distribution.

1. What is $\mathrm{E}(Y)$?
2. What is $\mathrm{E}(Y^2)$ and hence what is $\mathrm{Var}(Y)$?

Exercise 1.8
Show the following properties of the normalizing factor (1.2) of the gamma PDF.

1. $\Gamma(1) = 1$.
2. $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

(Manipulation of an integrand to make it have same form as a well-known PDF is a tactic much used in Section 1.8 to evaluate moment generating functions.)

Exercise 1.9
Let Y have a $\mathsf{Gamma}\,(\nu, \lambda)$ distribution, i.e.,
$$f_Y(y) = \frac{1}{\Gamma(\nu)}\lambda(\lambda y)^{\nu-1}e^{-\lambda y} \quad (0 < y < \infty; \nu > 0; \lambda > 0),$$
where ν and λ are shape and rate parameters.
Use the result in (1.3) to show that the PDF of $Z = 1/Y$ is
$$f_Z(z) = \frac{1}{\Gamma(\nu)}\frac{1}{z}\left(\frac{\lambda}{z}\right)^{\nu} e^{-\lambda/z} \quad (0 < z < \infty; \nu > 0; \lambda > 0).$$
This is called an inverse-gamma distribution with shape parameter ν and scale parameter λ.

Exercise 1.10

Let Y be the lifetime of an item. It has an $\mathsf{Expon}\,(\lambda)$ distribution.

1. Find the survival function, $S_Y(y) = \Pr(Y > y)$.
2. Suppose at time y_0 the item is still alive, and we want to condition on the fact that $Y > y_0$. Find $\Pr(Y > y_0 + y \mid Y > y_0)$, the probability of surviving an additional time y given survival to time y_0.
3. Hence, given that $Y > y_0$, what is the distribution of the additional lifetime?

Exercise 1.11

Let $Y \sim \mathsf{logN}\,(\mu, \sigma^2)$, i.e., $Z = \ln(Y)$ has a $\mathsf{N}\,(\mu, \sigma^2)$ distribution. Using properties of the normal distribution, show that the the median of Y is e^{μ}.

Exercise 1.12

A mail-order company sends an offer to its population of customers. Let B_1 be a random variable taking the values 0 (does not buy) or 1 (buys) for a randomly chosen customer. At a later date the company sends out another offer; let B_2 be the analogous 0/1 random variable for the second offering.

The probabilities for the joint distribution of B_1 and B_2 are given in the following table.

		B_2 = 0	B_2 = 1
B_1	0	0.3	0.0
	1	0.1	0.6

1. Compute $\mathrm{E}(B_1)$ and $\mathrm{E}(B_2)$.
2. Compute $\mathrm{Var}(B_1)$ and $\mathrm{Var}(B_2)$.
3. Compute $\mathrm{Cov}(B_1, B_2)$.
4. Let $B = B_1 + B_2$ be the total number of purchases by a randomly selected customer. Compute $\mathrm{E}(B)$ and $\mathrm{Var}(B)$:
   (a) by first enumerating the distribution of B (i.e., computing $f_B(b)$ for all values b of B); and
   (b) by using results on linear combinations of random variables.

Exercise 1.13
For a specific type of property insurance claim, an actuary models the customer's loss by the random variable $X \sim \mathsf{Expon}(\lambda)$. But the particular policy has a limit k on the amount that the insurance company has to pay. Thus, when a claim is made the company pays out $Y = \min(X, k)$. What is $\mathrm{E}(Y)$?

Exercise 1.14
Let X and Y be continuous random variables with finite expectations, and let a, b, and c be finite constants. From the definition of expectation, prove the following results.

1. $\mathrm{E}(a + X) = a + \mathrm{E}(X)$.
2. $\mathrm{E}(bX) = b\mathrm{E}(X)$.
3. $\mathrm{E}(a + bX) = a + b\mathrm{E}(X)$.
4. $\mathrm{E}(X + Y) = \mathrm{E}(X) + \mathrm{E}(Y)$.
5. $\mathrm{E}(a + bX + cY) = a + b\mathrm{E}(X) + c\mathrm{E}(Y)$.
6. If X and Y are discrete random variables, how are these proofs changed?

Exercise 1.15
Let X and Y be random variables with finite covariance, and let a and b be finite constants. From the definition of covariance, prove the result

$$\mathrm{Cov}(a + bY, c + dZ) = bd\mathrm{Cov}(Y, Z)$$

in (1.11).

Exercise 1.16
Let X and Y be random variables with finite variances, and let a, b, and c be finite constants. Starting from the definition of variance, i.e., $\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2$, prove the following results. (Hint: The definition of variance is in terms of expectations; use the results of Exercise 1.14.)

1. $\mathrm{Var}(a + X) = \mathrm{Var}(X)$.
2. $\mathrm{Var}(bX) = b^2\mathrm{Var}(X)$.
3. $\mathrm{Var}(a + bX) = b^2\mathrm{Var}(X)$.
4. $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y)$.
5. $\mathrm{Var}(a + bX + cY) = b^2\mathrm{Var}(X) + c^2\mathrm{Var}(Y) + 2bc\mathrm{Cov}(X, Y)$. (This is special case of the more general result in Section 1.7.7.)

Exercise 1.17

Let $Y_1, \ldots, Y_n$ be independent random variables, each taking the values 0 or 1 with probabilities $1 - \pi$ and π, respectively. Here π, the probability that $Y = 1$, is an unknown parameter to be estimated.

1. Show that $\mathrm{E}(Y_i) = \pi$ and $\mathrm{Var}(Y_i) = \pi(1 - \pi)$.
2. Consider the estimator $\tilde{\pi} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ of π. (This is simply the proportion of 1's amongst $Y_1, \ldots, Y_n$. It is a random variable because the Y_i are random.)
   (a) Show that $\mathrm{E}(\tilde{\pi}) = \pi$, i.e., $\tilde{\pi}$ is an unbiased estimator of π.
   (b) Show that $\mathrm{Var}(\tilde{\pi}) = \pi(1 - \pi)/n$.

Exercise 1.18

[Quiz #1, 2009-10, Term 1] Let B be a Bernoulli random variable taking values $b = 0, 1$. Its PMF is given by $f_B(0) = \Pr(B = 0) = 1 - \pi$ and $f_B(1) = \Pr(B = 1) = \pi$. Thus, $B \sim \mathsf{Bern}\,(\pi)$.

Show each of the following results. For full marks you need to be explicit about the mathematical definition of the quantity involved (E(), Var() or MGF) and how the definition is used for this specific problem.

1. Show $\mathrm{E}(B) = \pi$.
2. Find $\mathrm{E}(10^B)$.
3. Show $\mathrm{Var}(B) = \pi(1 - \pi)$.
4. Show that the moment generating function (MGF) of B is $M_B(t) = 1 - \pi + \pi e^t$.
5. Check that the MGF exists for t in an open neighbourhood of zero.
6. Use the MGF to find $\mathrm{E}(B)$.

Exercise 1.19

[Quiz #1, 2009-10, Term 1] Let $Y = B_1 + \cdots + B_n$, where the random variables $B_1, \ldots, B_n$ are independent and each has a $\mathsf{Bern}\,(\pi)$ distribution. You may use the results in Exercise 1.18 that B_i has mean π, variance $\pi(1 - \pi)$, and MGF $1 - \pi + \pi e^t$. Also n is some fixed number.

1. Find $\mathrm{E}(Y)$.
2. Find $\mathrm{Var}(Y)$.
3. Find the moment generating function of Y.
4. Hence, what is the distribution of Y?

Exercise 1.20
Let $Y \sim \mathsf{Pois}\,(\mu)$. Thus, the PMF of Y is

$$f_Y(y) = \frac{e^{-\mu}\mu^y}{y!} \quad (y = 0, 1, \ldots, \infty; \mu > 0).$$

1. Show that Y has the MGF
$$M_Y(t) = e^{\mu(e^t - 1)}.$$
2. Let $Y_1, \ldots, Y_n$ be independent Poisson random variables, where Y_i has mean μ_i, i.e., $Y_i \sim \mathsf{Pois}\,(\mu_i)$. Thus, the random variables may have different means and are not necessarily identically distributed. What is the MGF of $\sum_{i=1}^n Y_i$?
3. Hence, what is the distribution of $\sum_{i=1}^n Y_i$?

Exercise 1.21
Let $Y \sim \mathsf{Unif}\,(a, b)$. Use the expansion of its MGF in Example 1.29 to show the following properties:

1. $\mathrm{E}(Y) = (a + b)/2$; and
2. $\mathrm{Var}(Y) = (b - a)^2/12$.

Exercise 1.22
Let $Y \sim \mathsf{Geom1}\,(\pi)$. Thus, the PMF of Y is

$$f_Y(y) = (1 - \pi)^{y-1}\pi \quad (y = 1, 2, \ldots, \infty; 0 < \pi < 1).$$

1. Show that Y has the MGF
$$M_Y(t) = \frac{e^t\pi}{1 - (1 - \pi)e^t}.$$
2. From the MGF show that
$$\mathrm{E}(Y) = \frac{1}{\pi} \quad \text{and} \quad \mathrm{Var}(Y) = \frac{1 - \pi}{\pi^2}.$$
3. Let $Y_1, \ldots, Y_n$ be IID $\mathsf{Geom1}\,(\pi)$ random variables. What is the MGF of $\sum_{i=1}^n Y_i$?
4. Hence, what is the distribution of $\sum_{i=1}^n Y_i$?

Exercise 1.23
Let $Y \sim \mathsf{N}(\mu, \sigma^2)$, i.e., the normal distribution with mean μ and variance σ^2. This exercise shows in two ways that the MGF of Y is

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

1. Apply the definition of the MGF in Definition 1.7 directly to the $\mathsf{N}(\mu, \sigma^2)$ PDF to find the MGF of Y.
2. Let Z have a standard normal distribution, i.e., $\mathsf{N}(0, 1)$. Its MGF is

   $$M_Z(t) = e^{\frac{1}{2}t^2}$$

   (see Example 1.24). Now let $Y = \mu + \sigma Z$.
   (a) Verify that $\mathrm{E}(Y) = \mu$ and $\mathrm{Var}(Y) = \sigma^2$ as required.
   (b) Find the MGF of Y from the MGF of Z.

Exercise 1.24
Let $Y \sim \mathsf{N}(\mu, \sigma^2)$. Starting from the MGF of Y, i.e.,

$$M_Y(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2},$$

this exercise verifies the first two moments.

1. Use the MGF to show that $\mathrm{E}(Y) = \mu$.
2. Use the MGF to show that $\mathrm{E}(Y^2) = \mu^2 + \sigma^2$, and hence that $\mathrm{Var}(Y) = \sigma^2$.

Exercise 1.25
Let $Y \sim \mathsf{Expon}(\lambda)$. Consider multiplying Y by a constant to give a new random variable, $Z = bY$, where $b > 0$.

1. Table 1.4 says the MGF of Y is $\lambda/(\lambda - t)$. What is the MGF of Z?
2. What is the distribution of Z?
3. Apply the same argument to the gamma distribution to show that if $Y \sim \mathsf{Gamma}(\nu, \lambda)$, then $Z = bY \sim \mathsf{Gamma}(\nu, \lambda/b)$.

Exercise 1.26
A random variable, Y, taking positive values is said to have a log-normal distribution if $Z = \ln(Y)$ has a $\mathsf{N}(\mu, \sigma^2)$ distribution. This exercise finds the expectation of Y from the MGF of Z.

1. The definition of the MGF of Z is $\mathrm{E}(e^{tZ})$. What expression do we get if we put $t = 1$ in this definition?
2. Look up the MGF of Z (see Table 1.4 or Exercise 1.23), and put $t = 1$ in it. Hence, what is $\mathrm{E}(Y)$?

Chapter 2

The Normal Distribution in Statistics

2.1 Introduction

The normal distribution, sometimes called the *Gaussian* distribution after Gauss, is ubiquitous in statistical analysis. It will arise in this book in several ways, including the following.

- Based on a random sample from a $\mathsf{N}(\mu, \sigma^2)$ distribution, the sample mean is often used to estimate μ. Exact properties of the normal distribution lead to a confidence interval for μ that accounts for the uncertainty in estimation, even if σ^2 is unknown. This analysis based on the t distribution is common in applications, as typified by Example 2.1.

- If a random sample is taken from a distribution with mean μ and variance σ^2, but the distribution is only approximately normal, use of the t distribution to provide a confidence interval for μ is often still approximately valid.

- For a large sample size, the normal distribution serves as an approximation to other distributions. For instance, the binomial distribution is a commonly used model for applications where estimating a population proportion is the focus. The error in using the sample proportion as an estimate is again quantified in a confidence interval, this time based on a normal approximation to the binomial distribution. More generally, under certain conditions, sample means, proportions, and totals have approximate normal distributions for a large enough sample size via the central limit theorem (Section 2.5.3).

- The method of maximum likelihood in Chapter 4 is a powerful generic method to estimate parameters for a wide range of probability models. An approximate confidence interval for the parameter is often available from an approximate normal distribution.

2.2 Some Properties of the Normal Distribution

Let Y be a normal random variable with mean μ and variance σ^2. We write $Y \sim \mathsf{N}(\mu, \sigma^2)$. The following properties were established using MGFs in Section 1.8 or in related exercises.

- The expected value (mean) of Y is $\mathrm{E}(Y) = \mu$, and the variance is $\mathrm{Var}(Y) = \sigma^2$ (Exercise 1.24).
- A linear function of Y is also normal: $a+bY \sim \mathsf{N}(a + b\mu, b^2\sigma^2)$ (Example 1.30). In particular,
$$Z = \frac{Y - \mu}{\sigma} = -\frac{\mu}{\sigma} + \frac{1}{\sigma}Y \sim \mathsf{N}(0, 1)$$
has the standard normal distribution $\mathsf{N}(0, 1)$, i.e., with mean 0 and variance 1 (Exercise 2.1).

Also, for n normally distributed random variables, we have the following properties.

- If the n random variables are *independent* (but not necessarily identically distributed), Example 1.32 showed that a linear combination of them also has a normal distribution.
- If the n random variables are *correlated* but follow a multivariate normal distribution, a linear combination of them still has a normal distribution. The covariances between the n variables affect only the variance of the linear combination, via (1.10).
- If the n random variables follow a multivariate normal distribution and all the pairwise covariances between them are zero, then they are mutually independent. This result is a generalization of Lemma 1.1, which said that if Y and Z are bivariate normal with $\mathrm{Cov}(Y, Z) = 0$, then Y and Z are independent.

2.3 Distributions Derived From the Normal

The normal distribution is important in itself and because other important distributions arise from it. The relationships among these distributions are summarized in Figure 2.1. We next give some details about these distributions.

2.3.1 The χ^2 distribution

A χ^2 ("chi-squared") random variable is generated by the sum of squares of independent standard normal random variables.

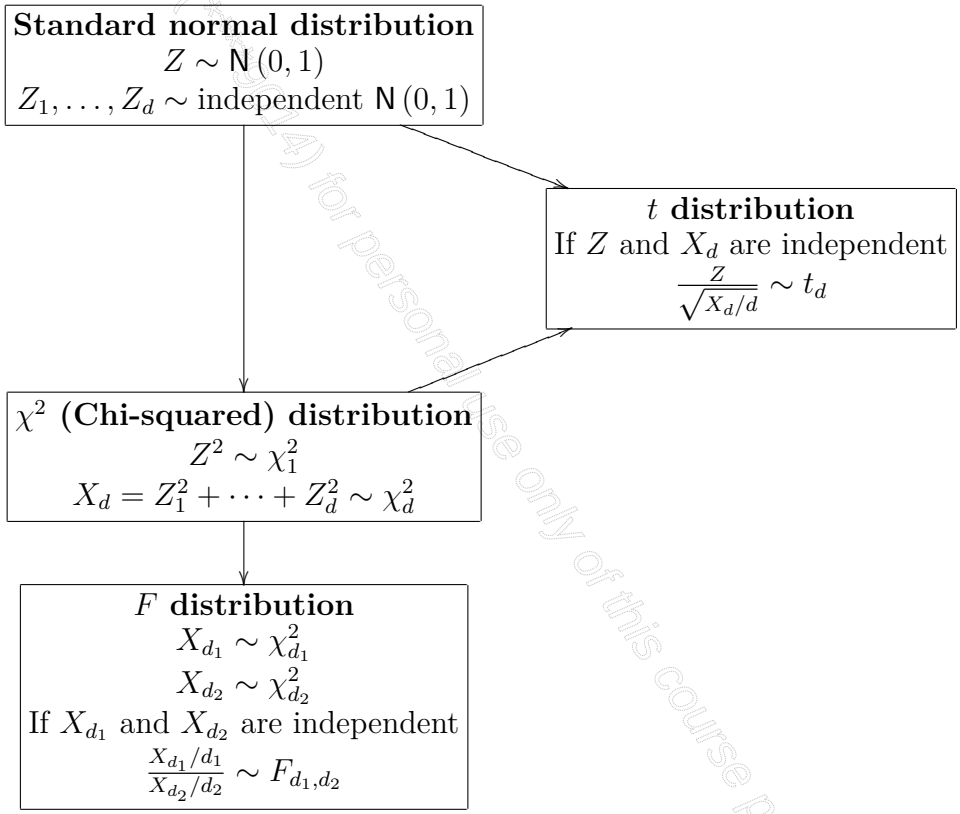


Figure 2.1: Relationships between distributions derived from the normal

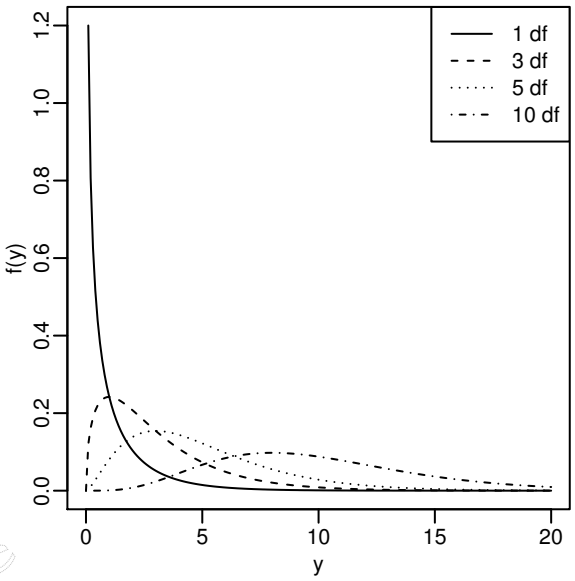


Figure 2.2: PDF of the χ^2 distribution with 1, 3, 5, and 10 degrees of freedom

Lemma 2.1 (χ^2 distribution)
If $Z_1, \ldots, Z_d$ are independent $\mathsf{N}(0,1)$ random variables, then their sum of squares,

$$X_d = Z_1^2 + \cdots + Z_d^2 \sim \chi_d^2,$$

has a χ^2 distribution with d degrees of freedom, which we write as χ_d^2.

A proof is developed in Exercise 2.6.

Figure 2.2 plots the PDF of the χ_d^2 distribution for $d = 1$, 3, 5, and 10. With $d = 1$ or $d = 2$ degrees of freedom, the PDF is monotonic decreasing. For $d \geq 3$, the PDF increases then decreases. For large d the shape of the distribution is approximately normal; this is starting to be evident for $d = 10$ in Figure 2.2. The limiting normality stems from the central limit theorem (Theorem 2.2), because of the sum in Lemma 2.1.

The χ^2 distribution is related to the gamma distribution. Let G be a $\mathsf{Gamma}(\nu = d/2, \lambda = 1)$ random variable (see Table 1.4), where $d > 0$ is an integer. Then $X = 2G$ is a χ_d^2 random variable.

The χ^2 distribution arises in inference about the variance of a normal distribution in Section 2.4.2. It is also heavily used for hypothesis testing in Chapters 7–9.

