Statistical Inference: A Primer on Likelihood and Bayesian Methods

William J. Welch

Version: August 14, 2019

Chapter 1

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

Copyright

Contents

Chapter 2

Probability Tools

2.1 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.

2.2 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.

2.2.1 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.

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 $\text{Pois}(\mu)$ 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 $1 + \mu + \frac{\mu^2}{2!} + \cdots$ for e^{μ}.

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. The density function is positive and integrates to 1 over the range of possible y values.

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 $\text{Expon}(\lambda)$, depends on the parameter $\lambda > 0$. The exponential PDF integrates to 1, as required:

$$\int_0^{\infty} f_Y(y)\, dy = \int_0^{\infty} \lambda e^{-\lambda y}\, dy = -e^{-\lambda y}\Big|_0^{\infty} = 0 - (-1) = 1.$$

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.

2.3 DISCRETE AND CONTINUOUS RANDOM VARIABLES

2.3.1 Example 1.3 (Normal distribution: symmetry)

A random variable with a normal distribution has PDF

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

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

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

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

2.3.2 Cumulative distribution function

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

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

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 $\Pr(Y \le y)$ 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

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

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.

2.3.3 Example 1.4 (Exponential distribution: CDF)

Let Y have an Expon(λ) distribution. From the definition of the CDF,

$$F_Y(y) = \int_0^y f_Y(t)dt = 1 - e^{-\lambda y},$$

where $f_Y(t) = \lambda e^{-\lambda t}$ for $t > 0$. (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:

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

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.

2.4 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".

2.4.1 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 $f_Y(y)$.

The expected value or mean of a random variable Y is given by the sum

$$E(Y) = \sum_y y f_Y(y)$$

if Y takes discrete values, or by the integral

$$E(Y) = \int y f_Y(y) dy$$

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

2.4.2 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

$$E(Y) = 60(0.2) + 90(0.8) = 84,$$

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.

2.4.3 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 Unif(a, b). From Definition 1.1, the expectation or mean of Y is

$$E(Y) = \int_a^b y f_Y(y) dy = \frac{a+b}{2}.$$

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

$$\sum |y| f_Y(y) < \infty \quad \text{(discrete)} \quad \text{or} \quad \int |y| f_Y(y) dy < \infty \quad \text{(continuous)}.$$

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

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

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 $\frac{1}{y!}$, 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

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

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

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

however, the sum does not converge ($\sum_{y=1}^{\infty} \frac{1}{y}$ 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.

2.5 Variances

2.5.1 Computation

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

The variance of Y is

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

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

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

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:

$$\sum \text{Var}(Y) = E((Y - E(Y))^2) = \sum (y - \mu)^2 f_Y(y).$$

Alternatively,

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

where $\mu = E(Y)$.

2.5.2 Example 1.10 (Final-exam grade: variance)

For the distribution $f_Y(y)$ of final grades in Table 1.1, we already computed $E(Y) = \mu = 84$. Hence,

$$\text{Var}(Y) = E((Y - \mu)^2) = (60 - 84)^2(0.2) + (90 - 84)^2(0.8) = 144.$$

Alternatively, using the second definition of variance,

$$\text{Var}(Y) = E(Y^2) - \mu^2 = (60)^2(0.2) + (90)^2(0.8) - (84)^2 = 144.$$

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:

$$E(Y-\mu)^2 = \sum (y-\mu)^2 f_Y(y) = \sum (y^2 - 2\mu y + \mu^2) f_Y(y) = \sum y^2 f_Y(y) - 2\mu \sum y f_Y(y) + \mu^2 = E(Y^2) - 2\mu E$$

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

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

Alternatively,

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

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

2.5.3 Example 1.11 (Uniform distribution: variance)

Let Y have a 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

$$E(Y) = \frac{a+b}{2}.$$

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

$$E(Y^2) = \int_a^b y^2 f_Y(y)dy = \int_a^b y^2 \frac{1}{b-a} dy.$$

Thus, the final variance is computed as

$$\mathrm{Var}(Y) = \frac{(b-a)^2}{12}.$$

2.5.4 Standard deviation

The standard deviation, often denoted by σ, is

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

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.

2.5.5 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. 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.

2.6 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).

The distributions in Table 1.3 are now briefly described.

2.6.1 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

$$f_B(b) = \pi^b(1 - \pi)^{1-b}$$

for $b = 0, 1$; $0 < \pi < 1$. The Bernoulli distribution Bern(π) 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.)

2.6.2 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, $Y \sim \text{Bin}(n, \pi)$ is $Y = \sum_{i=1}^{n} B_i$, where the B_i are independent Bern(π). The binomial distribution is perhaps the most important discrete distribution, because Y/n is the sample proportion, of interest in numerous applications.

2.6.3 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.

2.6.4 Negative-binomial distribution

A negative-binomial random variable $Y \sim \text{NegBin}(n, \pi)$ arises as the sum of n independent $\text{Geom1}(\pi)$ random variables. Thus, it counts the number of Bernoulli trials until n successes have occurred.

2.6.5 Poisson distribution

A Poisson random variable can be thought of as a limiting case of the binomial. If $Y \sim \text{Bin}(n, \pi)$, and we take the limits $n \to \infty$ and $\pi \to 0$ such that $\mu = n\pi$ tends to a constant, then $Y \sim \text{Pois}(\mu)$ is the limiting distribution.

© Copyright William J. Welch 2009–2019. All rights reserved. Not to be copied, used, or revised without explicit written permission from the copyright owner. 2019.8.14

1.11. EXERCISES

1-49

Exercise 1.13
For a specific type of property insurance claim, an actuary models the customer's loss by the random variable $X \sim \text{Expon}(\lambda)$. 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 $E(Y)$?
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. $E(a + X) = a + E(X)$.
2. $E(bX) = bE(X)$.
3. $E(a + bX) = a + bE(X)$.
4. $E(X + Y) = E(X) + E(Y)$.
5. $E(a + bX + cY) = a + bE(X) + cE(Y)$.
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:

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

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., $\text{Var}(X) = E(X^2) - (E(X))^2$, prove the following results. (Hint: The definition of variance is in terms of expectations; use the results of Exercise 1.14.)

1. $\text{Var}(a + X) = \text{Var}(X)$.
2. $\text{Var}(bX) = b^2\text{Var}(X)$.
3. $\text{Var}(a + bX) = b^2\text{Var}(X)$.
4. $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$.
5. $\text{Var}(a + bX + cY) = b^2\text{Var}(X) + c^2\text{Var}(Y) + 2bc\text{Cov}(X, Y)$.

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 $E(Y_i) = \pi$ and $\text{Var}(Y_i) = \pi(1-\pi)$.
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 $E(\tilde{\pi}) = \pi$, i.e., $\tilde{\pi}$ is an unbiased estimator of π.
   (b) Show that $\text{Var}(\tilde{\pi}) = \frac{\pi(1-\pi)}{n}$.

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, $B \sim \text{Bern}(\pi)$. 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 $E(B) = \pi$.
2. Find $E(10B)$.
3. Show $\text{Var}(B) = \pi(1-\pi)$.
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 $E(B)$.

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 Bern(π) distribution. You may use the results in Exercise 1.18 that B_i has mean π, variance $\pi(1-\pi)$, and MGF $M_B(t) = 1 - \pi + \pi e^t$. Also n is some fixed number.

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

Exercise 1.20
Let $Y \sim \text{Pois}(\mu)$. 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., $Y_i \sim \text{Pois}(\mu_i)$. 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 $Y \sim \text{Unif}(a, b)$. Use the expansion of its MGF in Example 1.29 to show the following properties:

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

Exercise 1.22
Let $Y \sim \text{Geom1}(\pi)$. 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
$$E(Y) = \frac{1}{\pi}$$
and
$$\text{Var}(Y) = \frac{1 - \pi}{\pi^2}.$$
3. Let $Y_1, \ldots, Y_n$ be IID $\text{Geom1}(\pi)$ random variables.
4. Hence, what is the distribution of $\sum_{i=1}^n Y_i$? What is the MGF of $\sum_{i=1}^n Y_i$?

Exercise 1.23
Let $Y \sim N(\mu, \sigma^2)$, 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 $N(\mu, \sigma^2)$ PDF to find the MGF of Y.