2.3.2 The t distribution

The following lemma is due to W.S. Gosset, who wrote under the pseudonym "Student".

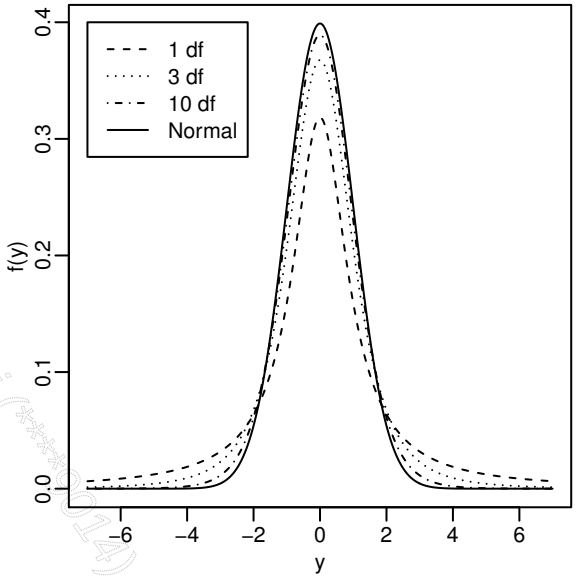


Figure 2.3: PDF of the t distribution with 1, 3, 10, and ∞ degrees of freedom; ∞ degrees of freedom gives the standard normal

Lemma 2.2 (t distribution (Student, 1908))
If $Z \sim \mathsf{N}(0,1)$, $X_d \sim \chi^2_d$, and Z and X_d are independent, then

$$\frac{Z}{\sqrt{X_d/d}} \sim t_d,$$

where t_d is Student's t distribution with d degrees of freedom.

Gosset's motivation for the lemma was the distribution of the sample mean of n IID $\mathsf{N}(\mu, \sigma^2)$ standardized using the *sample* variance as an estimator of σ^2, leading to confidence intervals and hypothesis tests for μ based upon the t distribution (Section 2.4.3). Contrary to Gosset's modest dismissal of his contribution in comments made to R.A. Fisher, the result is one of the most commonly applied in all of the statistical sciences.

Figure 2.3 plots the t_d PDF for $d = 1$, 3, and 10. The PDF has a bell shape similar to the normal, but has wider tails than the normal. As $d \to \infty$, the t_d PDF approaches the standard normal PDF. With 1 degree of freedom, the distribution is known as the Cauchy distribution.

An insightful way to prove Lemma 2.2 is to first write $Z/\sqrt{X_d/d}$ in the lemma as a mixture of normal random variables. Define

$$W = \frac{X_d}{d} \quad \text{and} \quad T = \frac{Z}{\sqrt{X_d/d}} = \frac{Z}{\sqrt{W}},$$

where T is the variable of interest and W appears in its denominator. (We write T following our convention of using upper case letters for random variables, whereas

many textbooks use t for the random variable and T for Hotelling's T^2 statistic.) Their joint distribution can be written in terms of the conditional distribution of T given W:

$$f_{T,W}(t, w) = f_{T|W}(t \mid w) f_W(w)$$

(see Section 1.7.2). The key simplification here is that conditional on W taking the value w, the variable T of interest becomes

$$\frac{Z}{\sqrt{w}} \sim \mathsf{N}\left(0, \frac{1}{w}\right),$$

i.e., $f_{T|W}(t \mid w)$ is simply the normal PDF with variance $1/w$ or standard deviation $1/\sqrt{w}$. Here we are using the assumed independence of Z and $\sqrt{W}$ in the lemma, so that Z still has a normal distribution conditional on W. Once we condition on $W = w$, there is a constant, not random, divisor in $Z/\sqrt{w}$, i.e., a simple linear transformation of a normal random variable, which is also normal. Similarly, W is a trivial rescaling of a χ^2_d random variable, and $f_W(w)$ is easily derived. The marginal distribution of T then follows by integrating out W from the joint distribution (see Section 1.7.1):

$$f_T(t) = \int_0^\infty f_{T,W}(t, w)\, dw = \int_0^\infty f_{T|W}(t \mid w) f_W(w)\, dw.$$

Written this way, the t_d PDF is an average or mixture of normal PDFs, averaged with respect to the distribution of W.

The second step is to carry out the integration. Readers interested in the formal details can find them in the Appendix of Section 2.9, but it is perhaps more instructive to demonstrate the averaging of normal random variables graphically. For definiteness, take $d = 3$ degrees of freedom. Furthermore, we will approximate the continuous values of $X_{d=3}$ and hence W by just five representative values. Figure 2.4(a) shows the χ^2_3 PDF of X_3. The distribution is divided into five equal probabilities by the dotted lines. These sub-intervals in the figure are represented by the numbers 0.6, 1.4, etc. For instance, the value 0.6 cuts off a probability of 0.1 to the left, and in this sense it is in the centre of the first interval of probability 0.2. The five representative values are obtained in `R` using the function `qchisq`.

```
> # Quantiles of X_3 cutting off probs 0.1, 0.3, etc. to the left
> x3 <- qchisq(c(0.1, 0.3, 0.5, 0.7, 0.9), df = 3)
> x3
[1] 0.5843744 1.4236522 2.3659739 3.6648708 6.2513886
```

As $W = X_3/3$ is a monotonic increasing function of X_3, we have $0.1 = \Pr(X_3 < 0.5843744) = \Pr(X_3/3 < 0.5843744/3)$, etc., and the five representative values of W are found by simple arithmetic.

```
> # Quantiles of W cutting off probs 0.1, 0.3, etc. to the left
> w <- x3 / 3
> w
[1] 0.1947915 0.4745507 0.7886580 1.2216236 2.0837962
```

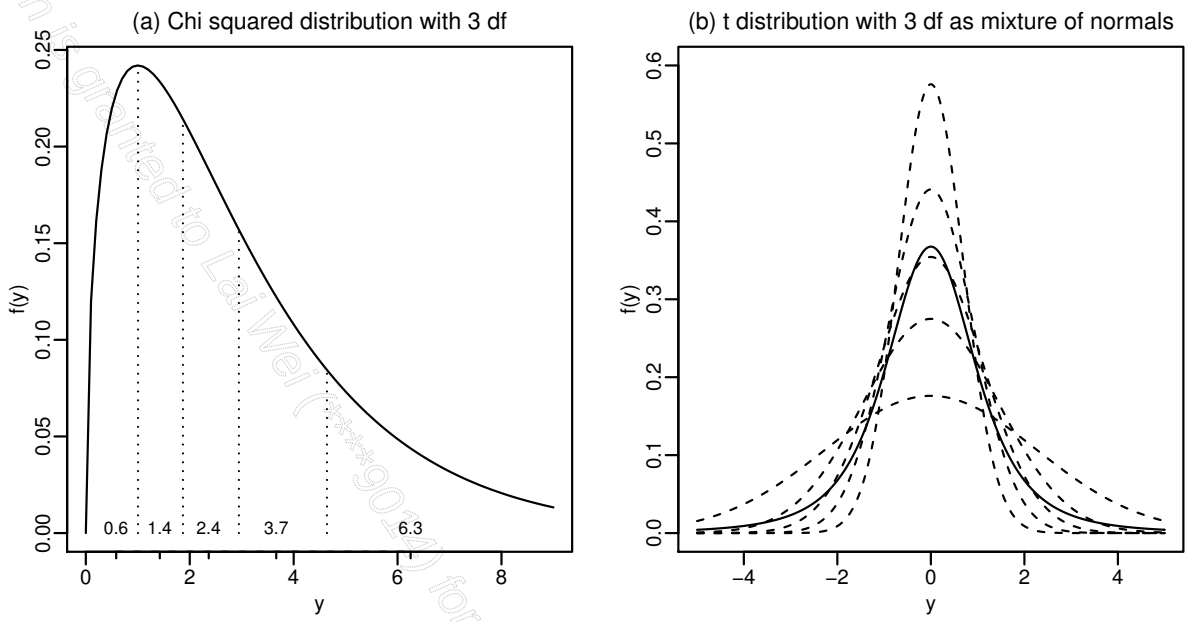


Figure 2.4: (a) χ^2_3 PDF, with the distribution divided into five subintervals of probability 0.2 each by the dashed lines. Each subinterval is represented by a single value. (b) The t_3 PDF shown as a solid line is to a good approximation given by the average of the five normal PDFs shown as dashed lines.

Finally, each representative value, w, leads to a standard deviation of $1/\sqrt{w}$ in the normal distribution.

```
> # Standard deviation of conditional normal
> sd.norm <- 1/sqrt(w)
> sd.norm
[1] 2.2657659 1.4516391 1.1260448 0.9047556 0.6927434
```

Figure 2.4(b) shows five normal PDFs as dashed lines with these five standard deviations. Note that the conditional normal distributions have considerable variation in their standard deviations. Visually averaging these five PDFs gives a curve that is hard to distinguish from the t_3 PDF shown by a solid line in the figure.

The t distribution will arise when we make inference about the mean, μ, of a normal distribution and the variance, σ^2, also has to be estimated.

2.3.3 The F distribution

The F distribution arises from the ratio of two independent χ^2 distributions.

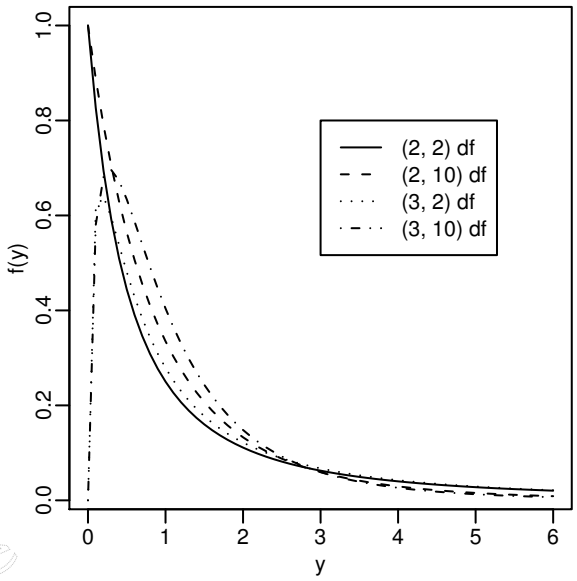


Figure 2.5: PDF of the F distribution with (2, 2) (2, 10), (3, 2), and (3, 10) degrees of freedom

Lemma 2.3 (F distribution)
If $X_{d_1} \sim \chi^2_{d_1}$, $X_{d_2} \sim \chi^2_{d_2}$, and X_{d_1} and X_{d_2} are independent, then

$$\frac{X_{d_1}/d_1}{X_{d_2}/d_2} \sim F_{d_1,d_2},$$

where F_{d_1,d_2} is an F distribution with d_1 and d_2 degrees of freedom.

Figure 2.5 plots the PDF of F_{d_1,d_2} for various values of d_1 and d_2. With $d_1 = 1$ or 2, the PDF is monotonic decreasing. For $d_1 > 2$ the distribution increases then decreases.

The F distribution is used to make inference when comparing two variances. It is also very important in inference about the parameters of linear regression models (STAT 306).

2.4 Estimating the Parameters of the Normal Distribution

Let $Y_1, \ldots, Y_n$ be independent $\mathsf{N}(\mu, \sigma^2)$ random variables. Such variables would arise if each is an independent draw from the same normal distribution. Equivalent to this probability model, a statistician might say $Y_1, \ldots, Y_n$ are assumed to be a random sample of size n from an $\mathsf{N}(\mu, \sigma^2)$ distribution. Here, the statistician has in mind an infinite—in practice, large—population of values taken to be normal.

Typically, both μ and σ^2 are unknown, and these quantities are estimated by the sample mean and sample variance, respectively.

2.4.1 Distribution of the sample mean (known variance)

Suppose we use the sample mean to estimate μ. What are its statistical properties?

The sample mean is

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

Clearly, $\bar{Y}$ is a linear combination of random variables. Hence,

$$\mathrm{E}(\bar{Y}) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(Y_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

and

$$\mathrm{Var}(\bar{Y}) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(Y_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}.$$

The distribution of $\bar{Y}$ will be exactly normal under the assumptions we are making or often approximately normal:

- The assumptions at the beginning of Section 2.4 included normality of $Y_1, \ldots, Y_n$. As a linear combination of normal random variables is normal (Example 1.32), $\bar{Y}$ is also normal. We can summarize by saying

  $$\bar{Y} \sim \mathsf{N}\left(\mu, \sigma^2/n\right),$$

  or equivalently if we standardize $\bar{Y}$ for its mean and variance,

  $$\frac{\bar{Y} - \mu}{\sqrt{\sigma^2/n}} \sim \mathsf{N}(0,1). \tag{2.1}$$

- Even if $Y_1, \ldots, Y_n$ are not normal, the CLT will often apply. The left-hand side of (2.1) is the same as the sample-mean version of the CLT in (2.10). Thus, if the conditions of the CLT hold, the left-hand side of (2.1) will converge in distribution to $\mathsf{N}(0,1)$ as $n \to \infty$ even if the Y_i are not normal.

If we know σ^2 we are ready for statistical inference (confidence intervals, hypothesis tests) for μ based on the normal distribution. Unfortunately, in practice, σ^2 is also usually unknown and we need to estimate it by the sample variance.

2.4.2 Distribution of the sample variance

The sample variance is

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2\right). \tag{2.2}$$

The equivalence of the two formulas for computing S^2 is seen by multiplying out the square in the first formula and collecting terms.

It is fairly straightforward to show that dividing by $n-1$ (and not n) makes S^2 an unbiased estimator of σ^2, i.e.,

$$\mathrm{E}(S^2) = \sigma^2.$$

This is proved as Exercise 2.9. The proof only requires that $Y_1, \ldots, Y_n$ are independent with mean μ and variance σ^2, i.e., it does not depend on having a normal distribution.

Finding the distribution of S^2 is more challenging in general but approachable when the Y_i are IID normal. In the first definition of S^2 in (2.2), write

$$Y_i - \bar{Y} = \sigma\left(\frac{Y_i - \mu}{\sigma} - \frac{\bar{Y} - \mu}{\sigma}\right) = \sigma(Z_i - \bar{Z}),$$

i.e., standardize Y_i to Z_i with a standard normal distribution. Then

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

or

$$\frac{S^2}{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \bar{Z})^2.$$

Thus, the distribution of S^2 is that of $\sum_{i=1}^{n}(Z_i - \bar{Z})^2$ up to constants.

The following theorem shows that the distribution of $\sum_{i=1}^{n}(Z_i - \bar{Z})^2$ is χ^2_{n-1}, i.e., the degrees of freedom are $n-1$ and not n. Each observation is "corrected" for the sample mean, i.e., $Y_i - \bar{Y}$ or $Z_i - \bar{Z}$. This is because the mean, μ, of the normal distribution is unknown and has to be estimated; the estimation of this one parameter leads to a correction of the degrees of freedom by 1.

Theorem 2.1 (χ^2 corrected degrees of freedom)
Let $Z_1, \ldots, Z_n$ be independent $\mathsf{N}(0,1)$ random variables. Then

$$\sum_{i=1}^{n}(Z_i - \bar{Z})^2 \sim \chi^2_{n-1},$$

i.e., χ^2 with $n-1$ degrees of freedom.

Proof. Write

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} (Z_i - \bar{Z})^2 + n\bar{Z}^2. \tag{2.3}$$

(This is just a restatement of the equivalent formulas in (2.2) with Z instead of Y.) Then use an argument based on MGFs.

1. On the right of (2.3), denote by $M(t)$ the MGF of $\sum_{i=1}^{n}(Z_i - \bar{Z})^2$. This sum is the random variable of interest in Theorem 2.1. By finding its MGF, we will identify its distribution.

2. In the second term on the right of (2.3), $\bar{Z} \sim \mathsf{N}(0, 1/n)$, so that $\sqrt{n}\bar{Z} \sim \mathsf{N}(0, 1)$. Hence, $n\bar{Z}^2 \sim \chi_1^2$ with MGF $(1-2t)^{-1/2}$ (see Exercise 2.3).

3. We also note that $Z_i - \bar{Z}$ and $\bar{Z}$ are independent (use the result of Exercise 2.8 in the special case of Z with $\mu = 0$ and $\sigma^2 = 1$ instead of Y). Thus, the two terms on the right of (2.3) are independent and by Lemma 1.3 their sum has MGF $M(t) \times (1-2t)^{-1/2}$.

4. On the left of (2.3), Lemma 2.1 says that $\sum_{i=1}^{n} Z_i^2$ has a χ_n^2 distribution with MGF $(1-2t)^{-n/2}$.

5. Hence, equating the MGFs of the left and right of (2.3),

$$(1-2t)^{-n/2} = M(t) \times (1-2t)^{-1/2},$$

and $M(t) = (1-2t)^{-(n-1)/2}$, which is the MGF of a χ_{n-1}^2 random variable.

The consequence of Theorem 2.1 is that S^2/σ^2 has the same distribution as that of a χ_{n-1}^2 random variable divided by $n-1$.

2.4.3 Distribution of the standardized sample mean (unknown variance)

At the end of Section 2.4.1 we looked ahead to making statistical inference about the mean μ based on the sample mean, $\bar{Y}$. We noted that in practice σ^2 will have to be estimated in the standardized sample mean,

$$\frac{\bar{Y} - \mu}{\sqrt{\sigma^2/n}},$$

in (2.1). The obvious strategy is to use the sample variance, S^2, as it is an unbiased estimator of σ^2. The standardized mean becomes

$$\frac{\bar{Y} - \mu}{\sqrt{S^2/n}}.$$

Replacing σ^2 by S^2 will change the standard normal distribution on the right of (2.1). It becomes a t_{n-1} distribution. To see this, we write

$$\frac{\bar{Y}-\mu}{\sqrt{S^2/n}} = \frac{(\bar{Y}-\mu)/\sqrt{\sigma^2/n}}{\sqrt{S^2/\sigma^2}}. \tag{2.4}$$

We then argue as follows about the numerator and denominator of this expression.

- The numerator, $(\bar{Y}-\mu)/\sqrt{\sigma^2/n}$ is back to the random variable in (2.1) and therefore has the same distribution as $Z \sim \mathsf{N}(0,1)$.
- The denominator is $\sqrt{S^2/\sigma^2}$. At the end of Section 2.4.2, we concluded that S^2/σ^2 has the same distribution as that of a χ^2_{n-1} random variable divided by $n-1$. Thus, the denominator has the same distribution as $\sqrt{X_{n-1}/(n-1)}$, where $X_{n-1} \sim \chi^2_{n-1}$.
- The random variables in the numerator and denominator are $\bar{Y}$ and S^2, respectively. Now, S^2 is a function of $Y_i - \bar{Y}$ $(i = 1, \ldots, n)$, and Exercise 2.8 shows that $\bar{Y}$ and $Y_i - \bar{Y}$ are independent. Therefore, $\bar{Y}$ and S^2 are independent, and the numerator and denominator of (2.4) are independent.

These properties of the numerator and denominator of 2.4 are those leading to the t distribution in Lemma 2.2. Thus,

$$\frac{\bar{Y}-\mu}{\sqrt{S^2/n}} \sim t_{n-1}. \tag{2.5}$$

In Lemma 2.2, the degrees of freedom are given by $d = n-1$, because we related S^2/σ^2 to χ^2_{n-1}. Thus, we will use the t_{n-1} distribution for inference about the mean of a normal distribution when the variance is unknown. This is the result derived by Student (1908), which was motivated by his analysis of experimental results at Guinness Breweries in Dublin.

Example 2.1 (Lung function: confidence interval for the normal mean) Schlaich et al. (1998) conducted a study on reduced pulmonary (i.e., lung) function in patients with spinal osteoporosis ("manifest osteoporosis"). Their objective was to compare pulmonary function between patients with this manifest osteoporosis and patients without the condition. For now we will consider only the manifest osteoporosis data here.

The measure of lung function was forced expiratory volume in 1 second (FEV1). The raw measure is adjusted for sex, age, and body height, leading to a percentage (FEV1%) relative to a standard, called y below. (The calculations in this example are based on data adjusted for current body height.) Data for $n = 34$ patients with manifest osteoporosis were collected, which we will treat as a random sample from a larger population of interest. The authors checked that the data are consistent

with arising from a normal distribution. They were interested in estimating the mean, μ, of the normal distribution when σ^2 is unknown.

The data summaries for the sample of size 34 are:

$$\bar{y} = 94.3 \quad \text{and} \quad s = 14.7,$$

where s is the sample standard deviation.

The estimate of μ is $\bar{y} = 94.3$ here. How much error could there be in this estimate just by chance due to random sampling? The analysis Schlaich et al. conducted was based on the exact t distribution in (2.5), i.e.,

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

Given $0 < \alpha < 1$, by definition the quantiles $t_{n-1,\alpha/2}$ and $t_{n-1,1-\alpha/2}$ cut off a probability of $\alpha/2$ in each tail of the t_{n-1} distribution, and

$$\Pr\left(t_{n-1,\alpha/2} < \frac{\bar{Y} - \mu}{S/\sqrt{n}} < t_{n-1,1-\alpha/2}\right) = 1 - \alpha.$$

Rearrangement gives

$$\Pr\left(\bar{Y} + t_{n-1,\alpha/2}\,\frac{S}{\sqrt{n}} < \mu < \bar{Y} + t_{n-1,1-\alpha/2}\,\frac{S}{\sqrt{n}}\right) = 1 - \alpha,$$

which is a probabilistic bound on μ. Figure 2.6 illustrates.

Note that the random variables here are the *estimators* $\tilde{\mu} = \bar{Y}$ and $\tilde{\sigma} = S$. When we replace them by the numerical *estimates* $\bar{y}$ and s from the data, we have a $100(1 - \alpha)\%$ confidence interval for μ:

$$\bar{y} + t_{n-1,\alpha/2}\,\frac{s}{\sqrt{n}} < \mu < \bar{y} + t_{n-1,1-\alpha/2}\,\frac{s}{\sqrt{n}}.$$

(Chapter 3 develops the distinction between estimators and estimates.) Because the t distribution is symmetric, $t_{n-1,\alpha/2} = -t_{n-1,1-\alpha/2}$.

With $n = 34$ as in the Schlaich et al. study, and $\alpha = 0.05$ for 95% confidence, `qt(0.975, df = 33)` in `R` gives $t_{n-1,1-\alpha/2} = t_{33,0.975} = 2.035$. A 95% confidence interval for μ is therefore

$$\bar{y} \pm t_{n-1,0.975}\frac{s}{\sqrt{n}} = 94.3 \pm 2.035\frac{14.7}{\sqrt{34}} = 94.3 \pm 5.1 = [89.2, 99.4]\%.$$

(The units of the interval, percentage points, are the same as for the FEV1% measurements.) ◇◇◇

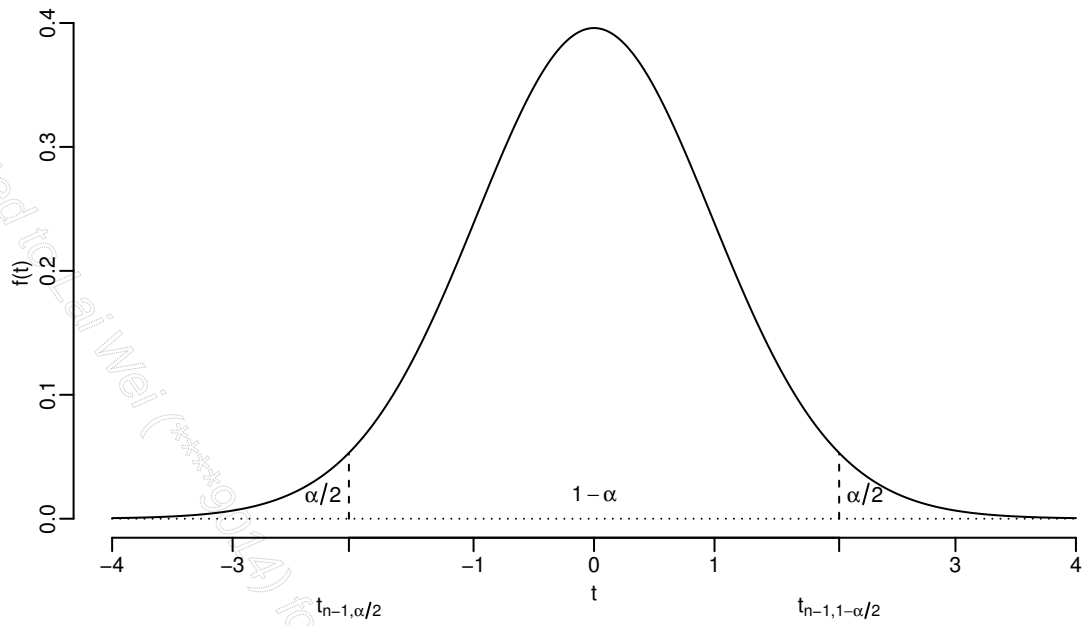


Figure 2.6: Quantiles of the t distribution with $n-1$ degrees of freedom. The quantiles $t_{n-1,\alpha/2} = -t_{n-1,1-\alpha/2}$ and $t_{n-1,1-\alpha/2}$ cut off the probability $\alpha/2$ in each tail. The PDF shown has $n-1 = 33$ degrees of freedom as in the Schlaich et al. study, and the quantiles shown are for $\alpha = 0.05$, leading to a 2-sided 95% confidence interval.

2.5 Limiting Normal Distributions

2.5.1 Convergence in distribution

Definition 2.1 (Convergence in distribution)
Let $X_1, X_2, \ldots$ be a sequence of random variables with CDFs $F_{X_1}(x)$, $F_{X_2}(x), \ldots$, respectively. Suppose there exists a CDF $F_X(x)$ such that

$$\lim_{n\to\infty} \Pr(X_n \le x) = F_X(x),$$

for all x where $F_X(x)$ is continuous. Then we say that the sequence of random variables X_n converges in distribution to $F_X(x)$.

The random variables $X_1, X_2, \ldots$ here are *not* the individual elements of a sample. Rather, X_n will be based on a summary like the sample mean of the entire sample. In the next example, we work with a standardized version of the sample total. The example also demonstrates that the limiting distribution may be found from the limiting MGF.

Example 2.2 (Binomial distribution: normal approximation)
Let $X_n \sim \mathsf{Bin}(n, \pi)$ with PMF

$$f_{X_n}(x) = \binom{n}{x} \pi^x (1-\pi)^{n-x} \quad (x = 0, 1, \ldots, n).$$

We will show that a *standardized version* of X_n has a distribution that converges to the standard normal distribution as $n \to \infty$.

We want the limiting distribution of X_n as $n \to \infty$. But the limit cannot possibly be a fixed distribution. We know that $\mathrm{E}(X_n) = n\pi$ and $\mathrm{Var}(X_n) = n\pi(1-\pi)$; both are changing (increasing) with n. But

$$Z_n = \frac{X_n - \mathrm{E}(X_n)}{\sqrt{\mathrm{Var}(X_n)}} = \frac{X_n - n\pi}{\sqrt{n\pi(1-\pi)}} \tag{2.6}$$

has mean 0 and variance 1. It is the distribution of this standardized quantity that converges to a fixed distribution, namely the standard normal.

The key steps are: (1) to find the MGF of Z_n in (2.6); and (2) to show that the MGF tends to that of the standard normal as $n \to \infty$.

Derivation of the MGF of Z_n is based on Lemma 1.2 for the MGF of a linear function of a random variable. Write Z_n in (2.6) as $Z_n = a_n + b_n X_n$, where

$$a_n = -\frac{\pi\sqrt{n}}{c} \quad \text{and} \quad b_n = \frac{1}{c\sqrt{n}},$$

and $c = \sqrt{\pi(1-\pi)}$ to simplify notation. From Table 1.3 the MGF of the binomial random variable X_n is

$$M_{X_n}(t) = (1 - \pi + \pi e^t)^n.$$

Applying Lemma 1.2 gives

$$M_{Z_n}(t) = e^{a_n t} M_{X_n}(b_n t) = e^{-\frac{\pi\sqrt{n}}{c}t}\left(1 - \pi + \pi e^{\frac{1}{c\sqrt{n}}t}\right)^n,$$

which can be rearranged as

$$\left(e^{-\frac{\pi}{c\sqrt{n}}t}\left(1 - \pi + \pi e^{\frac{1}{c\sqrt{n}}t}\right)\right)^n = \left((1-\pi)e^{-\frac{\pi}{c\sqrt{n}}t} + \pi e^{\frac{(1-\pi)}{c\sqrt{n}}t}\right)^n.$$

This completes the first step.

To find the limiting MGF of Z_n as $n \to \infty$, first make Taylor series expansions of the exponential functions in $M_{Z_n}(t)$:

$$\begin{aligned} M_{Z_n}(t) &= \left((1-\pi)\left(1 - \frac{\pi}{c\sqrt{n}}t + \frac{\pi^2}{2c^2 n}t^2 - O\left(\frac{1}{n^{3/2}}\right)\right)\right. \\ &+ \left.\pi\left(1 + \frac{1-\pi}{c\sqrt{n}}t + \frac{(1-\pi)^2}{2c^2 n}t^2 + O\left(\frac{1}{n^{3/2}}\right)\right)\right)^n. \end{aligned}$$

The MGF here is a function of t and n, but for any fixed value of t, we are interested in the behaviour as $n \to \infty$. The notation $O(1/n^{3/2})$ represents further terms that are decreasing at rate $1/n^{3/2}$ or faster. Technically, if $m(t,n)$ is the sum of all terms after the t^2 term, then $m(t,n)$ is $O(1/n^{3/2})$ says that $|m(t,n)| < a(t)/n^{3/2}$ for some positive constant $a(t)$ and sufficiently large n. Thus, $m(t,n)$ gives a

negligible contribution for large n, providing the leading terms in $1/\sqrt{n}$ and $1/n$ do not both cancel. Collecting terms of order t^0, t^1, and t^2 gives

$$M_{Z_n}(t) = \left(1 + \frac{(1-\pi)\pi^2 + \pi(1-\pi)^2}{2c^2 n} t^2 + O\left(\frac{1}{n^{3/2}}\right)\right)^n,$$

which further simplifies to

$$M_{Z_n}(t) = \left(1 + \frac{1}{n}\frac{t^2}{2} + O\left(\frac{1}{n^{3/2}}\right)\right)^n,$$

recalling that $c = \sqrt{\pi(1-\pi)}$. As $n \to \infty$, the $O(1/n^{3/2})$ term can be ignored, and $(1 + (t^2/2)/n)^n \to e^{t^2/2}$ (recall that $(1 + x/n)^n \to e^x$ as $n \to \infty$). Thus,

$$M_{Z_n}(t) \to e^{t^2/2} \quad \text{as } n \to \infty,$$

i.e., the limiting MGF is that of the standard normal (see Table 1.4). Theorem 1.3 says that the MGF uniquely identifies a distribution, and hence the distribution of Z_n converges to the standard normal distribution as $n \to \infty$. ◊◊◊

2.5.2 Limiting distributions and large-sample approximations in statistics

Results on limiting distributions like the result in Example 2.2 show convergence in distribution for a standardized version of a random variable. The sample size, n, becomes infinitely large. In statistical practice, however, sample sizes are finite. Furthermore, to a statistician the statistic of interest is often an unstandardized random variable. Limiting distributions of standardized random variables justify the use of approximate distributions for finite samples and unstandardized quantities.

For instance, let X_n be a sample from a $\mathsf{Bin}(n, \pi)$ distribution. The sample proportion, X_n/n, is often be used to estimate π. This is not the standardized random variable Z_n in Example 2.2. We can write X_n/n in terms of Z_n, however, by rearranging (2.6):

$$X_n = n\pi + \sqrt{n\pi(1-\pi)} Z_n,$$

and hence

$$\frac{X_n}{n} = \pi + \sqrt{\frac{\pi(1-\pi)}{n}} Z_n.$$

For a *finite* sample, Z_n has an *approximately* standard normal distribution, and the theory says that the approximation will become better as the sample size increases. Furthermore, X_n/n is a linear function of Z_n, and using the result that a linear function of a normal random variable also has a normal distribution (Example 1.30), X_n/n has the following approximate distribution:

$$\frac{X_n}{n} \sim \mathsf{N}\left(\pi, \frac{\pi(1-\pi)}{n}\right). \tag{2.7}$$

This is the argument for quantifying how accurate X_n/n is as an estimator of π in a statistical sense.

Example 2.3 (Opinion polls: margin of error (confidence interval)) Opinion polls are typically conducted with sample sizes like $n = 1000$ or $n = 3000$ from a large population such as all adult Canadians. Often, the results will be qualified with a statement like, "These results have a margin of error of 3 percentage points 19 times out of 20." Let's see how such a statement can be justified.

For definiteness, take the Nanus poll taken in September 2016 commissioned by Clean Energy Canada (available at `http://cleanenergycanada.org/wp-content/uploads/2016/09/Clean-Energy-Canada-Nanos-Climate-Policy-Polling-Report-Oct-2016.pdf`). It asked 1000 randomly selected Canadian adults a number of questions about climate change. The findings included:

- 48% agreed with the statement, "A changing climate presents a significant threat to our economic future" (and a further 23% somewhat agreed).
- 33% supported, "Having a price on carbon to reduce the use of fossil fuels such as coal, oil or natural gas" (and a further 26% somewhat supported the statement).

In the methodology section of the report, we find, "The margin of error is ±3.1 percentage points 19 times out of 20." The margin of error is a measure of sampling variability. Thus, to check the calculation, we treat the sample as a random sample from a large population. Then, the number of people agreeing with a particular statement (e.g., the question about significant threat to our economic future) is a binomial random variable with $n = 1000$ (the sample size here) and probability π. The parameter π represents the unknown proportion of people in the population who agree with this particular statement. The Nanos poll was stratified by age, gender, and region and not a simple random sample, which implies our binomial model probability is oversimplified. Nonetheless, the impact on the margin of error calculation is usually small.

The approximate distribution of X_n/n in (2.7) implies that approximately

$$\frac{X_n}{n} - \pi \sim \mathsf{N}\left(0, \frac{\pi(1-\pi)}{n}\right).$$

On the left is the *error* in estimating π by X_n/n. The normal distribution has the property that 95% of the probability ("19 times out of 20") lies within ±1.96 standard deviations of the mean. Thus, the error is within

$$\pm 1.96\sqrt{\frac{\pi(1-\pi)}{n}} \tag{2.8}$$

approximately 19 times out of 20.

There are two further practical difficulties in completing the computation.

- What value should be used for π in the error calculation? Recall that the purpose of the poll is to estimate π, which is unknown.
- Most polls, like the Nanos poll, ask several questions. Each question could have a different true value of π. It would be overly complicated to give a different error calculation for every question.

To overcome both of these difficulties, pollsters will often report (2.8) for $\pi = 0.5$. This is the value of π that maximizes the variance in the approximate normal distribution of the error and hence gives the widest, worst-case bounds on the margin of error.

Returning to the Nanos poll, with $n = 1000$ the worst-case margin of error from (2.8) is

$$\pm 1.96\sqrt{\frac{0.5(1-0.5)}{1000}} = \pm 1.96(0.0158) = \pm 0.031,$$

i.e., 3.1%, which is the margin of error reported. ♢♢♢

Error bounds like those in Example 2.3 are also called *confidence intervals* and are tackled more generally in Section 4.6.

2.5.3 Central limit theorem

The normal approximation to the binomial distribution is just a special case of the central limit theorem (CLT).

Theorem 2.2 (Central limit theorem (CLT))
Let $Y_1, Y_2, \ldots$ be a sequence of IID random variables with mean μ, variance σ^2, and MGF defined in a neighbourhood of zero. Define the sum

$$X_n = \sum_{i=1}^{n} Y_i.$$

The standardized sum,

$$Z_n = \frac{X_n - n\mu}{\sigma\sqrt{n}}, \tag{2.9}$$

converges in distribution to $\mathsf{N}(0,1)$.

Referring back to Definition 2.1 for convergence in distribution, the CLT says that the CDF of Z_n approaches that of the standard normal as $n \to \infty$, i.e., $\Pr(Z_n < z) \to \Pr(Z < z)$, where $Z \sim \mathsf{N}(0,1)$. This is sufficient for statistical purposes: in Example 2.3, for instance, the limits for the 95% confidence interval use $z_{0.975} = 1.96$, a quantile of the standard normal defined by the CDF. Note, however, that the speed of convergence will be slower in the extreme tails of the Z_n distribution.

We can also express the CLT in terms of the arithmetic mean, $\bar{Y}_n = X_n/n$. Dividing the top and bottom of (2.9) by n, we have

$$Z_n = \frac{\bar{Y}_n - \mu}{\sigma/\sqrt{n}} \tag{2.10}$$

converges in distribution to $\mathsf{N}(0,1)$. We can use either form depending on whether its easier to work with a sum or mean of random variables.

The proof of the CLT, given in Section 2.10, is based on a generalization of the strategy in Example 2.2 for the binomial distribution.

Example 2.4 (Binomial distribution: normal approximation via CLT) In Example 2.2 the normal approximation to the binomial distribution was developed by working directly with the binomial MGF. The CLT, a more general result, gives the same limiting normal approximation almost immediately.

Let $B_1, \ldots, B_n$ be n independent $\mathsf{Bern}(\pi)$ random variables. Then

$$X_n = \sum_{i=1}^{n} B_i$$

has the distribution $X_n \sim \mathsf{Bin}(n, \pi)$ (Exercise 1.19). We have expressed the binomial random variable as a sum, so the sum version of the CLT in (2.9) is more convenient. As $\mathrm{E}(B_i) = \pi$ and $\mathrm{Var}(B_i) = \pi(1-\pi)$, the CLT says that

$$\frac{X_n - n\pi}{\sqrt{n\pi(1-\pi)}}$$

converges in distribution to the standard normal. This is the result obtained in Example 2.2. ♢♢♢

2.6 Getting It Done in R

2.6.1 Sample mean, standard deviation, and variance

The functions `mean`, `sd`, and `var` are useful for obtaining the sample mean, standard deviation, and variance, respectively.

2.6.2 Quantiles of the t distribution

R has functions to compute quantiles for commonly used distributions. Corresponding to the functions in Table 1.9 with names starting with `d` for PMFs and PDFs, names starting with `q` compute quantiles.

For instance, to compute a confidence interval in the lung-function study of Example 2.1, quantiles of the t distribution are computed by `qt`.

```
> qt(p = 0.025, df = 33)
[1] -2.034515
> qt(p = 0.975, df = 33)
[1] 2.034515
```

Here, `p` is the probability to the left of the quantile, and there are 33 degrees of freedom because $n = 34$ in the particular study.

Whereas `pt(y, ...)` returns a CDF probability for a given value or quantile of a random variable with a t distribution, the function `qt(p, ...)` returns a quantile for a given probability. Thus, the quantile function is the inverse cumulative distribution function. In general, for given probability p, the quantile function returns y such that $\Pr(Y \leq y) = F_Y(y) = p$, or equivalently $y = F^{-1}(p)$, where F^{-1} is the inverse function of the CDF.

2.6.3 Limiting normal distributions

Similarly, in Example 2.3 we needed quantiles of the standard normal distribution, which are available from `qnorm`.

```
> qnorm(p = 0.025)
[1] -1.959964
> qnorm(p = 0.975)
[1] 1.959964
```

Quantiles of the normal distribution with non-standard mean and variance are obtained via the arguments `mean` and `sd` of `qnorm`.

2.7 Learning Outcomes

On completion of this chapter you should be able to demonstrate the following skills.