2. Let Z have a standard normal distribution, i.e., $N(0, 1)$. Its MGF is

   $$M_Z(t) = e^{\frac{1}{2}t^2} \quad \text{(see Example 1.24).}$$

   Now let $Y = \mu + \sigma Z$.

   (a) Verify that $E(Y) = \mu$ and $\text{Var}(Y) = \sigma^2$ as required.
   (b) Find the MGF of Y from the MGF of Z.

Exercise 1.24
Let $Y \sim N(\mu, \sigma^2)$. 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 $E(Y) = \mu$.

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

Exercise 1.25
Let $Y \sim \text{Expon}(\lambda)$. 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 $\frac{\lambda}{\lambda - t}$. 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 $Y \sim \text{Gamma}(\nu, \lambda)$, then $Z = bY \sim \text{Gamma}(\nu, \lambda/b)$.

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

1. The definition of the MGF of Z is $E(e^{tZ})$. 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 $E(Y)$?

2-1
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 $N(\mu, \sigma^2)$ 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 $Y \sim N(\mu, \sigma^2)$. The following properties were established using MGFs in Section 1.8 or in related exercises.

- The expected value (mean) of Y is $E(Y) = \mu$, and the variance is $\text{Var}(Y) = \sigma^2$ (Exercise 1.24).

- A linear function of Y is also normal: $a + bY \sim N(a + b\mu, b^2\sigma^2)$ (Example 1.30). In particular,

$$Z = \frac{Y - \mu}{\sigma} \sim N(0, 1$$

  has the standard normal distribution $N(0, 1)$, 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 $\text{Cov}(Y, Z) = 0$, 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.

Figure 2.1: Relationships between distributions derived from the normal

The χ^2 distribution

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

2.7 SEVERAL VARIABLES

2.7.1 Example 1.21 (Gamma distribution: mean and variance)

The gamma distribution (see Table 1.4) has PDF

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

for $0 < y < \infty; \nu > 0; \lambda > 0$,

which we write as Gamma (ν, λ). 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 Gamma (ν, λ) random variable Y can be generated by:

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

where the Y_i are independent Expon(λ) random variables. (This result will be proved in Example 1.31.) We know Y_i has mean $\frac{1}{\lambda}$ and variance $\frac{1}{\lambda^2}$. Hence, it immediately follows that a Gamma (ν, λ) random variable has mean $\frac{\nu}{\lambda}$ (from (1.9)) and variance $\frac{\nu}{\lambda^2}$ (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 $\text{Cov}(Y_i, Y_j)$ 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.

2.7.2 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.

2.7.3 Chapter 1. Probability Tools

Similarly, for linear combinations of random variables,

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

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

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

2.7.4 Bivariate normal distribution

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

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

where

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

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 $\det(\Sigma)$ and Σ^{-1} denote matrix determinant and inverse of Σ, respectively. The off-diagonal element $\rho\sigma_1\sigma_2$ in the covariance matrix Σ 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.

2.7.5 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

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

and hence $\det^2(\Sigma) = \sigma_1 \sigma_2$. Second,

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

Substituting these two results into the joint distribution gives

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

This is the product of a $N(\mu_1, \sigma_1^2)$ PDF for Y_1 and a $N(\mu_2, \sigma_2^2)$ 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.

2.7.6 Moment Generating Functions

2.7.7 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).

2.7.8 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.

(Moments of a random variable) Let Y be a random variable. Its kth moment for $k = 1, 2, \ldots$ is $E(Y^k)$, which exists if the expectation is finite. Thus, the first moment with $k = 1$ is simply $E(Y)$. The first two moments, $E(Y^1)$ and $E(Y^2)$, 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.

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

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

if it exists for t in a neighbourhood of 0, i.e., for t in the open interval $(-T, T)$, 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 $t = 0$.