1. **Normal distribution**

   (a) Use the MGF to derive the mean, variance, and distribution of a linear function of a normal random variable. (You can interpret "derive" here as justify via an appropriate result.)

   (b) Derive the mean, variance, and distribution of a linear combination of normal random variable.

   (c) Standardize a normal random variable to have a standard normal distribution.

2. χ^2 **distribution**

(a) Derive the MGF of the χ^2_1 distribution.

(b) Show via the MGF that the square of a standard normal random variable has a χ^2_1 distribution.

(c) Find the MGF of the χ^2_d distribution.

(d) Show that the sum of squares of d independent standard normal random variables has a χ^2_d distribution.

3. **Random sample from a normal distribution**
Let $Y_1, \ldots, Y_n$ be a random sample of independent $\mathsf{N}(\mu, \sigma^2)$ random variables.

(a) Derive the mean and variance of the sample mean, $\bar{Y}$.

(b) Write down unstandardized and standardized distributions of $\bar{Y}$ when σ^2 is known.

(c) Write down the distribution of the sample variance, S^2.

(d) Show that the sample mean and sample variance are statistically independent.

(e) Write down a standardized version of $\bar{Y}$ when σ^2 is unknown. Argue that the standardized random variable has the properties leading to a t distribution with specified degrees of freedom.

(f) Use the t distribution to compute a confidence interval for μ when σ^2 is unknown for given data.

4. **Central Limit Theorem**

(a) Apply the CLT to establish convergence in distribution to the standard normal for a random variable with a specified distribution (e.g., binomial). Included here is the formulation of the random variable as a sample mean or sample sum, standardizing it, and checking the conditions of the CLT. A detailed proof of the theorem is not expected.

(b) Under a binomial probability model, convert the standard normal distribution of a standardized version of the sample proportion as $n \to \infty$ to an approximate normal distribution of the unstandardized sample proportion and a finite sample size.

(c) Compute the "margin of error" in estimating the parameter π in the binomial distribution for given data.

5. **Explanation**
Explain your reasoning by describing the results you are using, along with any assumptions that are necessary.

2.8 Exercises

Exercise 2.1
Let $Y \sim \mathsf{N}(\mu, \sigma^2)$. Show that the standardized random variable $Z = (Y - \mu)/\sigma$ has a $\mathsf{N}(0, 1)$ distribution.

Exercise 2.2
Let $Y_1, \ldots, Y_n$ be independent $\mathsf{N}(\mu, \sigma^2)$ random variables.

1. Write down the MGF of Y_i.
2. Derive the MGF of $Y_1 + \cdots + Y_n$.
3. Hence, derive the MGF of $\bar{Y} = (Y_1 + \cdots + Y_n)/n$.
4. Hence, derive the MGF of $Z = (\bar{Y} - \mu)/(\sigma/\sqrt{n})$ and identify its distribution.

In parts 2–4, you may use without proof general properties in Chapter 1 on MGFs. When you use a property, however, remember to state it, make clear how it is being applied, and check any assumptions required for the result.

Exercise 2.3
Let $Y \sim \chi^2_1$. Show that the MGF of Y is $(1 - 2t)^{-1/2}$.

Exercise 2.4
[Quiz #1, 2010-11, Term 2, except that showing "$\text{Var}(Z^2) = 2$" was not included] Let Z have a standard normal distribution with expectation 0 and variance 1, i.e., $Z \sim \mathsf{N}(0, 1)$.

1. Write down the *definition* of $M_{Z^2}(t)$, the moment generating function (MGF) of Z^2.
2. From the definition, show $M_{Z^2}(t) = (1 - 2t)^{-1/2}$.
3. Use $M_{Z^2}(t)$ to show that $\text{E}(Z^2) = 1$ and $\text{Var}(Z^2) = 2$.
4. Argue that the distribution of Z^2 is χ^2_1.

Exercise 2.5
Show that the MGF of a random variable with a χ^2_d distribution is $(1 - 2t)^{-d/2}$.

Exercise 2.6
[Quiz #1, 2011-12, Term 1, except that part 5 was not included] Let $Z_1, \ldots, Z_d$ be d independent $\mathsf{N}(0, 1)$ random variables, and let

$$Y = Z_1^2 + \cdots + Z_d^2.$$

You may use without proof the result that $Z_i^2 \sim \chi^2_1$ for $i = 1, \ldots, d$, and that the moment generating function (MGF) of the χ^2_1 distribution is $(1 - 2t)^{-1/2}$ (see Exercise 2.4).

1. Write down the *definition* of $M_Y(t)$, the MGF of Y.
2. Show that the MGF of Y is $(1-2t)^{-d/2}$, either directly starting from the definition or by stating and applying an appropriate result. In either case be sure to explain the assumption(s) you are using.
3. Hence, what is the distribution of Y? Explain briefly.
4. From the MGF find $\mathrm{E}(Y)$.
5. From the MGF find $\mathrm{Var}(Y)$.

Exercise 2.7

Let X_1 and X_2 have independent χ^2 distributions with degrees of freedom d_1 and d_2, respectively. Show that $X_1 + X_2$ has a $\chi^2_{d_1+d_2}$ distribution.

Exercise 2.8

Let $Y_1, \ldots, Y_n$ be independent random variables with mean μ and variance σ^2, and let $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$.

1. Show that the covariance between $\bar{Y}$ and $Y_i - \bar{Y}$ is zero.
2. Assume also that $Y_1, \ldots, Y_n$ are normally distributed. Are $\bar{Y}$ and $Y_i - \bar{Y}$ independent? Why?

Exercise 2.9

Let $Y_1, \ldots, Y_n$ be independent random variables, each with mean μ and variance σ^2. The values of μ and σ^2 are both unknown. Consider the sum of squares

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

1. Show that $\mathrm{E}(X) = (n-1)\sigma^2$.
2. Hence give an estimator of σ^2 based on X that has expectation σ^2.

Exercise 2.10

In this exercise we calculate further confidence intervals for the study in Example 2.1. Recall that Schlaich et al. (1998) collected data on adjusted forced expiratory volume for $n = 34$ patients with manifest osteoporosis. The data summaries for the sample are:

$$\bar{y} = 94.3 \quad \text{and} \quad s = 14.7,$$

where the units are percentage points.

As before, we will assume that the $n = 34$ data values are a random sample from a $\mathsf{N}(\mu, \sigma^2)$ distribution, and that the objective is to estimate μ.

1. Compute 90% and 99% confidence intervals for μ based on Student's t distribution. What are their widths?

2. Before Student derived the t distribution, common practice was to carry out calculations as above using the standard normal distribution rather than the t distribution.

   (a) Use `R` to plot the PDF of the standard normal for values in the range $[-4, 4]$. You can set up such a grid of values at spacing of 0.01 using
   `x <- seq(-4, 4, by = 0.01)`
   When you use `plot` with `x` on the x-axis and the corresponding values of the normal PDF on the y-axis, include the argument `type = "l"` to tell `R` to join the coordinates as lines to create a curve.

   (b) Add the PDF of the t distribution (with appropriate degrees of freedom) to your plot. Using `lines` rather than `plot` will add to the current plot rather than generating a new one. To distinguish the two curves, the argument `lty = 2` will make the new curve from dashed lines.

   (c) Comment on how well the standard normal approximates the t distribution here.

3. Recompute the 90% and 99% confidence intervals in part 1 but use the standard normal rather than the t distribution. What are their widths?

4. How much wider are the confidence intervals using the t distribution relative to those using the standard normal? ("Relative" here is a ratio of widths.)

5. Look again at your plots of the standard normal and t PDFs. Why is there more more discrepancy in the confidence interval from the t distribution relative to the normal distribution as the confidence level increases?

Exercise 2.11

Example 2.1 analyzed data collected by Schlaich et al. (1998) on lung function in patients with manifest osteoporosis. The investigators also collected data on a second sample of $n = 51$ patients without manifest osteoporosis. The second sample is a "control" group for comparison. The definition of the measure FEV1% we will use for the control group is the same as in Example 2.1 and is again called y.

The control sample gives the following data summaries:

$$\bar{y} = 96.1 \quad \text{and} \quad s = 14.4,$$

where s is the sample standard deviation. We will again assume the data are a random sample from a normal distribution and that interest centres on estimation of the mean of the distribution.

1. Based on the above description, write down a formal probability model for the way that the control-sample data, $y_1, \ldots, y_{51}$, arose. Be sure to specify:

   (a) the random variable(s);
   (b) the distribution of the random variable(s);
   (c) a description of any parameters of the distribution;
   (d) any other assumption(s) about the random variable(s).

2. Assuming the probability model holds, calculate:
   (a) an estimate of the mean of the distribution;
   (b) an estimate of the standard deviation of the sample mean over repeated samples;
   (c) a 95% confidence interval for the mean of the assumed distribution.

3. Again assuming the probability model is correct, are there any approximations in the confidence-interval calculation? Briefly explain why or why not.

4. Compare the confidence intervals in Example 2.1 and in this exercise, and comment briefly.

Exercise 2.12

[Parts 1–5 appeared on Quiz #1, 2010-11, Term 2] Suppose a random sample of size $n = 2$ is drawn from a $\mathsf{N}(\mu, \sigma^2)$ distribution to estimate μ when σ^2 is unknown. There is a big impact on the 95% confidence interval for μ from using the t distribution instead of the standard normal.

```
> qnorm(0.975)
[1] 1.959964
> qt(0.975, df = 1)
[1] 12.7062
```

Thus, a confidence interval for μ based on the t distribution will be much, much wider. Why? And why does the t distribution have 1 degree of freedom here (and not 2 from $n = 2$)? The exercise sheds some light on these questions.

Let Y_1 and Y_2 be independent random variables sampled from a $\mathsf{N}(\mu, \sigma^2)$ distribution. For such a sample of size $n = 2$ it is easily shown that the sample variance,

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

simplifies to

$$S^2 = \left(\frac{Y_1 - Y_2}{\sqrt{2}}\right)^2.$$

You may use this result without proof.

1. Let $V = (Y_1 - Y_2)/\sqrt{2}$. Show the following properties of V. Carefully state any result you are using (no proof of the result required) and how it is applied here.

(a) $\mathrm{E}(V) = 0$.

(b) $\mathrm{Var}(V) = \sigma^2$.

(c) V has a normal distribution.

2. Explain why the distribution of V/σ is standard normal.

3. Hence argue that when $n = 2$, the distribution of S^2/σ^2 is χ^2_1. (A result on the connection between $\mathsf{N}(0,1)$ and χ^2_1 random variables may be stated and used without proof.)

4. Using (without proof) the properties of the χ^2_1 distribution, what are the expectation and variance of S^2/σ^2 when $n = 2$?

5. Hence, what is the expectation of S^2 when $n = 2$?

6. What is the variance of S^2 when $n = 2$?

7. Use `qchisq` in `R` to find quantiles l and u such that

$$\Pr(S^2/\sigma^2 < l) = 0.025 \quad \text{and} \quad \Pr(S^2/\sigma^2 > u) = 0.025.$$

Hence, l and u are lower and upper bounds on S^2/σ^2 in the sense that

$$\Pr(l < S^2/\sigma^2 < u) = 0.95.$$

8. Suppose S^2 is used to estimate σ^2 from a sample of size $n = 2$. Comment on the values of S^2/σ^2 that could occur.

Exercise 2.13

[Parts 1–3 appeared on the final exam, 2010-11, Term 1.] Let $Y_1, \ldots, Y_n$ be independent $\mathsf{N}(\mu, \sigma^2)$ random variables. Their sample variance is

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

This question explores the properties of S^2/σ^2 and why the t_{n-1} distribution approaches the standard normal as $n \to \infty$. Section 2.4.2 argued that S^2/σ^2 has the same distribution as $X/(n-1)$, where $X \sim \chi^2_{n-1}$. You may use this result and properties of the χ^2 distribution without proof.

1. Find $\mathrm{E}(S^2/\sigma^2)$.

2. Find $\mathrm{Var}(S^2/\sigma^2)$.

3. Let $\epsilon > 0$ be a fixed constant representing an arbitrarily small "error".

(a) As $n \to \infty$, what is the limiting probability

$$\Pr(1 - \epsilon < S^2/\sigma^2 < 1 + \epsilon)?$$

(b) Briefly describe how you would justify the limiting probability (a complete proof is not required).

4. In Section 2.4.3 the sample mean was standardized by its expectation and *sample* variance and then expanded in equation (2.4):

$$\frac{\bar{Y} - \mu}{\sqrt{S^2/n}} = \frac{(\bar{Y} - \mu)/\sqrt{\sigma^2/n}}{\sqrt{S^2/\sigma^2}}.$$

It was shown that this quantity has a t_{n-1} distribution. As $n \to \infty$, it is known that the t_{n-1} distribution converges to $\mathsf{N}(0, 1)$. Use the above results to justify this convergence.

5. Using the `R` function `rnorm`, simulate 1000 samples of size $n = 10$ from the normal distribution with $\mu = 0$ and $\sigma = 2$. For each sample, compute its sample variance using `var`.

6. Construct a histogram of the 1000 sample variances. Does it have a shape that looks like one of the distributions in Figure 2.2? If so, which?

7. Compute the sample mean and sample variance of the 1000 sample variances using `mean` and `var`. Compare to the theoretical mean and variance of the sample variance you found in parts 1 and 2.

Exercise 2.14

Let X have a χ^2_d distribution. Show that a standardized version of X has a limiting standard normal distribution as $d \to \infty$. Be sure to be specific about the standardization of X and to check the conditions of any result on limiting distribution that you use.

Exercise 2.15

This exercise demonstrates the CLT via simulation.

1. In `R`, generate a sample of 1000 independent $\mathsf{Unif}(-1, 1)$ random variables.

```
n <- 1000
x <- runif(n, min = -1, max = 1)
```

Take a look at the first 10 elements of the vector `x` that contains the sample using `x[1:10]`, and make sure they look good. For example, all values should be in $[-1, 1]$!

2. Use `hist` to draw a histogram of all the data in `x`. Look at `help(hist)` to find out how to do this.

3. Compute the sample mean and sample variance of the data. Compare with the theoretical mean and variance of the $\mathsf{Unif}\,(-1, 1)$ distribution. Why do the sample and theoretical quantities not agree exactly?

4. Generate a second, independent sample

```
y <- runif(n, min = -1, max = 1)
```

and then compute the sums `z[1] = x[1] + y[1], z[2] = x[2] + y[2]`, etc. using

```
z <- x + y
```

Note that `R` will apply the sum operator element-wise to the vectors. Take a look at the first few elements of `x, y, z` to make sure the summation has worked correctly. Thus, `z` contains 1000 values, where each element is generated as the *sum of a sample of two independent* $\mathsf{Unif}\,(-1, 1)$ *random variables.*

5. Draw a histogram of the sample data in `z`. Does the histogram look more normal than a single sample from the uniform distribution?

6. Repeat Steps 4–5 to generate a total of 5 independent samples, and compute `z` as the sum across all 5 vectors. Does the histogram of `z` have a shape that looks fairly normal?

7. Why do the `z` values not appear to be from a *standard* normal distribution? What would we have to do to the `z` values to standardize them?

Exercise 2.16

Throughout this question, whenever you use a general result, make sure you state it clearly and check its conditions (if any).

1. Let $U \sim \mathsf{Unif}\,(-1, 1)$, i.e., U has a uniform distribution with parameters $a = -1$ and $b = 1$ in Table 1.4.

   (a) From the definition of expectation, find $\mathrm{E}(U)$.

   (b) From the definition of variance, find $\mathrm{Var}(U)$.

   (c) From the definition of the moment generating function, find $M_U(t)$.

   (d) From the moment generating function, find $\mathrm{E}(U)$ and $\mathrm{Var}(U)$. (As in Example 1.29, you may find it easier to expand the exponential functions, collect leading terms, then differentiate.)

2. Let $U_1, \ldots, U_n$ be independent $\mathsf{Unif}\,(-1, 1)$ random variables. Consider the random variable

$$Y = \sum_{i=1}^{n} U_i.$$

(a) What is $\mathrm{E}(Y)$?

(b) What is $\mathrm{Var}(Y)$?

(c) What is the MGF of Y?

3. Now standardize Y to find a new random variable, Z, with mean 0 and variance 1.

(a) What is Z?

(b) What is the MGF of Z?

(c) What is the MGF of Z as $n \to \infty$? One approach is to expand the exponential functions and collect terms before taking the limit. Note also that $(1 + a/n)^n \to e^a$ as $n \to \infty$.

(d) What is the distribution of Z?

(e) You have just proved a special case of a more general theorem. What is it?

Exercise 2.17

Example 2.2 worked through the normal approximation to X_n, a $\mathsf{Bin}\,(n, \pi)$ random variable. It was shown that

$$Z = \frac{X_n - n\pi}{\sqrt{n\pi(1-\pi)}}$$

is approximately $\mathsf{N}\,(0, 1)$ for large n. Here, X_n is a discrete random variable, whereas Z is continuous. How can a random variable with a discrete PMF be approximated by another with a continuous PDF?

Exercise 2.18

[This exercise appeared on Quiz #1, 2011-12, Term 1 without Parts 3 and 5. The quiz included the fact that the `R` function `qnorm(0.975)` returns 1.959964.] Let $Y_1, \ldots, Y_n$ be independent random variables, each with mean μ and variance σ^2. Note that we are not necessarily assuming any distribution for the Y_i yet.

Consider using $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ to estimate μ.

1. Show that $\mathrm{E}(\bar{Y}) = \mu$.

2. Show that $\mathrm{Var}(\bar{Y}) = \sigma^2/n$.

3. What does Chebyshev's inequality give for the probability $\Pr(|\bar{Y} - \mu| > \epsilon)$, where $\epsilon = 1.96\sqrt{\sigma^2/n}$?

4. What does the CLT say about the distribution of $\bar{Y}$ as $n \to \infty$?

5. Give an approximation based on the CLT to $\Pr(|\bar{Y} - \mu| > \epsilon)$, where $\epsilon = 1.96\sqrt{\sigma^2/n}$. Explain briefly.

6. Suppose $Y_1, \ldots, Y_n$ also have a normal distribution.

(a) What is the distribution of $\bar{Y}$? Explain briefly.

(b) What is $\Pr(|\bar{Y} - \mu| > \epsilon)$, where $\epsilon = 1.96\sqrt{\sigma^2/n}$? Explain briefly.

(c) Is the calculated probability a large-sample approximation? Explain briefly.

Exercise 2.19

Example 2.2 argued that a standardized version of a binomial random variable has a limiting standard normal distribution as $n \to \infty$. Outline the key steps in the argument, pointing to results that would be used, without tedious algebraic detail. For instance, you might start with

> Step 1. Let $X_n \sim \mathsf{Bin}(n, \pi)$. Its MGF, $M_{X_n}(t)$, can be obtained from Table 1.3.

In other words, there is no need to derive $M_{X_n}(t)$ for this first step. Indeed, there is no need even to give an explicit expression for $M_{X_n}(t)$ as you will not be manipulating it algebraically in subsequent steps. On the other hand, you will need to define carefully and mathematically various terms like $M_{X_n}(t)$ as you go along, just to make your argument clear.

Exercise 2.20

This exercise explores the shape of the Poisson distribution, via simulation and via a limiting-distribution argument.

1. Using `rpois` in `R`, generate a random sample of 1000 values from a $\mathsf{Pois}(\mu = 0.35)$ distribution and plot the values using `hist`. Does the empirical distribution have a roughly normal shape?

2. Repeat part 1 but sample from a $\mathsf{Pois}(\mu = 25)$ distribution.

3. What do the two simulations suggest about the condition(s) for the normal distribution to be a good approximation to the Poisson distribution.

4. Let $Y \sim \mathsf{Pois}(\mu)$. The standardized variable $Z = (Y - \mu)/\sqrt{\mu}$ has mean 0 and variance 1. The MGF of Z can be written as

$$M_Z(t) = \exp\left(\frac{t^2}{2} + O\left(\frac{1}{\sqrt{\mu}}\right)\right).$$

   The notation $O\left(1/\sqrt{\mu}\right)$ says that, for any t, the sum of all terms after the t^2 term becomes negligible for sufficiently large μ. Also, see the last part of this question for the derivation of the expansion here.

   What is the limiting distribution of Z as $\mu \to \infty$?

5. In Worksheet 1.3 we ended up with a random variable Y with a $\mathsf{Pois}\,(\mu)$ distribution (approximately). Based on the data in Worksheet 3.1, μ appears to be fairly small, of the order $\mu \simeq 0.35$. Comment on whether the result in part 4 of this exercise justifies further approximating the distribution of Y by a normal distribution.

6. In Worksheet 3.1 we assume the data are a realization of $Y_1, \ldots, Y_n$, where $n = 74$ and the Y_i are IID $\mathsf{Pois}\,(\mu)$. Exercise 1.20 showed that $X_n = \sum_{i=1}^{n} Y_i$ has a $\mathsf{Pois}\,(n\mu)$ distribution. As $n = 74$, we clearly have $n\mu \gg \mu$. Comment on whether a normal distribution might be a good approximation to the distribution of X_n and hence the sample mean, $\bar{Y} = X/n$.

7. Derive the expansion in part 4.

   (a) Write down the MGF of Y.

   (b) The standardized variable $Z = (Y - \mu)/\sqrt{\mu}$ has mean 0 and variance 1. What is the MGF of Z?

   (c) Using a series expansion of the exp function, show that the exponent $\mu\left(\exp\left(t/\sqrt{\mu}\right) - 1\right)$ appearing in $M_Z(t)$ can be written as

$$\sqrt{\mu}t + \frac{t^2}{2} + O\left(\frac{1}{\sqrt{\mu}}\right).$$

   (d) Hence collect terms and obtain the expansion of $M_Z(t)$.

2.9 Appendix: Proof of Lemma 2.2

From Section 2.3.2 we already established that the PDF of $T = Z/\sqrt{X_d/d}$ in Lemma 2.2 is given by

$$f_T(t) = \int_0^\infty f_{T|W}(t \mid w) f_W(w)\, dw,$$

where $W = X_d/d$. It was also argued that $f_{T|W}(t \mid w)$ is the $\mathsf{N}\,(0, 1/w)$ PDF. Hence

$$f_{T|W}(t \mid w) = \frac{w^{\frac{1}{2}}}{\sqrt{2\pi}} e^{-(w/2)t^2}$$

by substituting $y = t$, $\mu = 0$, and $\sigma^2 = 1/w$ in the normal PDF of Table 1.4. Furthermore, $W = X_d/d$ is a simple linear transformation of the χ^2_d random variable X_d. From the χ^2_d PDF in Table 1.4 transformed according to (1.3), the PDF of W is

$$f_W(w) = d\frac{1}{2^{d/2}\Gamma\left(\frac{d}{2}\right)}(dw)^{d/2-1} e^{-dw/2}.$$

(The factor d comes from the derivative of the linear inverse transformation $X_d = dW$.)

Combining the two PDFs, the required integral is

$$\begin{aligned} f_T(t) &= \int_0^\infty \frac{w^{\frac{1}{2}}}{\sqrt{2\pi}} e^{-(w/2)t^2} \times d \frac{1}{2^{d/2}\Gamma\left(\frac{d}{2}\right)} (dw)^{d/2-1} e^{-dw/2}\, dw, \\ &= \frac{d^{d/2}}{\sqrt{2\pi}2^{d/2}\Gamma\left(\frac{d}{2}\right)} \int_0^\infty w^{(d+1)/2-1} e^{-w(d+t^2)/2}\, dw. \end{aligned}$$

Up to constants, the integrand is the gamma PDF in Table 1.4 with $\nu = (d+1)/2$ and $\lambda = (d+t^2)/2$. Inserting the required constants $\Gamma(\nu) = \Gamma\left(\frac{d+1}{2}\right)$ and $\lambda^\nu = ((d+t^2)/2)^{(d+1)/2}$ gives

$$f_T(t) = \frac{d^{d/2}}{\sqrt{2\pi}2^{d/2}\Gamma\left(\frac{d}{2}\right)} \frac{\Gamma\left(\frac{d+1}{2}\right)}{((d+t^2)/2)^{(d+1)/2}} \int_0^\infty \frac{1}{\Gamma(\nu)} \lambda(\lambda w)^{\nu-1} e^{-\lambda w}\, dw.$$

The integrand is now a $\mathsf{Gamma}\,(\nu, \lambda)$ PDF, which integrates to 1, leaving only the constants outside the integral. Those constants can be simplified using properties of the gamma function in Section 1.5.3: $\sqrt{\pi} = \Gamma\left(\frac{1}{2}\right)$ and $\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{d}{2}\right)/\Gamma\left(\frac{d+1}{2}\right) = B\left(\frac{1}{2}, \frac{d}{2}\right)$, where $B(\cdot,\cdot)$ is the beta function. Hence,

$$f_T(t) = \frac{d^{d/2}}{B\left(\frac{1}{2}, \frac{d}{2}\right) 2^{(d+1)/2}((d+t^2)/2)^{(d+1)/2}} = \frac{1}{B\left(\frac{1}{2}, \frac{d}{2}\right)\sqrt{d}} \left(1 + \frac{t^2}{d}\right)^{-\frac{d+1}{2}},$$

which is the PDF of the t_d distribution in Table 1.4.

2.10 Appendix: Proof of the Central Limit Theorem

The Central limit theorem (CLT) in Theorem 2.2 is proved in the following steps.

1. Rewrite Z_n in terms of $Y_i - \mu$.
2. Check that Z_n in the theorem is standardized.
3. Find the MGF of $Y_i - \mu$.
4. Find the MGF of $S_n = \sum_{i=1}^n (Y_i - \mu)$
5. Find the MGF of Z_n from that of S_n.
6. Show the limiting MGF of Z_n converges to that of the standard normal as $n \to \infty$.

1. **Rewrite Z_n in terms of $Y_i - \mu$**

   Write

   $$Z_n = \frac{X_n - n\mu}{\sigma\sqrt{n}} = \frac{\sum_{i=1}^n Y_i - n\mu}{\sigma\sqrt{n}} = \frac{\sum_{i=1}^n (Y_i - \mu)}{\sigma\sqrt{n}}. \tag{2.11}$$

2. **Check that Z_n is standardized**

   We then note that

   $$\mathrm{E}(Y_i - \mu) = \mathrm{E}(Y_i) - \mu = \mu - \mu = 0$$

   and

   $$\mathrm{Var}(Y_i - \mu) = \mathrm{Var}(Y_i) = \sigma^2,$$

   based on the expectation and variance of a linear function of a random variable. It then follows that

   $$\mathrm{E}\left(\sum_{i=1}^n (Y_i - \mu)\right) = \sum_{i=1}^n \mathrm{E}(Y_i - \mu) = 0,$$

   and

   $$\mathrm{Var}\left(\sum_{i=1}^n (Y_i - \mu)\right) = \sum_{i=1}^n \mathrm{Var}(Y_i - \mu) = \sum_{i=1}^n \mathrm{Var}(Y_i) = n\sigma^2.$$

   These expressions also use results for the expectation and variance of a sum of random variables. For the variance calculation only, independence of $Y_1, \ldots, Y_n$ is also required. Hence, immediately $\mathrm{E}(Z_n) = 0$, and

   $$\mathrm{Var}(Z_n) = \left(\frac{1}{\sigma\sqrt{n}}\right)^2 \mathrm{Var}\left(\sum_{i=1}^n (Y_i - \mu)\right) = \frac{1}{\sigma^2 n} n\sigma^2 = 1,$$

   which again use results for a linear function of a random variable. Thus, Z_n has mean 0 and variance 1 and is indeed a standardized random variable.

3. **Find the MGF of $Y_i - \mu$**

   In the theorem, $Y_1, \ldots, Y_n$ have identical distributions, so they share a common MGF, $M_Y(t)$. Recall that $M_Y(t)$ is, by definition, $\mathrm{E}(e^{tY})$, and expanding the exponential function gives

   $$M_Y(t) = \mathrm{E}(e^{tY}) = \mathrm{E}\left(1 + tY + \frac{(tY)^2}{2!} + \frac{(tY)^3}{3!} + \cdots\right).$$

   Similarly, $Y_i - \mu$ in (2.11) has MGF

   $$M_{Y-\mu}(t) = \mathrm{E}(e^{t(Y-\mu)}) = \mathrm{E}\left(1 + t(Y-\mu) + \frac{(t(Y-\mu))^2}{2!} + \frac{(t(Y-\mu))^3}{3!} + \cdots\right).$$

Treating this as the expectation of a linear combination of random variables, we have

$$M_{Y-\mu}(t) = 1 + t\mathrm{E}(Y-\mu) + \frac{t^2}{2!}\mathrm{E}(Y-\mu)^2 + \frac{t^3}{3!}\mathrm{E}(Y-\mu)^3 + \cdots .$$

The expression simplifies by noting that $\mathrm{E}(Y-\mu) = 0$ and $\mathrm{E}(Y-\mu)^2 = \mathrm{Var}(Y) = \sigma^2$, whereupon

$$M_{Y-\mu}(t) = 1 + t0 + \frac{t^2}{2!}\sigma^2 + \frac{t^3}{3!}\mathrm{E}(Y-\mu)^3 + \cdots = 1 + \frac{t^2}{2!}\sigma^2 + \frac{t^3}{3!}\mathrm{E}(Y-\mu)^3 + \cdots .$$

4. **Find the MGF of $S_n = \sum_{i=1}^n (Y_i - \mu)$**
   Write $S_n = \sum_{i=1}^n (Y_i - \mu)$, which has MGF

$$M_{S_n}(t) = (M_{Y-\mu}(t))^n = \left(1 + \frac{t^2}{2!}\sigma^2 + \frac{t^3}{3!}\mathrm{E}(Y-\mu)^3 + \cdots\right)^n ,$$

   using Lemma 1.3 on the MGF of a sum of independent random variables.

5. **Find the MGF of Z_n from that of S_n**
   Write

$$Z_n = \frac{1}{\sigma\sqrt{n}} S_n ,$$

   whereupon

$$M_{Z_n}(t) = M_{S_n}\left(\frac{1}{\sigma\sqrt{n}}t\right) = \left(1 + \left(\frac{1}{\sigma\sqrt{n}}t\right)^2 \frac{\sigma^2}{2!} + \left(\frac{1}{\sigma\sqrt{n}}t\right)^3 \frac{\mathrm{E}(Y-\mu)^3}{3!} + \cdots\right)^n ,$$

   using Lemma 1.2 on the MGF of a linear function of random random variables. This simplifies to

$$M_{Z_n}(t) = \left(1 + \frac{1}{n}\frac{t^2}{2} + O\left(\frac{1}{n^{3/2}}\right)\right)^n ,$$

   where $O\left(\frac{1}{n^{3/2}}\right)$ represents further terms that are decreasing at rate $1/n^{3/2}$ or faster.

6. **Find the limiting MGF of Z_n as $n \to \infty$**
   As $n \to \infty$,

$$M_{Z_n}(t) = \left(1 + \frac{1}{n}\frac{t^2}{2} + O\left(\frac{1}{n^{3/2}}\right)\right)^n \to e^{\frac{1}{2}t^2},$$

   from the result that $(1 + x/n)^n \to e^x$ as $n \to \infty$. This is the MGF of a $\mathsf{N}(0,1)$ random variable, and the proof is complete.

Chapter 3

Statistical Estimation

Statistical methods often involve estimation of unknown parameters in probability models, as in Example 2.3 where the parameter π of the binomial distribution was estimated. Furthermore, the example showed how the properties of the estimation method lead to a probabilistic bound on the margin of error.

This chapter introduces statistical estimation more generally. It starts with some philosophy relating to the *frequentist* view: that statistical properties like bias and variance are defined by considering how estimates and other quantities would change by chance over repeated random samples from the probability model. This paves the way for the general method of maximum likelihood (Chapter 4) and analysis of its properties from this frequentist perspective.

3.1 Statistical Models: The Role of Probability

A probability calculation typically starts with a probability model (a distribution) for some random variable(s), Y, and calculates quantities like $\Pr(Y \geq c)$ or $\mathrm{E}(Y)$. These calculations say something about the values that Y generates. Mathematically, the manipulations involved may be lengthy and involve much special knowledge, especially of integration and summation, but in one sense they are easy. Given a well-defined probability model and a well-defined task like "find $\Pr(Y \geq c)$" there is only one answer. This is called *deductive* logic. It is important to note that to carry out such calculations numerically, the *values of the model's parameters must be known.*

Statistical inference also starts with a probability model but essentially uses it in a reverse process: We start with data (i.e., observed values y) from some distribution and then try to *infer* properties of the distribution. If the form of the distribution is "known" (e.g., the Poisson) the values of its parameter (e.g., μ) or parameters will usually have to be estimated from the data. In practice, one might not even know the form of the distribution, and the data will be used to infer or at least check the distribution. This *inductive* process is much less well defined and generates much

work for statisticians!

Thus, common steps in statistical inference are:

1. Choose a probability model (e.g., the binomial for discrete values or the normal for continuous values). Often, the context will suggest a distribution from first principles, but in complex problems specifying a probability distribution is usually difficult.

2. Estimate the parameter(s) of the distribution (e.g., π for the binomial or μ and σ^2 for the normal) from a sample of y values. This gives a fitted distribution, fitted to the data.

3. Check that the fitted distribution is in reasonable agreement with the data. Again, particularly for complex problems, this is not easy.

4. Use the probability model to answer questions of scientific interest, make predictions, etc.

3.2 The Frequentist Philosophy

This chapter concentrates on the second of the steps outlined above: estimating the parameters of a given (chosen) distribution. In general, let θ be a parameter of interest (e.g., π for the binomial distribution). We use some method to generate an estimate, $\hat{\theta}$, from the data.

When we compute a numeric value for $\hat{\theta}$ this is an *estimate*. For instance, consider Example 2.3 where the parameter π of interest was the proportion in a large population agreeing with the statement that Trudeau was the best Prime Minister. The binomial distribution was the probability model, and π had the estimate $\hat{\pi} = 0.36$. In general, an estimate of a parameter is based on one or more *statistics*, functions of the data like the sample mean or sample proportion. Without knowing the true value of the parameter we cannot say how good an estimate is. Furthermore, a number has no statistical properties. It is just a number.

In the frequentist philosophy of statistics, we consider the values of $\hat{\theta}$ that would have occurred in repeated random samples from the assumed probability distribution for the data. (The term "frequentist" here parallels the frequentist interpretation of probability, where probabilities are defined by long-run relative frequencies in repetitions of an experiment like rolling a die.) Often, this is a hypothetical argument as there is only one sample of data values. Nonetheless, if we consider other samples that might have occurred, each sample would generate its own value of $\hat{\theta}$, and now $\hat{\theta}$ has a *sampling distribution*, i.e., it is a random variable. The random variable is called an *estimator* of θ and will be denoted by $\tilde{\theta}$. The statistical properties of the random variable $\tilde{\theta}$, explored in the next subsection, can be used to make statements about how accurate $\tilde{\theta}$ is in estimating θ over repeated random samples of data.

Faults	Frequency	
(y)	observed	expected
0	1	1.5
1	5	3.7
2	3	4.4
3	4	3.6
4	2	2.2
5	2	1.0
> 5	0	0.6
	17	17.0

Table 3.1: Observed numbers of faults on data lines of length about 170 km, and the expected numbers under a Poisson probability model

Many textbooks and articles use the notation $\hat{\theta}$ both for an estimate, i.e., a number for a specific data set, and for the estimator. The context has to guide the reader whether this refers to an estimate or an estimator. In our text, however, the distinction between $\tilde{\theta}$ and $\hat{\theta}$ is just like that between a random variable, Y, and one of its values, y, and hopefully helps the reader better understand the frequentist concept of properties over hypothetical repeated samples. The $\tilde{\theta}$ versus $\hat{\theta}$ notation is borrowed from my colleagues Professors MacKay and Oldford, with whom I taught at the University of Waterloo.

Example 3.1 (Faults on data lines: estimating the Poisson mean)
The number of faults (y) over a period of time was collected for a sample of 17 data-transmission lines, all of length about 170 km. The observed frequencies are in Table 3.1. For instance, there are 5 lines with exactly 1 fault. The data are displayed as a histogram in Figure 3.1(a).

Suppose the number of faults per line in the population of all lines is represented by a Poisson distribution, i.e., $\mathsf{Pois}\,(\mu)$, which has PMF

$$f_Y(y) = \frac{e^{-\mu}\mu^y}{y!} \quad (y = 0, 1, \ldots, \infty; \mu > 0).$$

(There are good engineering reasons for using the Poisson distribution here.) The parameter μ is the expectation or mean of the $\mathsf{Pois}\,(\mu)$ distribution, which we need to estimate to assess the reliability of the system. As μ is the mean of the assumed distribution, we use the sample mean, $\bar{y}$, as an obvious estimate of μ from the data. The observed sample mean is

$$\bar{y} = \frac{1}{17}\sum_{i=1}^{17} y_i = \frac{0\times 1 + 1\times 5 + 2\times 3 + 3\times 4 + 4\times 2 + 5\times 2}{17} = \frac{41}{17} = 2.41$$

for the faults data. We write $\hat{\mu} = 2.41$ for the estimate of μ here.

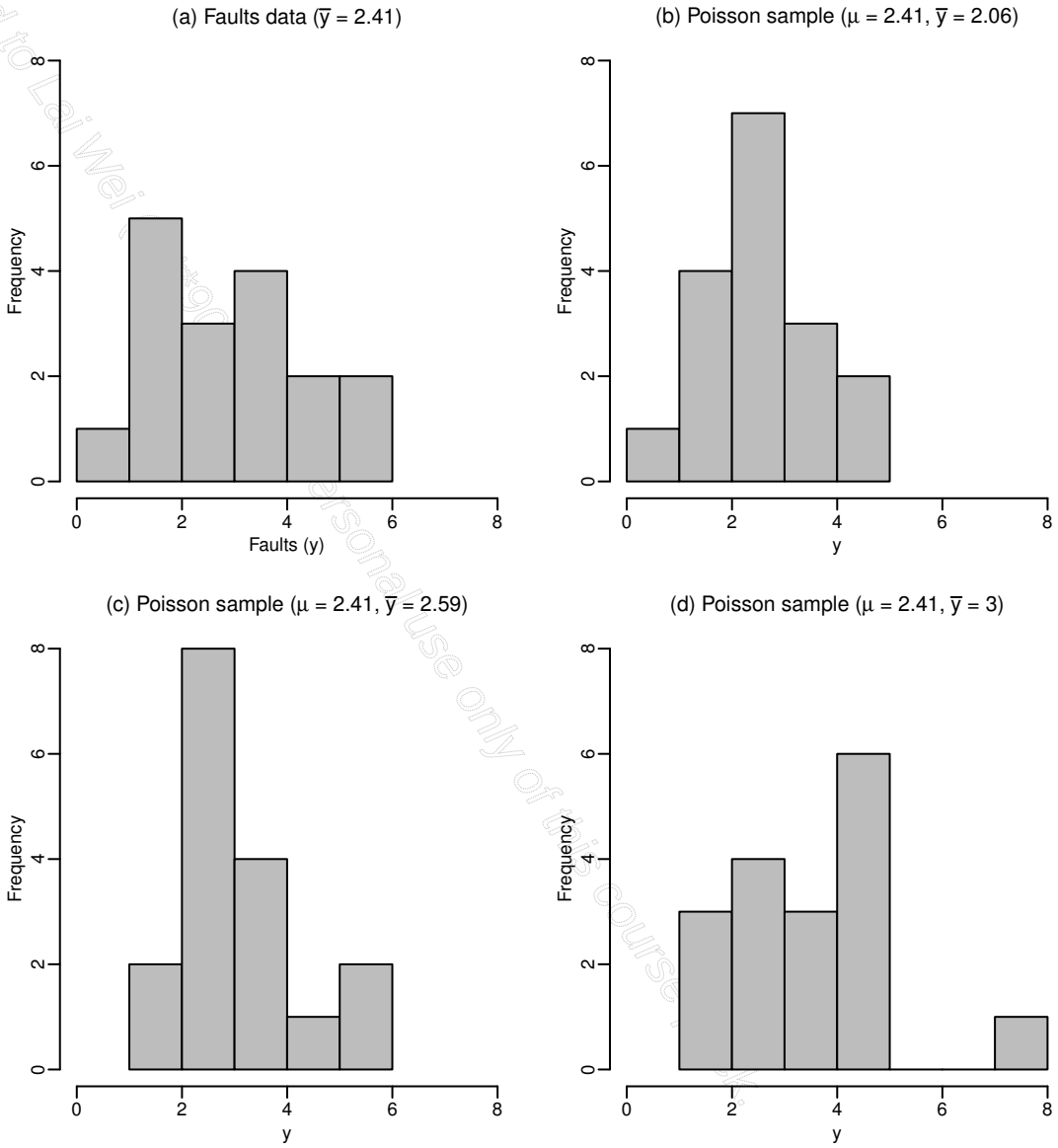


Figure 3.1: (a) Histogram of the faults data in Table 3.1. (b), (c), and (d) Histograms of three random samples of size $n = 17$ from a Poisson distribution with the parameter μ set to 2.41, which is the value of $\bar{y}$ observed for the faults data. The sample mean, $\bar{y}$, is also given for each random sample.