2.7.9 Example 1.22 (Exponential distribution: MGF)

Let Y be distributed Expon(λ). As this is a continuous random variable, we compute the expectation in the MGF via integration:

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

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

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

The integration to find the MGF of the Expon (λ) distribution is

$$M_Y(t) = \frac{\lambda}{\lambda - t}.$$

3-10

Chapter 3

STATISTICAL ESTIMATION

A standard answer is that σ_e^2 is biased, whereas S^2 is not (Exercise 2.9). The bias turns out to be small relative to $sd(\sigma_e^2)$, however, in the important special case that the Y_i are also normally distributed. Then, via the arguments of Section 2.4.2,

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

and

$$Var(X) = 2(n-1)\sigma^4,$$

whereupon

$$Bias(\sigma_e^2) = E(\sigma_e^2) - \sigma^2 = \frac{(n-1)\sigma^2 - \sigma^2}{n} = -\frac{\sigma^2}{n}$$

and

$$Var(X) = \frac{1}{n^2} \cdot \frac{2(n-1)\sigma^4}{n^2}.$$

We see that for any $n \geq 2$, the bias of σ_e^2 is smaller in magnitude than its standard deviation:

$$\frac{Bias(\sigma_e^2)}{sd(\sigma_e^2)} = \frac{-\sigma^2/n}{\sqrt{2(n-1)\sigma^2/n^2}} = -\sqrt{\frac{1}{2(n-1)}}.$$

Of course, no bias is better than a "small" bias, but Exercise 3.4 shows that σ_e^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 sources 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.

3.1 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

$$Pr(|\tilde{\theta}_n - \theta| < \epsilon) \to 1 \text{ as } n \to \infty$$

for any fixed error, $\epsilon > 0$.

3.3. PROPERTIES OF AN ESTIMATOR

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 $\epsilon > 0$,

$$Pr(|\bar{Y}_n - \mu| < \epsilon) \to 1 \text{ as } n \to \infty.$$

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

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

and similarly, we use (1.10) to get $Var(\bar{Y}_n)$. Because we are assuming the Y_i are independent, all the covariance terms, $Cov(Y_i, Y_j)$ for $i \neq j$, in (1.10) are zero. Thus,

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

Clearly, $Var(\bar{Y}_n) \to 0$ 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.

3-12

CHAPTER 3. STATISTICAL ESTIMATION

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., $Y_i \sim \text{Bern}(\pi)$. Note that $E(Y_i) = \pi$ and $Var(Y_i) = \pi(1-\pi)$. 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,

$$E(\bar{Y}_n) = \pi,$$

and

$$Var(\bar{Y}_n) = \frac{\pi(1-\pi)}{n}.$$

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.

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,

$$Var(\tilde{\theta}) = E\left[(\tilde{\theta} - \theta)^2\right],$$

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,

$$\frac{\tilde{\theta} - \theta}{\theta} = \frac{\tilde{\theta}}{\theta} - 1,$$

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 $\text{Bin}(n, \pi)$ probability model. Suppose the requirement is to determine n such that $sd(\tilde{\pi}) \leq 0.015$. 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

$$sd(\tilde{\pi}) = \sqrt{\frac{\pi(1-\pi)}{n}} \leq 0.015$$

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 $sd(\tilde{\pi}) \leq 0.015$ for any π, and a 95% confidence interval using the standard error and a normal approximation will be no wider than $\pm z_{0.975} \cdot se(\tilde{\pi}) = \pm 1.96 \cdot 0.015 = \pm 0.0294$, 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 argument 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 $sd(\tilde{\pi}) \leq 0.015$ 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., $sd(\tilde{\pi}) \leq 0.03\pi$. Applying the rule

for the variance (hence standard deviation) of a linear function of a random variable, the new requirement is equivalent to

$$sd\left(\frac{\tilde{\pi}-\pi}{\pi}\right) = 0.03.$$

Solving for n yields

$$n = \frac{\pi(1-\pi)}{(0.03)^2},$$

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 $n_{\text{rel}} = 21112$, increasing to 110000 at $\pi = 0.01$. Relaxing the requirement so that $sd(\tilde{\pi}) \leq 0.1\pi$ when $\pi = 0.01 still needs \text{n}_{\text{rel}} = 9900$. Estimating a small probability with good relative accuracy requires a huge sample size.

□□□

3.2 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.

3.3 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.

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 $Var(\tilde{\mu})$ (an exact formula for the variance).
3. Is $\tilde{\mu}$ a consistent estimator of μ?
4. Consider $Var(\tilde{\mu})$ as an estimator of $Var(\tilde{\mu})$.
   (a) Show that $Var(\tilde{\mu})$ is an unbiased estimator of $Var(\tilde{\mu})$.
   (b) What is the variance of $Var(\tilde{\mu})$?
   (c) Is $Var(\tilde{\mu})$ a consistent estimator of $Var(\tilde{\mu})$?