How good is this estimate? This question is difficult to answer for the specific sample, but we can look at properties over random samples. Such properties are defined mathematically in Section 3.3. For now, we can informally compare the sample data with random samples of size $n = 17$ drawn from a $\mathsf{Pois}(\mu)$ distribution. The value of μ is unknown—the objective here is to estimate it—but we proceed by setting μ to 2.41, the estimate from the data. This gives an idea of how much the data and hence $\hat{\mu}$ vary from one sample to another just by chance if μ is about 2.41. The histograms in Figures 3.1(b), (c), and (d) show three such random samples created via `rpois` in `R`. Two features of the plots are worth noting.

- There is a fair amount of difference in shape in the histograms of Figures 3.1(b), (c), and (d) because the sample size is small. The histogram of the faults data in Figure 3.1(a) does not stand out compared to the Poisson samples.
- The sample mean, $\bar{y}$, is reported for the three samples from a $\mathsf{Pois}(\mu = 2.41)$ distribution. It ranges from 2.06 to 3.00. Thus, just by chance, the sample mean varies from one random sample to another. Such *sampling variation across repeat random samples* is the basis for treating the sample mean as a random variable, $\bar{Y}$. The variation in $\bar{Y}$ across samples defines the uncertainty attached to an estimate in this chapter: the frequentist paradigm.

Although Figure 3.1 demonstrates considerable variation in the sample mean here just due to chance, the data are good enough to narrow down the range of possible values of μ in a statistical sense. Figure 3.2 repeats Figure 3.1, but now $\mu = 5$ when generating three random samples from the $\mathsf{Pois}(\mu)$ distribution. The histogram of the data in Figure 3.2(a) now stands out: The other three histograms are shifted to the right. Furthermore, the sample means from the $\mathsf{Pois}(\mu = 5)$ distributions still vary but take values much larger than $\bar{y} = 2.41$ for the faults data. We really need more random samples to say much more, but it looks like the data rule out $\mu = 5$ via this statistical sampling argument.

Is the Poisson distribution a reasonable probability model here? If we set μ to the estimate 2.41, the Poisson PMF is

$$f_Y(y) = \frac{e^{-2.41}(2.41)^y}{y!}.$$

Thus, in a sample of size 17 we would expect $17 f_Y(0) = 17 \times 0.0898 = 1.5$ values of 0, $17 f_Y(1) = 17 \times 0.216 = 3.7$ values of 1, etc. These expected frequencies are also given in Table 3.1. The agreement between observed and expected frequencies appears good. *Goodness of fit* between data and a hypothesized distribution is taken up more formally in Section 8.4. ◇◇◇

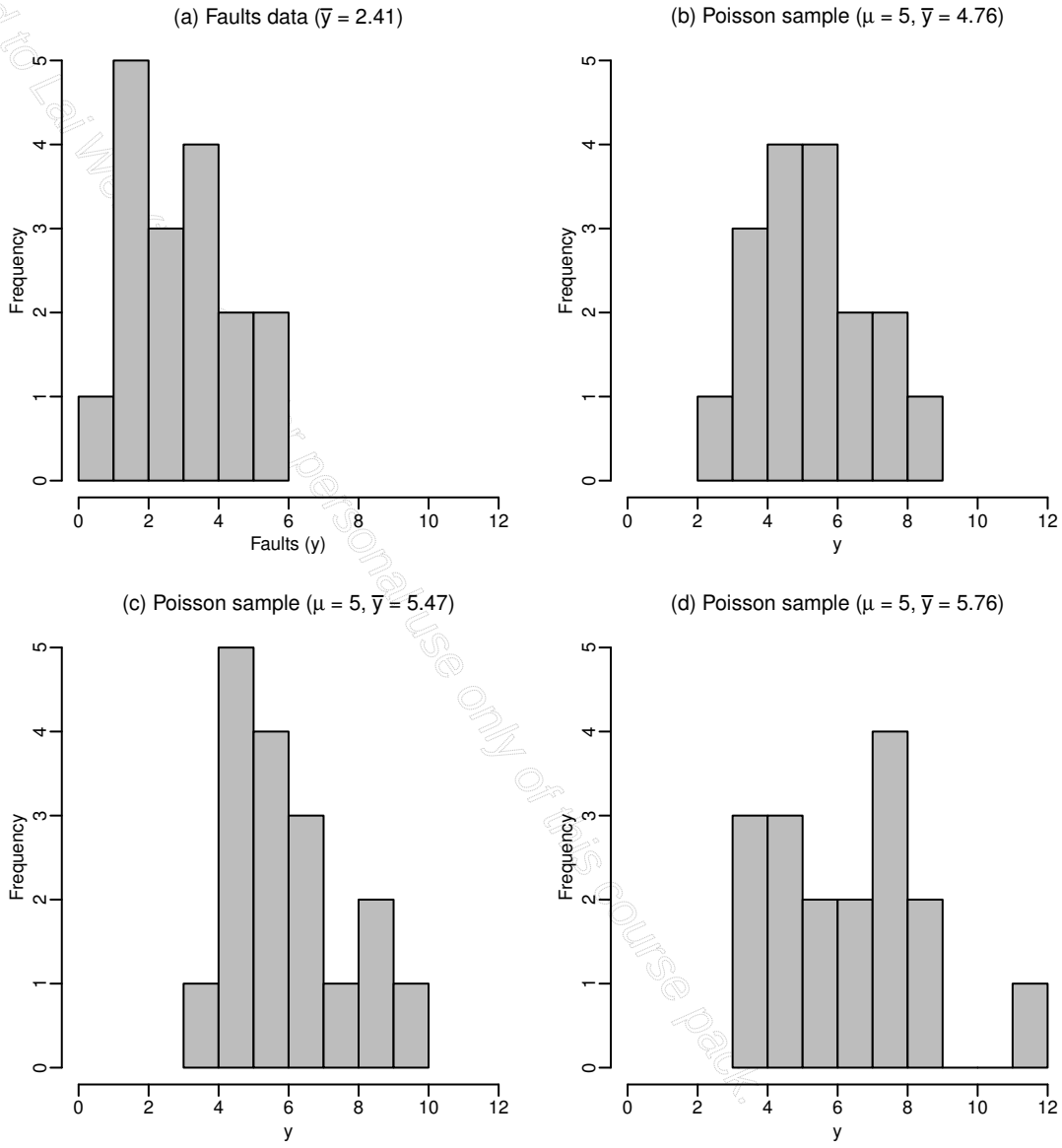


Figure 3.2: (a) Histogram of the faults data in Table 3.1. (b), (c), and (d) Histograms of three random samples of size $n = 17$ from a Poisson distribution with the parameter μ set to 5, which is much larger than the $\bar{y} = 2.41$ observed for the faults data. The sample mean, $\bar{y}$, is also given for each random sample.

3.3 Properties of an Estimator

Let $\tilde{\theta}$ be an estimator of the parameter θ. How good is $\tilde{\theta}$: How much error does it have in estimating θ? We can answer this question, at least probabilistically, by examining the statistical properties of $\tilde{\theta}$ and hence its error, $\tilde{\theta} - \theta$. All the properties, like expectation and variance of $\tilde{\theta}$, are with respect to its sampling distribution over repeated random samples.

3.3.1 Bias

The bias of $\tilde{\theta}$ as an estimator of θ follows from its expectation.

Definition 3.1 (Bias)
The bias of the estimator $\tilde{\theta}$ of a parameter θ is

$$\mathit{Bias}(\tilde{\theta}) = \mathrm{E}(\tilde{\theta}) - \theta.$$

Rewriting the bias as

$$\mathrm{Bias}(\tilde{\theta}) = \mathrm{E}(\tilde{\theta}) - \theta = \mathrm{E}(\tilde{\theta} - \theta),$$

we see that the bias is the mean of the error distribution.

If $\mathrm{E}(\tilde{\theta}) = \theta$, then the bias is zero, and we say that $\tilde{\theta}$ is unbiased. Unbiasedness means that the sampling distribution of $\tilde{\theta}$ is centred on the true value, θ, in the sense that the error might be positive or negative from one sample to another, but these errors cancel out on average. Thus, unbiasedness is desirable.

3.3.2 Variance

The variance of $\tilde{\theta}$ as estimator of θ is defined just like the variance of any random variable (Definition 1.3). Small values of $\mathrm{Var}(\tilde{\theta})$ are desirable in the sense that the estimator has small variability over random samples.

3.3.3 Mean squared error

A measure of accuracy summarizing the distribution of the error $\tilde{\theta} - \theta$ is the *mean squared error* (MSE), the expectation of the squared error.

Definition 3.2 (Mean squared error (MSE))
The mean squared error of $\tilde{\theta}$ as an estimator of θ is

$$MSE(\tilde{\theta}) = \mathrm{E}(\tilde{\theta} - \theta)^2.$$

It can be decomposed as

$$MSE(\tilde{\theta}) = Var(\tilde{\theta}) + Bias^2(\tilde{\theta}).$$

The MSE combines the bias and variance properties of $\tilde{\theta}$ in a single measure. The decomposition of the MSE into its two components can be seen by writing:

$$\begin{aligned}
\mathrm{MSE}(\tilde{\theta}) =& \mathrm{E}\big(\tilde{\theta} - \theta\big)^2 = \mathrm{E}\big(\tilde{\theta} - \mathrm{E}(\tilde{\theta}) + \mathrm{E}(\tilde{\theta}) - \theta\big)^2 \\
=& \mathrm{E}\big(\tilde{\theta} - \mathrm{E}(\tilde{\theta}) + \mathrm{Bias}(\tilde{\theta})\big)^2 \quad \text{(from the definition of Bias}(\tilde{\theta}) \\
=& \mathrm{E}\big((\tilde{\theta} - \mathrm{E}(\tilde{\theta}))^2 + 2(\tilde{\theta} - \mathrm{E}(\tilde{\theta}))\,\mathrm{Bias}(\tilde{\theta}) + \mathrm{Bias}^2(\tilde{\theta})\big) \\
=& \mathrm{E}\big(\tilde{\theta} - \mathrm{E}(\tilde{\theta})\big)^2 + 2\mathrm{E}\big(\tilde{\theta} - \mathrm{E}(\tilde{\theta})\big)\,\mathrm{Bias}(\tilde{\theta}) + \mathrm{Bias}^2(\tilde{\theta}) \\
& \text{(from the expectation of a linear combination of random variables} \\
& \text{and noting that E}(\tilde{\theta}) \text{ and hence Bias}(\tilde{\theta}) \text{ are constants)} \\
=& \mathrm{Var}(\tilde{\theta}) + 2 \times 0 \times \mathrm{Bias}(\tilde{\theta}) + \mathrm{Bias}^2(\tilde{\theta}) \\
& \text{(from the definition of variance and E}\big(\tilde{\theta} - \mathrm{E}(\tilde{\theta})\big) = \mathrm{E}(\tilde{\theta}) - \mathrm{E}(\tilde{\theta}) = 0) \\
=& \mathrm{Var}(\tilde{\theta}) + \mathrm{Bias}^2(\tilde{\theta}).
\end{aligned}$$

Thus, mean squared error penalizes an estimator both for having bias (i.e., the wrong expectation or mean) and for its variance. As MSE decreases, the accuracy of $\tilde{\theta}$ in estimating θ increases in this statistical sense.

If $\tilde{\theta}$ is unbiased, then $\mathrm{MSE}(\tilde{\theta}) = \mathrm{Var}(\tilde{\theta})$. This is often the case, exactly or approximately, hence the emphasis on variance calculations in statistical inference about parameters.

Example 3.2 (Faults on data lines: properties of the estimator of μ)
Example 3.1 used $\hat{\mu} = \bar{y} = 2.41$ as an obvious estimate of the Poisson parameter, μ, for the faults data with $n = 17$. To assess the accuracy of this estimate we can think in terms of repeated random samples of the data and the estimator $\tilde{\mu}$. How far can $\tilde{\mu}$ stray from μ in a probabilistic sense just by chance sampling variation?

The sample mean $\bar{Y}$ is the mean of 17 IID Poisson random variables here. From the properties in Table 1.3, a Poisson random variable Y has mean μ and variance μ. The following properties of $\bar{Y}$ are shown more generally in Section 2.4.1.

- **Unbiasedness.** As $\bar{Y} = \sum_{i=1}^{17} Y_i/17$ has the same expectation as $\mathrm{E}(Y_i) = \mu$, we have $\mathrm{E}(\bar{Y}) = \mu$ and $\tilde{\mu}$ is an unbiased estimator of μ for any value of n.
- **Variance.** As $\bar{Y}$ has variance $\mathrm{Var}(Y_i)/17$, we have $\mathrm{Var}(\tilde{\mu}) = \mu/17$.

- **Estimated standard deviation or standard error.** Because $\text{Var}(\tilde{\mu}) = \mu/17$ depends on μ, in practice it has to be estimated. An obvious estimate is $\widehat{\text{Var}}(\tilde{\mu}) = \hat{\mu}/17 = 2.41/17 = 0.142$. Equivalently, the estimated standard deviation of $\tilde{\mu}$ is $\widehat{\text{sd}}(\tilde{\mu}) = \sqrt{0.142} = 0.38$. An estimated standard deviation is often called a *standard error,* which we write as $\text{se}(\tilde{\mu})$ here. Note that only a random variable, here $\tilde{\mu}$, can have a variance or standard deviation, whether it is estimated or not. Thus, $\widehat{\text{sd}}(\tilde{\mu}) = \text{se}(\tilde{\mu}) = 0.38$ makes statistical sense, but $\widehat{\text{sd}}(\hat{\mu})$ or $\text{se}(\hat{\mu})$ do not.

  Authors will often write a looser statement like, "an estimate $\hat{\mu} = 2.41$ with a standard error of 0.38," but we need to interpret "with" in the sense of a property of $\tilde{\mu}$ not $\hat{\mu}$.

- **Mean squared error.** Because $\tilde{\mu}$ is an unbiased estimator of μ, the MSE of $\tilde{\mu}$ is also $\mu/17$.

To summarize, while the estimate $\hat{\mu} = 2.41$ is a number with no statistical properties, the corresponding estimator $\tilde{\mu}$ is unbiased over repeated samples of size $n = 17$. Hence, the MSE of the estimator is entirely due to its variance, namely $\mu/17$. The estimator has an estimated variance of 0.142 or equivalently an estimated standard deviation of 0.38. ♢♢♢

Several facts specific to the Poisson distribution were used in Example 3.2 to obtain statistical properties of the estimator of μ. Chapter 4 takes a general approach, the method of maximum likelihood, for estimation and quantification of the uncertainty of estimators. It relies less on knowledge of specific properties and can be extended to applications where exact properties are unavailable.

3.3.4 Practical perspective

Estimation bias is usually of little practical consequence, as it is typically zero or small relative to the standard deviation of an estimator. The following example questions whether a familiar unbiasedness argument is compelling.

Example 3.3 (Sample variance: divisor of $n-1$ or n?)
Let $Y_1, \ldots, Y_n$ have constant mean μ and constant variance σ^2. A vexing question to countless students of statistics is why not define their sample variance with a divisor of n, i.e.,

$$\widetilde{\sigma^2} = \frac{1}{n} X,$$

where

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

instead of the familiar $S^2 = X/(n-1)$?

A standard answer is that $\widetilde{\sigma^2}$ is biased, whereas S^2 is not (Exercise 2.9). The bias turns out to be small relative to $\text{sd}(\widetilde{\sigma^2})$, however, in the important special case that the Y_i are also normally distributed. Then, via the arguments of Section 2.4.2,

$$\text{E}(X) = (n-1)\sigma^2 \quad \text{and} \quad \text{Var}(X) = 2(n-1)\sigma^4,$$

whereupon

$$\text{Bias}(\widetilde{\sigma^2}) = \text{E}(\widetilde{\sigma^2}) - \sigma^2 = \frac{1}{n}(n-1)\sigma^2 - \sigma^2 = -\frac{1}{n}\sigma^2$$

and

$$\text{Var}(\widetilde{\sigma^2}) = \frac{1}{n^2}\text{Var}(X) = \frac{2(n-1)}{n^2}\sigma^4.$$

We see that for any $n \geq 2$, the bias of $\widetilde{\sigma^2}$ is smaller in magnitude than its standard deviation:

$$\frac{\text{Bias}(\widetilde{\sigma^2})}{\text{sd}(\widetilde{\sigma^2})} = \frac{-\sigma^2/n}{\sqrt{2(n-1)}\sigma^2/n} = -\frac{1}{\sqrt{2(n-1)}}.$$

Of course, no bias is better than a "small" bias, but Exercise 3.4 shows that $\widetilde{\sigma^2}$ has a smaller standard deviation and smaller MSE than S^2 and is more accurate overall. The most compelling case for the use of S^2 with divisor $n-1$ is that, for IID normal random variables, X has a χ^2_{n-1} distribution with $n-1$ degrees of freedom, and hence S^2 fits the mathematical requirements of the t distribution (Section 2.4.3). ♢♢♢

Usually, the MSE of an estimator either equals its variance (unbiased estimation) or is not much larger than the variance (small squared bias). Hence, we will see that when a confidence interval is calculated to quantify error it is nearly always based on the estimator's standard deviation only.

Estimation bias may be ignorable, but there are many other possible source of bias in an empirical study. They include sampling from a population other than the target one or bias in the measurement system producing data. Moreover, other sources are difficult to study in a quantitative way by analysis of the data (and hence will not be pursued much in this text). Best practice is to mitigate sources of bias proactively by careful study design and objective measurement.

3.3.5 Consistency

Definition 3.3 (Consistency)
The estimator $\tilde{\theta}_n$ of a parameter θ based on a sample of size n is consistent for estimating θ if

$$\Pr(|\tilde{\theta}_n - \theta| < \epsilon) \to 1 \quad \textit{as } n \to \infty$$

for any fixed error, $\epsilon > 0$.

Consistency of $\tilde{\theta}_n$ is a special case of *convergence in probability* of a random variable ($\tilde{\theta}_n$ here) to a constant (θ).

Consistency requires that both the bias and variance of $\tilde{\theta}$ go to zero as $n \to \infty$. Hence, a necessary and sufficient condition is that the mean squared error goes to zero.

Many estimators are of the form of a sample mean, which includes the sample proportion, or a simple function of the sample mean. If the objective is to estimate the mean of the underlying distribution, and the sample consists of independent observations, then consistency of such estimators is easily established as a special case of the Weak Law of Large Numbers (WLLN).

Theorem 3.1 (Weak law of large numbers (WLLN))
Let

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

where $Y_1, \ldots, Y_n$ are n independent random variables, each with mean μ and variance σ^2 (both of which must exist). Then, for any $\epsilon > 0$,

$$\Pr(|\bar{Y}_n - \mu| < \epsilon) \to 1 \quad \text{as } n \to \infty.$$

The law of large numbers is not about large observations! It is concerned with a *large sample size, n*. Also, the random variables are *not* the individual elements of the sample. Rather, as the sample size, n, increases, there is a sequence of sample means, $\bar{Y}_n$, computed from more and more elements.

In the WLLN, ϵ can be made arbitrarily small as long as it is positive. Thus, the theorem says that the distribution of $\bar{Y}_n$ is more and more concentrated around an arbitrarily small neighbourhood of μ as the sample size grows. Thus, $\bar{Y}_n$ is a consistent estimator of μ.

The proof of the WLLN is straightforward. From (1.9) we have

$$\mathrm{E}(\bar{Y}_n) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \sum_{i=1}^{n}\frac{1}{n}\mathrm{E}(Y_i) = \sum_{i=1}^{n}\frac{1}{n}\mu = \mu.$$

Similarly, we use (1.10) to get $\mathrm{Var}(\bar{Y}_n)$. Because we are assuming the Y_i are independent, all the covariance terms, $\mathrm{Cov}(Y_i, Y_j)$ for $i \neq j$, in (1.10) are zero. Thus,

$$\mathrm{Var}(\bar{Y}_n) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \sum_{i=1}^{n}\frac{1}{n^2}\mathrm{Var}(Y_i) = \sum_{i=1}^{n}\frac{1}{n^2}\sigma^2 = \frac{\sigma^2}{n}.$$

Clearly, $\mathrm{Var}(\bar{Y}_n) \to 0$ as $n \to \infty$, and the result follows by putting $t = \epsilon$ in Chebyshev's inequality (Theorem 1.1).

Example 3.4 (Opinion polls: weak law of large numbers)
Each voter in the population of eligible voters intends to vote for the Statistics for Everybody Party ($y = 1$) or will not ($y = 0$) in the next federal election.

To estimate the population proportion, π, intending to vote for Statistics for Everybody, a random sample of n eligible voters is taken.

Let Y_i be the voting intention for person i in the sample; it is a Bernoulli random variable with $\Pr(Y_i = 1) = \pi$, i.e., $\mathsf{Bern}\,(\pi)$. Note that $\mathrm{E}(Y_i) = \pi$ and $\mathrm{Var}(Y_i) = \pi(1-\pi)$. We further assume that $Y_1, \ldots, Y_n$ are independent random variables (which would be approximately true if the population size is large relative to the sample size, which it usually is).

Consider estimating π using the sample mean,

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

where the subscript n on $\bar{Y}_n$ emphasizes that we are investigating the impact of increasing the sample size. (The sample mean is also a sample proportion here: As Y_i take values 0 and 1 only, $\bar{Y}$ is the proportion of 1's in the sample.) Then, from Exercise 1.17,

$$\mathrm{E}(\bar{Y}_n) = \pi,$$

and

$$\mathrm{Var}(\bar{Y}_n) = \frac{\pi(1-\pi)}{n}.$$

The variance clearly tends to zero as n increases. Thus, by Chebyshev's inequality (Theorem 1.1) the estimator $\bar{Y}$ is in an arbitrarily small neighbourhood of the true value π with probability approaching 1 as the sample size n grows, and $\bar{Y}_n$ is a consistent estimator of π.

We could argue the same result directly from the WLLN. The conditions of the WLLN are easily verified: $Y_1, \ldots, Y_n$ are independent and the mean and variance of Y_i both exist. Hence, via the WLLN, the sample mean of $Y_1, \ldots, Y_n$ is a consistent estimator of $\mathrm{E}(Y_i) = \pi$. ♢♢♢

Consistency of $\tilde{\theta}$ is a natural requirement: an estimator that does not converge to θ for an infinite sample size should be questioned.

3.3.6 Relative Error

Implicitly, the measures of error used so far have related to *absolute* error. For example,

$$\mathrm{Var}(\tilde{\theta}) = \mathrm{E}\left(\left(\tilde{\theta} - \theta\right)^2\right),$$

where positive and negative errors $\tilde{\theta} - \theta$ are treated the same due to squaring. Similarly, in MSE and its squared bias component, the sign of the bias is immaterial.

For some applications, *relative* error,

$$\frac{\tilde{\theta} - \theta}{\theta} = \frac{\tilde{\theta}}{\theta} - 1,$$

π	0.01	0.02	0.05	0.10	0.20	0.30	0.40	0.50
n_{abs}	44	88	212	400	712	934	1067	1112
n_{rel}	110 000	54 445	21 112	10 000	4445	2593	1667	1112

Table 3.2: Sample size to estimate the binomial parameter π: n_{abs} achieves $\text{sd}(\tilde{\pi}) \leq 0.015$ and n_{rel} achieves $\text{sd}(\tilde{\pi})/\pi \leq 0.03$

and summary measures based on it are more compelling, however. The following sample-size calculation illustrates that the different definitions of error can have important consequences.

Example 3.5 (Binomial distribution: sample size determination)
A common question is, “What should the sample size be?” For instance, the opinion poll of Example 2.3 had $n = 1000$. Why are the sample sizes of opinion polls typically of order one thousand?

Example 2.3 boils down to estimation of π in a $\textsf{Bin}\,(n, \pi)$ probability model. Suppose the requirement is to determine n such that $\text{sd}(\tilde{\pi}) \leq 0.015$. That requirement is often stated as 1.5 percentage *points,* to emphasize that it is an absolute number of “points” on the percentage scale. Rearrangement of

$$0.015 = \text{sd}(\tilde{\pi}) = \sqrt{\frac{\pi(1-\pi)}{n}}$$

yields

$$n = \frac{\pi(1-\pi)}{(0.015)^2},$$

as shown in the n_{abs} row of Table 3.2. The required sample size depends on the true value π. It is maximized when $\pi = 0.5$ and decreases as π decreases. Similarly n_{abs} decreases as π increases above 0.5 (not shown) in a symmetric way. Thus, a sample size $n_{\text{abs}} = 1112$ will give $\text{sd}(\tilde{\pi}) \leq 0.015$ for any π, and a 95% confidence interval using the standard error and a normal approximation will be no wider that $\pm z_{0.0975}\text{se}(\tilde{\pi}) = \pm 1.96 \times 0.015 = \pm 0.0294$, or about plus or minus 3 percentage points. As n_{abs} does not change much for, say, $0.2 < \pi < 0.8$, a sample size based on the worst case, $\pi = 0.5$, is often employed when π is anticipated in such a range.

The previous arguement becomes less relevant when π is small, however. Table 3.2 gives $n_{\text{abs}} = 44$ for $\pi = 0.01$, for instance, but the requirement $\text{sd}(\tilde{\pi}) \leq 0.015$ probably needs tightening: the standard deviation is larger than the true value. More natural is to control the standard deviation to be small relative to the true value, e.g., $\text{sd}(\tilde{\pi}) \leq 0.03\pi$. Applying the rule for the variance (hence standard deviation) of a linear function of a random variable, the new requirement is equivalent to

$$0.03 = \frac{\text{sd}(\tilde{\pi})}{\pi} = \text{sd}\left(\frac{\tilde{\pi}}{\pi}\right) = \text{sd}\left(\frac{\tilde{\pi}}{\pi} - 1\right) = \text{sd}\left(\frac{\tilde{\pi}-\pi}{\pi}\right),$$

i.e., the standard deviation of the relative error.

Solving

$$0.03\pi = \mathrm{sd}(\tilde{\pi}) = \sqrt{\frac{\pi(1-\pi)}{n}},$$

for n yields

$$n = \frac{(1-\pi)}{(0.03)^2\pi},$$

as shown in the n_{rel} row of Table 3.2. For $\pi = 0.5$, there is no impact on sample size as the absolute and relative requirements are chosen to be equivalent at $\pi = 0.5$. But n_{rel} increases rapidly as π approaches zero: even for a moderately small $\pi = 0.05$ we need a sample size of $n_{\mathrm{rel}} = 21\,112$, increasing to 110 000 at $\pi = 0.01$. Relaxing the requirement so that $\mathrm{sd}(\tilde{\pi}) \leq 0.1\pi$ when $\pi = 0.01$ still needs $n_{\mathrm{rel}} = 9900$. Estimating a small probability with good relative accuracy requires a huge sample size. ♢♢♢

3.4 Comparing Estimators

Bias and variance properties allow comparison of candidate estimators when the choice of estimator is not obvious.

An interesting case is the Laplace (double-exponential) distribution. As noted in Section 1.5.4, the distribution is symmetric around the location parameter μ, which is therefore both the mean and median. Should the sample mean or the sample median from a random sample be used as its estimator? (The sample median is the "middle" value in the data.) It turns out that both estimators are unbiased, but the sample median has the smaller variance for $n \geq 3$ (Sarhan, 1954). These results are demonstrated numerically in Exercise 3.5. The sample median is naturally more robust to unusual outlying observations that can arise due to the Laplace PDF's fat tails, and this intuition is borne out by the theoretical properties.

3.5 Getting It Done in R

In Example 3.1 random samples were drawn from the Poisson distributions. R has functions to generate random numbers for all the distributions listed in Table 1.9. They have names starting with `r`.

Thus, R has functions with names starting with `d` for the PDF or PMF, `p` for the CDF, `q` for a quantile, and `r` for generating random numbers. Table 3.3 lists these functions for the normal distribution, for instance. The function `rnorm` has an argument `n` for the sample size. Hence, `rnorm(n = 10)` would generate a sample of size $n = 10$ from the standard normal.

Purpose	R function
PDF	`dnorm(y, mean = 0, sd = 1)`
CDF	`pnorm(y, mean = 0, sd = 1)`
Quantile	`qnorm(p, mean = 0, sd = 1)`
Random number	`rnorm(n, mean = 0, sd = 1)`

Table 3.3: R functions to return the PDF, CDF, quantile, or random numbers for the normal distribution

3.6 Learning Outcomes

On completion of this chapter you should be able to carry out the following tasks.

1. Explain the difference between an estimate and an estimator of a parameter and why the distinction is important to define statistical properties.
2. Derive the following properties for a specific estimator: its bias (which might be zero); its variance; its mean squared error; and whether or not it is consistent.
3. Use the WLLN (Theorem 3.1 in Section 3.3.5), and check the conditions, to show the sample mean is a consistent estimator of the mean of the underlying distribution.
4. Explain your reasoning. When using a result such as the expectation or variance of a linear combination of random variables to derive a property of an estimator, briefly state the result you are using. If the result depends on an assumption such as statistical independence of random variables, remind the reader that you are using the assumption.

3.7 Exercises

Exercise 3.1
Frequentist inference considers variation across hypothetical repeated random samples. In some cases the *design* of a study involves a step with a probabilistic mechanism to select a sample of data. Briefly describe the probabilistic mechanism and the set of possible random samples for the following studies:

1. The opinion poll in Example 2.3; and
2. The clinical trial in Example 1.15.

Exercise 3.2

Suppose we obtain n independent observations, $Y_1, \ldots, Y_n$, from a Poisson probability model to estimate the Poisson parameter μ. Consider the estimator $\tilde{\mu} = \bar{Y}$.

1. Show $\tilde{\mu}$ is unbiased.
2. Find $\text{Var}(\tilde{\mu})$ (an exact formula for the variance).
3. Is $\tilde{\mu}$ a consistent estimator of μ?
4. Consider $\widetilde{\text{Var}}(\tilde{\mu}) = \tilde{\mu}/n$ as an estimator of $\text{Var}(\tilde{\mu})$.
   (a) Show that $\widetilde{\text{Var}}(\tilde{\mu})$ is an unbiased estimator of $\text{Var}(\tilde{\mu})$.
   (b) What is the variance of $\widetilde{\text{Var}}(\tilde{\mu})$?
   (c) Is $\widetilde{\text{Var}}(\tilde{\mu})$ a consistent estimator of $\text{Var}(\tilde{\mu})$?

Exercise 3.3

Let $Y_1, \ldots, Y_n$ be independent $\mathsf{N}(\mu, \sigma^2)$ random variables. Their sample variance is

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

Treat S^2 as an estimator, i.e., a random variable. Is it a consistent estimator of σ^2?

Exercise 3.4

[Final exam, 2011-12, Term 1] Let $Y_1, \ldots, Y_n$ be independent normal random variables, each with mean μ and variance σ^2. We want to estimate σ^2 from such a sample of size $n \geq 2$ when μ is also unknown.

This question investigates the exact properties of $\widetilde{\sigma^2} = X/n$, where $X = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$. You may use without proof: (1) the result that X/σ^2 has a χ^2_{n-1} distribution; and (2) statistical properties of the χ^2 distribution.

1. Show that the expectation of $\widetilde{\sigma^2}$ is $\sigma^2(n-1)/n$.
2. Show that the variance of $\widetilde{\sigma^2}$ is $2(n-1)\sigma^4/n^2$.
3. Hence, give and simplify an expression that summarizes the accuracy of $\widetilde{\sigma^2}$ as an estimator of σ^2.
4. Is $\widetilde{\sigma^2}$ a consistent estimator of σ^2? Explain briefly.
5. The sample variance with divisor $n-1$, namely $S^2 = X/(n-1)$, has $\text{E}(S^2) = \sigma^2$ and $\text{Var}(S^2) = 2\sigma^4/(n-1)$. Give one advantage and one disadvantage of S^2 relative to $\widetilde{\sigma^2}$ as an estimator of σ^2.
6. Explain which estimator of σ^2 is the more accurate, S^2 or $\widetilde{\sigma^2}$.

```
# Number of repeat samples
n.reps <- 10000
# Vectors to store the sample means and medians
sample.mean   <- rep(0, times = n.reps)
sample.median <- rep(0, times = n.reps)

# Generate repeat samples
library(rmutil)
for (k in 1:n.reps) {
   # Random sample k
   a.sample <- rlaplace(...)
   # Save mean and median
   sample.mean[k]   <- mean(a.sample)
   sample.median[k] <- median(a.sample)
}
```

Figure 3.3: R code for Exercise 3.5

Exercise 3.5

A random variable Y with a Laplace (double-exponential) distribution has PDF

$$f(y \mid \mu, \phi) = \frac{1}{2\phi} e^{-\frac{|y-\mu|}{\phi}} \quad (-\infty < y < \infty; -\infty < \mu < \infty; \phi > 0)$$

(see Table 1.4). As the Laplace distribution is symmetric, μ is the expected value (mean) and median. Hence, in this exercise we compare as estimators of μ the sample mean and the sample median from a random sample, $Y_1, \ldots, Y_n$.

We will also need the fact that $2\phi^2$ is the variance of the Laplace distribution.

1. Consider first the sample mean as an estimator, i.e., $\tilde{\mu} = \bar{Y}$. Its properties can be found theoretically.

   (a) What is the expected value of $\tilde{\mu}$? Is it unbiased as an estimator of μ?

   (b) Give a formula for $\text{sd}(\tilde{\mu})$. What is the numerical value of $\text{sd}(\tilde{\mu})$ if $n = 25$ and $\phi = 10$, say?

2. Now consider the sample median as an estimator, i.e., $\tilde{\mu} = \text{median}(Y_1, \ldots, Y_n)$. Note that $\tilde{\mu}$ is now a different estimator. It would be a little more difficult to establish its theoretical properties, and we resort to numerical simulation. Simulation has a learning bonus, however. We will be generating repeat samples, the principle underlying the frequentist philosophy of statistical inference.

   (a) Use `rlaplace` in `library(rmutil)` to generate a random sample of size $n = 25$ from the Laplace distribution with $\mu = 1$ and $\phi = 10$. Repeat for 10 000 samples as in the R code of Figure 3.3. You need to give values to the parameters passed to `rlaplace`. The code also stores the sample means to check a theoretical property of $\bar{Y}$ here.

After running your code, you have an empirical distribution of the sample median (and mean) from many repeat samples. Because the number of repeat samples is fairly large, the empirical distribution is a good approximation to the true distribution of the sample median. We will estimate the properties of the sample median from this empirical distribution.

(b) Apply `mean` to the 10 000 values of the sample median from the random samples to find the (approximate) expected value of the sample median. Does it appear that $\tilde{\mu} = \text{median}(Y_1, \ldots, Y_n)$ is an unbiased estimator?

(c) Apply `sd` to the 10 000 values of the sample median to find the (approximate) standard deviation of the sample median.

3. Check the theoretical property in part 1b by applying `sd` to the 10 000 values of the sample mean.

4. Which estimator appears to be a more accurate estimator of the parameter μ of the Laplace distribution, the sample mean or the sample median?

Chapter 10

Solutions

Solution 1.1

1. 0.001102, 0.004888, and 0.007978, respectively

2. 0.01102, 0.04888, and 0.07978, respectively

3. For the picture, think of the midpoint rule (also known as the rectangle method) for approximating a definite integral.

Solution 1.4

1. The Poisson distribution has PMF

$$f_Y(y) = \frac{e^{-\mu}\mu^y}{y!} \quad (y = 0, 1, \ldots, \infty;\ \mu > 0).$$

Therefore, from Definition 1.1, the expectation is

$$\mathrm{E}(Y) = \sum_{y=0}^{\infty} y \frac{e^{-\mu}\mu^y}{y!}.$$

The sum can be computed by noting there is no contribution from $y = 0$, cancelling y in the numerator and denominator of the summand, and then going back to a sum starting at $y = 0$:

$$E(Y) = \sum_{y=0}^{\infty} y \frac{e^{-\mu}\mu^y}{y!} = \sum_{y=1}^{\infty} y \frac{e^{-\mu}\mu^y}{y!} = e^{-\mu}\mu \sum_{y=1}^{\infty} \frac{\mu^{y-1}}{(y-1)!}$$
$$= e^{-\mu}\mu \sum_{y=0}^{\infty} \frac{\mu^y}{y!} = e^{-\mu}\mu e^{\mu} = \mu.$$

2. The simplest version of Definition 1.3 to use here is $\mathrm{Var}(Y) = \mathrm{E}(Y^2) - (\mathrm{E}(Y))^2$. It requires computation of $\mathrm{E}(Y^2)$, which proceeds in a similar way to $\mathrm{E}(Y)$:

$$E(Y^2) = \sum_{y=0}^{\infty} y^2 \frac{e^{-\mu}\mu^y}{y!} = \sum_{y=1}^{\infty} y^2 \frac{e^{-\mu}\mu^y}{y!} = \mu \sum_{y=1}^{\infty} y \frac{e^{-\mu}\mu^{y-1}}{(y-1)!} = \mu \sum_{y=0}^{\infty} (y+1) \frac{e^{-\mu}\mu^y}{y!}$$
$$= \mu \left(\sum_{y=0}^{\infty} y f_Y(y) + \sum_{y=0}^{\infty} f_Y(y) \right) = \mu(\mathrm{E}(Y) + 1) = \mu(\mu + 1),$$

using $\mathrm{E}(Y)$ from part 1 and noting that the PMF sums to 1.

Hence,

$$\mathrm{Var}(Y) = \mathrm{E}(Y^2) - (\mathrm{E}(Y))^2 = \mu(\mu + 1) - \mu^2 = \mu.$$

Solution 1.5

4. $(1-\pi)^y$

Solution 1.6

1. $1-\pi$
3. $\mathsf{Geom1}\,(\pi)$

Solution 1.7

1. The exponential distribution has PDF

$$f_Y(y) = \lambda e^{-\lambda y} \quad (0 < y < \infty;\, \lambda > 0).$$

Therefore, from Definition 1.1,

$$E(Y) = \int_0^\infty y\lambda e^{-\lambda y}\, dy.$$

We can carry out the integration in several ways.