Exercise 3.3 Let $Y_1, \ldots, Y_n$ be independent $N(\mu, \sigma^2)$ 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 $\sigma_e^2 = \frac{X}{n}$, 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 σ_e^2 is $\frac{\sigma^2(n-1)}{n}$.
2. Show that the variance of σ_e^2 is $\frac{2(n-1)\sigma^4}{n^2}$.
3. Hence, give and simplify an expression that summarizes the accuracy of σ_e^2 as an estimator of σ^2.
4. Is σ_e^2 a consistent estimator of σ^2? Explain briefly.
5. The sample variance with divisor $n-1$, namely $S^2 = \frac{X}{n-1}$, has $E(S^2) = \sigma^2$ and $Var(S^2) = \frac{2\sigma^4}{n-1}$. Give one advantage and one disadvantage of S^2 relative to σ_e^2 as an estimator of σ^2.
6. Explain which estimator of σ^2 is the more accurate, S^2 or σ_e^2.

3.7. EXERCISES

2.8. EXERCISES

2-23

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 $E(Y)$.

5. From the MGF find $\text{Var}(Y)$.

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

1. Show that the covariance between $\overline{Y}$ and $Y_i - \overline{Y}$ is zero.

2. Assume also that $Y_1, \ldots, Y_n$ are normally distributed. Are $\overline{Y}$ and $Y_i - \overline{Y}$ 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

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

1. Show that $E(X) = (n-1)\sigma^2$.

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:

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

where the units are percentage points. As before, we will assume that the $n = 34$ data values are a random sample from a $N(\mu, \sigma^2)$ 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

$$x \leftarrow seq(-4, 4, by = 0.01)$$

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 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:

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

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 $N(\mu, \sigma^2)$ 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 $N(\mu, \sigma^2)$ distribution. For such a sample of size $n = 2$, it is easily shown that the sample variance,

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})^2,$$

simplifies to

$$S^2 = \left(\frac{Y_1 - Y_2}{\sqrt{2}} \right)^2.$$

You may use this result without proof.

1. Let $V = \frac{(Y_1 - Y_2)}{\sqrt{2}}$. Show the following properties of V. Carefully state any result you are using (no proof of the result required). (a) $E(V) = 0$. (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 $\frac{S^2}{\sigma^2}$ is χ^2_1. (A result on the connection between $N(0,1)$ 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 $\frac{S^2}{\sigma^2}$ 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

$$P\left(\frac{S^2}{\sigma^2} < l\right) = 0.025 \text{ and } P\left(\frac{S^2}{\sigma^2} > u\right) = 0.025.$$

Hence, l and u are lower and upper bounds on $\frac{S^2}{\sigma^2}$ in the sense that

$$P(l < \frac{S^2}{\sigma^2} < u) = 0.95.$$

8. Suppose S^2 is used to estimate σ^2 from a sample of size $n = 2$. Comment on the values of $\frac{S^2}{\sigma^2}$ 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 $N(\mu, \sigma^2)$ random variables. Their sample variance is

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

This question explores the properties of $\frac{S^2}{\sigma^2}$ and why the t_{n-1} distribution approaches the standard normal as $n \to \infty$. Section 2.4.2 argued that $\frac{S^2}{\sigma^2}$ has the same distribution as $\frac{X}{n-1}$, where $X \sim \chi^2_{n-1}$. You may use this result and properties of the χ^2 distribution without proof.

1. Find $E\left(\frac{S^2}{\sigma^2}\right)$.

2. Find $\text{Var}\left(\frac{S^2}{\sigma^2}\right)$.

3. Let $\epsilon > 0$ be a fixed constant representing an arbitrarily small "error". (a) As $n \to \infty$, what is the limiting probability

$$P(1 - \epsilon < \frac{S^2}{\sigma^2} < 1 + \epsilon)?$$

(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):

$$\frac{\overline{Y} - \mu}{\sqrt{\frac{S^2}{n}}}$$

was shown to have a t_{n-1} distribution. As $n \to \infty$, it is known that the t_{n-1} distribution converges to $N(0,1)$. 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 $Unif(-1,1)$ random variables.

$$n \leftarrow 1000$$

$$x \leftarrow runif(n, min = -1, max = 1)$$

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 $Unif(-1,1)$ distribution. Why do the sample and theoretical quantities not agree exactly?

4. Generate a second, independent sample

$$y \leftarrow runif(n, min = -1, max = 1)$$

and then compute the sums

$$z[1] = x[1] + y[1], z[2] = x[2] + y[2], \ldots$$

using

$$z \leftarrow x + y$$

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 $Unif(-1, 1)$ 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 $U \sim Unif(-1, 1)$, 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 $E(U)$. (b) From the definition of variance, find $\text{Var}(U)$. (c) From the definition of the moment generating function, find $M_U(t)$. (d) From the moment generating function, find $E(U)$ and $\text{Var}(U)$. (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 $Unif(-1, 1)$ random variables. Consider the random variable

$$Y = \sum_{i=1}^{n} U_i.$$

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 $Bin(n, \pi)$ random variable. It was shown that

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

is approximately $N(0, 1)$ 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 $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ to estimate μ. 1. Show that $E(\overline{Y}) = \mu$. 2. Show that $\text{Var}(\overline{Y}) = \frac{\sigma^2}{n}$. 3. What does Chebyshev's inequality give for the probability $P(|\overline{Y} - \mu| > \epsilon)$, where $\epsilon = \frac{1.96\sigma}{\sqrt{n}}$? 4. What does the CLT say about the distribution of $\overline{Y}$ as $n \to \infty$? 5. Give an approximation based on the CLT to $P(|\overline{Y} - \mu| > \epsilon)$, where $\epsilon = \frac{1.96\sigma}{\sqrt{n}}$. Explain briefly. 6. Suppose $Y_1, \ldots, Y_n$ also have a normal distribution.

(a)

What is the distribution of $\overline{Y}$? Explain briefly. (b) What is $P(|\overline{Y} - \mu| > \epsilon)$, where $\epsilon = \frac{1.96\sigma}{\sqrt{n}}$? 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 $X_n \sim Bin(n, \pi)$. 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 $Pois(\mu = 0.35)$ distribution and plot the values using *hist*. Does the empirical distribution have a roughly normal shape? 2. Repeat part 1 but sample from a $Pois(\mu = 25)$ 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 $Y \sim Pois(\mu)$. The standardized variable $Z = \frac{(Y-\mu)}{\sqrt{\mu}}$ has mean 0 and variance 1. The MGF of Z can be written as

$$M_Z(t) = \exp\left(\frac{t^2}{2\mu} + O\left(\frac{1}{\mu}\right)\right).$$

The notation $O\left(\frac{1}{\mu}\right)$ 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 $\mu \to \infty$?

Appendix: Proof of Lemma 2.2

From Section 2.3.2 we already established that the PDF of $T = Z/\sqrt{X_d/d}$ is given by

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

where $W = X_d/d$. It was also argued that $f_{T|W}(t|w)$ is the $N(0, 1/w)$ PDF. Hence

$$f_{T|W}(t|w) = \frac{1}{\sqrt{2\pi w^2}} e^{-\frac{w}{2}t^2}$$

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

$$f_W(w) = \frac{d^{d/2}}{2^{d/2}\Gamma(d/2)} w^{d/2-1} e^{-dw/2}.$$

Combining the two PDFs, the required integral is

$$f_T(t) = \int_0^\infty \frac{1}{\sqrt{2\pi w^2}} e^{-\frac{w}{2}t^2} \cdot \frac{d^{d/2}}{2^{d/2}\Gamma(d/2)} w^{d/2-1} e^{-dw/2}\, dw.$$

Up to constants, the integrand is the gamma PDF in Table 1.4 with $\nu = (d+1)/2$ and $\lambda = (d+t)/2$. Inserting the required constants $\Gamma(\nu) = \Gamma(1/2)$ and $\lambda\nu = \frac{(d+t^2)}{2}(d+1)/2$ gives

$$f_T(t) = \frac{d^{d/2}}{2^{d/2}\Gamma(d/2)} \cdot \frac{1}{2\sqrt{2\pi}} \sqrt{w^{d+1-1} e^{-\frac{(d+t^2)}{2}}}.$$

In other words, this establishes the PDF of the t_d distribution in Table 1.4.