Integration by parts gives

$$\begin{aligned} E(Y) &= \int_0^\infty y\lambda e^{-\lambda y}\, dy = \int_0^\infty (-y)\frac{de^{-\lambda y}}{dy}\, dy = -ye^{-\lambda y}\big|_0^\infty - \int_0^\infty (-1)e^{-\lambda y}\, dy \\ &= (0-0) + \int_0^\infty e^{-\lambda y}\, dy = -\frac{1}{\lambda}e^{-\lambda y}\bigg|_0^\infty = 0 - \left(-\frac{1}{\lambda}\right) = \frac{1}{\lambda}. \end{aligned}$$

Alternatively,

$$\begin{aligned} E(Y) &= \int_0^\infty y\lambda e^{-\lambda y}\, dy = -\lambda \int_0^\infty \frac{de^{-\lambda y}}{d\lambda}\, dy = -\lambda \frac{d}{d\lambda}\int_0^\infty e^{-\lambda y}\, dy \\ &= -\lambda \frac{d}{d\lambda}\left(-\frac{1}{\lambda}e^{-\lambda y}\bigg|_0^\infty\right) = -\lambda \frac{d}{d\lambda}\left(-\frac{1}{\lambda}(0-1)\right) = -\lambda \frac{d}{d\lambda}\frac{1}{\lambda} = -\lambda\left(-\frac{1}{\lambda^2}\right) \\ &= \frac{1}{\lambda}. \end{aligned}$$

2. Again using integration by parts,

$$\begin{aligned} E(Y^2) &= \int_0^\infty y^2\lambda e^{-\lambda y}\, dy = \int_0^\infty (-y^2)\frac{de^{-\lambda y}}{dy}\, dy \\ &= -y^2e^{-\lambda y}\big|_0^\infty - \int_0^\infty (-2y)e^{-\lambda y}\, dy = (0-0) + 2\int_0^\infty ye^{-\lambda y}\, dy \\ &= \frac{2}{\lambda}\int_0^\infty y\lambda e^{-\lambda y}\, dy = \frac{2}{\lambda}\mathrm{E}(Y) = \frac{2}{\lambda}\frac{1}{\lambda} = \frac{2}{\lambda^2}, \end{aligned}$$

with $\mathrm{E}(Y)$ taken from part 1.

Hence,

$$\mathrm{Var}(Y) = \mathrm{E}(Y^2) - (\mathrm{E}(Y))^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}.$$

Solution 1.8

1. Think of the integrand in $\Gamma(1)$ as the PDF of one of the continuous distributions in Table 1.4, which must integrate to 1.

2. Substitute $x^2 = 2y$ in the integrand and introduce further constants to make the integrand one of the continuous distributions in Table 1.4.

Solution 1.9

The transformation here is $Z = g(Y) = 1/Y$. Hence $Y = g^{-1}(Z) = 1/Z$, and

$$\frac{dg^{-1}(z)}{dz} = -\frac{1}{z^2}.$$

We also know that Y has PDF

$$f_Y(y) = \frac{1}{\Gamma(\nu)}\lambda(\lambda y)^{\nu-1}e^{-\lambda y} \quad (0 < y < \infty; \nu > 0; \lambda > 0).$$

Applying the result in (1.3),

$$\begin{aligned} f_Z(z) &= \left|\frac{dg^{-1}(z)}{dz}\right| f_Y(g^{-1}(z)) = \frac{1}{z^2}f_Y(1/z) = \frac{1}{z^2}\frac{1}{\Gamma(\nu)}\lambda\left(\frac{\lambda}{z}\right)^{\nu-1}e^{-\lambda/z} \\ &= \frac{1}{\Gamma(\nu)}\frac{1}{z}\left(\frac{\lambda}{z}\right)^{\nu}e^{-\lambda/z} \quad (0 < y < \infty; \nu > 0; \lambda > 0). \end{aligned}$$

Solution 1.10

1. $e^{-\lambda y}$ from the CDF in Example 1.4

2. $e^{-\lambda y}$ for $y > 0$

3. $\mathsf{Expon}\,(\lambda)$

Solution 1.11

Because of the symmetry of the normal PDF around μ,

$$\Pr(Z < \mu) = 0.5.$$

As $\exp(\cdot)$ is a monotonically increasing transformation,

$$\Pr(e^Z < e^\mu) = 0.5,$$

and hence from the definition of the median, $Y = e^Z$ has median e^μ.

Solution 1.12

1. 0.7 and 0.6, respectively
2. 0.21 and 0.24, respectively
3. 0.18
4. $\mathrm{E}(B) = 1.3$ and $\mathrm{Var}(B) = 0.81$

Solution 1.13

As $Y = \min(X, k)$ is a function of the random variable X, compute the expectation of a function of a random variable:

$$E(Y) = \int_0^\infty \min(x, k)\lambda e^{-\lambda x}\, dx.$$

Then break the integral into two parts:

$$E(Y) = \int_0^k x\lambda e^{-\lambda x}\, dx + \int_k^\infty k\lambda e^{-\lambda x}\, dx = \lambda \int_0^k x e^{-\lambda x}\, dx + k \int_k^\infty \lambda e^{-\lambda x}\, dx. \quad (10.1)$$

Using the result

$$\int x e^{ax}\, dx = e^{ax}\left(\frac{x}{a} - \frac{1}{a^2}\right),$$

with $a = -\lambda$, the first integral in (10.1) is

$$\int_0^k x\lambda e^{-\lambda x}\, dx = e^{-\lambda x}\left(\frac{x}{-\lambda} - \frac{1}{\lambda^2}\right)\Bigg|_0^k = e^{-\lambda k}\left(\frac{k}{-\lambda} - \frac{1}{\lambda^2}\right) + \frac{1}{\lambda^2}.$$

The second integral in (10.1) is immediate from the exponential distribution's survival function:

$$\int_k^\infty \lambda e^{-\lambda x} = \Pr(X > k) = e^{-\lambda k}.$$

Putting these results together,

$$E(Y) = \lambda\left(e^{-\lambda k}\left(\frac{k}{-\lambda} - \frac{1}{\lambda^2}\right) + \frac{1}{\lambda^2}\right) + k e^{-\lambda k} = -\frac{e^{-\lambda k}}{\lambda} + \frac{1}{\lambda} = \frac{1 - e^{-\lambda k}}{\lambda}.$$

Solution 1.14

3. Use the two previous results rather than the definition of expectation.
4. The proof should use $f_{X,Y}(x, y)$, the joint PDF of X and Y, without making any assumptions about the joint distribution such as independence.
5. Use the previous results rather than the definition of expectation.

Solution 1.15

There is no need to manipulate integrals or sums. Instead work with expectations and use the results of Exercise 1.14.

Solution 1.16

3. Use the two previous results rather than the definition of variance.

5. Use the previous results and the result of Exercise 1.15 rather than the definition of variance.

Solution 1.19

1. Using the result for the expectation of a linear combination of random variables,

$$\mathrm{E}(Y) = \mathrm{E}(B_1 + \cdots + B_n) = \mathrm{E}(B_1) + \cdots + \mathrm{E}(B_n) = n\pi.$$

2. Using the result for the variance of a linear combination of random variables,

$$\mathrm{Var}(Y) = \mathrm{Var}(B_1 + \cdots + B_n) = \mathrm{Var}(B_1) + \cdots + \mathrm{Var}(B_n) = n\pi(1-\pi),$$

because the B_i are independent and hence all covariance terms are zero.

3. Using the result for the MGF of a sum of independent random variables,

$$M_Y(t) = \prod_{i=1}^{n} M_{B_i}(t) = (1 - \pi + \pi e^t)^n.$$

4. $\mathsf{Bin}\,(n, \pi)$, because the MGF identifies the distribution.

Solution 1.20

2. $\exp\left(\sum_{i=1}^{n} \mu_i(e^t - 1)\right)$

3. Poisson. Be sure to give the value of the Poisson parameter.

Solution 1.22

4. $\mathsf{NegBin}\,(n, \pi)$

Solution 1.23

1. Follow the steps of Example 1.24.

2. Note that Y is a linear function of Z.

   (a) Use the results on expectation and variance of a linear function of a random variable.

   (b) Apply Lemma 1.2.

Solution 1.25

1. $\lambda/(\lambda - bt)$
2. Expon (λ/b)

Solution 1.26

1. $\mathrm{E}(Y)$
2. $\mathrm{E}(Y) = e^{\mu+\sigma^2/2}$; note that e^{μ} is the median of Y (Exercise 1.11).

Solution 2.1

Write Z as a linear function of Y, i.e., $Z = a + bY$, then apply the method of Example 1.30 to establish the distribution of Z.

Solution 2.2

1. The MGF of Y_i is $\exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right) \quad (-\infty < t < +\infty)$.
2. Since we have a sum of independent random variables Y_i here, we can use the result on the MGF of a sum of independent random variables:

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

Then, substituting the MGF from part 1:

$$M_X(t) = \prod_{i=1}^{n} M_{Y_i}(t) = \prod_{i=1}^{n} \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right) = \exp\left(n\mu t + \frac{1}{2}n\sigma^2 t^2\right)$$
$$(-\infty < t < +\infty).$$

3. In general, if the MGF of Y is $M_Y(t)$, then $Z = a + bY$ has MGF $M_Z(t) = \exp(at)M_Y(bt)$. As $\bar{Y} = X/n$, and we already have $M_X(t)$ from part 1, we have

$$M_{\bar{Y}}(t) = M_X(t/n) = \exp\left(\mu t + \frac{1}{2}\frac{\sigma^2}{n}t^2\right) \quad (-\infty < t < +\infty).$$

4. Apply the general result in part 3 again, this time to the linear function

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} = \frac{-\mu\sqrt{n}}{\sigma} + \frac{\sqrt{n}}{\sigma}\bar{Y}.$$

From $M_{\bar{Y}}(t)$ in part 3, we have

$$\begin{aligned} M_Z(t) &= \exp\left(\frac{-\mu\sqrt{n}}{\sigma}t\right) M_{\bar{Y}}\left(\frac{\sqrt{n}}{\sigma}t\right) \\ &= \exp\left(\frac{-\mu\sqrt{n}}{\sigma}t\right) \exp\left(\frac{\mu\sqrt{n}}{\sigma}t + \frac{1}{2}\frac{\sigma^2}{n}\frac{nt^2}{\sigma^2}\right) = \exp(t^2/2) \\ &\quad (-\infty < t < +\infty). \end{aligned}$$

This is the MGF of a normal distribution (see part 1) with $\mu = 0$ and $\sigma^2 = 1$, and the MGF uniquely identifies a distribution. Thus, Z has a standard normal distribution.

Solution 2.3

The χ^2_1 PDF is obtained by putting $d = 1$ in the χ^2_d PDF given in Table 1.4. Rearrange the integral in the definition of the MGF so that it includes the integral of a gamma PDF.

Solution 2.5

The χ^2_d PDF is given in Table 1.4. Rearrange the integral so that it includes the integral of a gamma PDF.

Solution 2.6

2. Do not start with an assumed PDF for Y, or you will have a circular argument in part 3. No integration is required.

3. Find this MGF in Table 1.4 or use the result of Exercise 2.5.

4. d

5. $2d$

Solution 2.7

Use the result of Exercise 2.6.

Solution 2.8

2. Yes

Solution 2.10

1. The widths are 8.53 percentage points for 90% confidence and 13.78 percentage points for 99% confidence.

3. The widths are 8.29 percentage points for 90% confidence and 12.99 percentage points for 99% confidence.

4. About 2.9% wider for 90% confidence and about 6.1% wider for 99% confidence.

Solution 2.11

2. (c) $[92.0, 100.2]$

Solution 2.12

4. 1 and 2, respectively
5. σ^2
6. $2\sigma^4$
7. $l = 0.000982$ and $u = 5.02$

Solution 2.13

1. 1
2. $2/(n-1)$
3. (a) 1
   (b) Find a relevant theorem or lemma and describe briefly how it applies to the random variable S^2/σ^2.
4. Consider the properties of S^2/σ^2 as $n \to \infty$.
7. Don't forget that parts 1 and 2 relate to the distribution of S^2 divided by σ^2.

Solution 2.14

Use the fact that a random variable with a χ^2_d distribution arises as the sum of squares of d independent standard normal random variables, and apply the CLT.

Solution 2.16

1. (a) 0
   (b) 1/3
   (c) $(e^t - e^{-t})/(2t)$
   (d) 0 and 1/3
2. (a) 0
   (b) $n/3$
   (c) $\left(\frac{e^t - e^{-t}}{2t}\right)^n$
3. (a) $Z = \sqrt{\frac{3}{n}}Y$
   (b) $\left(\frac{\exp\left(t\sqrt{3/n}\right) - \exp\left(-t\sqrt{3/n}\right)}{2t\sqrt{3/n}}\right)^n$
   (c) $e^{t^2/2}$
   (d) Standard normal

Solution 2.17

Find an example where the normal approximation to the binomial is used. Argue that the use is compatible with Definition 2.1 on convergence in distribution. In particular, does Definition 2.1 mention a PMF or a PDF?

Solution 2.18

3. Less than or equal to 0.26

5. 0.05

6. (a) $\mathsf{N}(\mu, \sigma^2/n)$
   (b) 0.05
   (c) No, it is an exact property of the normal distribution.

Solution 3.2

1. Applying the result for the expectation of a linear combination of random variables to $\bar{Y}$ gives $\mathrm{E}(\tilde{\mu}) = \mu$.

2. Applying the result for the variance of a linear combination of independent random variables to $\bar{Y}$ gives $\mathrm{Var}(\tilde{\mu}) = \mu/n$.

3. Yes, it is unbiased, and its variance goes to zero as $n \to \infty$.

4. (a) Using the result for the expectation of a linear function of a random variable,
$$\mathrm{E}(\widetilde{\mathrm{Var}}(\tilde{\mu})) = \mathrm{E}(\tilde{\mu}/n) = \mathrm{E}(\tilde{\mu})/n = \mu/n = \mathrm{Var}(\tilde{\mu}).$$
   (b) Using the result for the variance of a linear function of a random variable,
$$\mathrm{Var}(\widetilde{\mathrm{Var}}(\tilde{\mu})) = \mathrm{Var}(\tilde{\mu}/n) = \mathrm{Var}(\tilde{\mu})/n^2 = \mu/n/n^2 = \mu/n^3.$$
   (c) Yes, it is unbiased, and its variance goes to zero as $n \to \infty$.

Solution 3.3

Yes. Use the results of Exercise 2.13.

Solution 3.4

3. The bias is $-\sigma^2/n$. Hence,
$$\mathrm{MSE}\left(\widetilde{\sigma^2}\right) = \mathrm{Bias}^2\left(\widetilde{\sigma^2}\right) + \mathrm{Var}\left(\widetilde{\sigma^2}\right) = \left(-\frac{\sigma^2}{n}\right)^2 + \frac{2(n-1)\sigma^4}{n^2} = \frac{(2n-1)\sigma^4}{n^2}.$$

4. Yes. The MSE goes to zero as $n \to \infty$.

5. S^2 is an unbiased estimator of σ^2, because $\mathrm{E}(S^2) = \sigma^2$, whereas $\widetilde{\sigma^2}$ is biased. On the other hand, $\mathrm{Var}(S^2) = 2\sigma^4/(n-1)$ is greater than

$$\mathrm{Var}\left(\widetilde{\sigma^2}\right) = \frac{2(n-1)\sigma^4}{n^2} = 2\sigma^4\left(\frac{1}{n} - \frac{1}{n^2}\right)$$

for all $n \geq 2$.

6. Compare the two MSEs:

$$\mathrm{MSE}(S^2) = \mathrm{Var}(S^2) = \frac{2}{n-1}\sigma^4$$

and

$$\begin{aligned}\mathrm{MSE}\left(\widetilde{\sigma^2}\right) &= \frac{2n-1}{n^2}\sigma^4 \quad \text{(from part 3)} \\ &= \left(\frac{2}{n} - \frac{1}{n^2}\right)\sigma^4.\end{aligned}$$

Clearly, $\mathrm{MSE}\left(\widetilde{\sigma^2}\right) < \mathrm{MSE}(S^2)$ for all $n \geq 2$, and $\widetilde{\sigma^2}$ is more accurate in the sense of MSE.

Solution 3.5

1. (a) Using the result on the expectation of a linear combination of random variables,

$$\begin{aligned}\mathrm{E}(\tilde{\mu}) = \mathrm{E}\left(\frac{Y_1 + Y_2 + \cdots + Y_n}{n}\right) &= \frac{1}{n}\left(\mathrm{E}(Y_1) + \mathrm{E}(Y_2) + \cdots + \mathrm{E}(Y_n)\right) \\ = \frac{1}{n}(n\mu) &= \mu.\end{aligned}$$

Hence, $\tilde{\mu} = \bar{Y}$ is an unbiased estimator of μ.

(b) Using the result on the variance of a linear combination of random variables and noting that all $\mathrm{Cov}(Y_i, Y_j)$ terms are zero because the Y_i are assumed independent,

$$\begin{aligned}\mathrm{Var}(\tilde{\mu}) &= \mathrm{Var}\left(\frac{Y_1 + Y_2 + \cdots + Y_n}{n}\right) \\ &= \frac{1}{n^2}\left(\mathrm{Var}(Y_1) + \mathrm{Var}(Y_2) + \cdots + \mathrm{Var}(Y_n)\right) \\ &= \frac{1}{n^2}\left(2n\phi^2\right) = \frac{2\phi^2}{n}.\end{aligned}$$

Therefore,

$$\mathrm{sd}(\tilde{\mu}) = \sqrt{\frac{2\phi^2}{n}}.$$

For the given numbers,

$$\mathrm{sd}(\tilde{\mu}) = \sqrt{\frac{2\phi^2}{n}} = \sqrt{\frac{200}{25}} = 2.828.$$

2. (b) One execution of the code produces 1.012 for the sample mean of the 10 000 sample medians. You will obtain a slightly different value from your (different) random samples. An unbiased estimator has an expected value of 1. Based on the simulation of the sampling distribution, the estimate of the expected value, 1.012, is close to 1. Hence, it appears that the bias is very small and could be zero.

   (c) An estimate of the true standard deviation of the sample median is 2.31. Again your estimate will be slightly different.

3. The estimate of the true standard deviation of the sample mean is 2.82, which agrees well with the theoretical calculation.

4. The sample mean is unbiased and the sample median appears to be unbiased or have negligible bias. However, the sample median has a much smaller standard deviation in the simulation. Hence, the sample median appears to be more accurate.

Solution 4.1

1. $-40\mu + 10\ln(\mu)$

4. 0.25

5. 0.0791

6. 0.25 ± 0.155

7. 0.38. Adapt the argument in Section 4.6. In particular, the tail probability $\alpha = 0.05$ will be entirely in one tail, instead of divided as it was in Figure 4.9.

8. 0.68

9. 0 faults has an expected frequency of 31.2, etc; yes

Solution 4.2

1. The likelihood function is

$$f_{Y_1,\ldots,Y_{298}}(y_1,\ldots,y_{298} \mid \mu) = \prod_{i=1}^{298} f_{Y_i}(y_i \mid \mu) = \prod_{i=1}^{298} \frac{e^{-\mu}\mu^{y_i}}{y_i!} = \frac{e^{-298\mu}\mu^{\sum_{i=1}^{298} y_i}}{\prod_{i=1}^{298} y_i!},$$

   since $Y_1, \ldots, Y_{298}$ are independent random variables.

2. The likelihood function is viewed as a function of μ.

3. The log likelihood is

$$\ln f_{Y_1,\ldots,Y_{298}}(y_1,\ldots,y_{298} \mid \mu) = -298\mu + \left(\sum_{i=1}^{298} y_i\right)\ln(\mu) + c,$$