Statistical Theory

Lecture Notes

Adolfo J. Rumbos

Contents

Chapter 1

Introduction

1.1 Introduction to statistical inference

The main topic of this course is statistical inference. Loosely speaking, statistical inference is the process of going from information gained from a sample to inferences about a population from which the sample is taken. There are two aspects of statistical inference that we'll be studying in this course: estimation and hypothesis testing. In estimation, we try to determine parameters from a population based on quantities, referred to as statistics, calculated from data in a sample. The degree to which the estimates resemble the parameters being estimated can be measured by ascertaining the probability that a certain range of values around the estimate will contain the actual parameter. The use of probability is at the core of statistical inference; it involves the postulation of a certain probability model underlying the situation being studied and calculations based on that model. The same procedure can in turn be used to determine the degree to which the data in the sample support the underlying model; this is the essence of hypothesis testing.

Before we delve into the details of the statistical theory of estimation and hypothesis testing, we will present a simple example which will serve to illustrate several aspects of the theory.

1.1.1 An Introductory Example

I have a hot–air popcorn popper which I have been using a lot lately. It is a small appliance consisting of a metal, cylindrical container with narrow vents at the bottom, on the sides of the cylinder, through which hot air is pumped. The vents are slanted in a given direction so that the kernels are made to circulate at the bottom of the container. The top of the container is covered with a hard-plastic lid with a wide spout that directs popped and unpopped kernels to a container placed next to the popper. The instructions call for one–quarter cup of kernels to be placed at the bottom of the container and the device to be plugged in. After a short while of the kernels swirling in hot air, a few of the kernels begin to pop. Pressure from the circulating air and other kernels popping an bouncing off around inside the cylinder forces kernels to the top of the container, then to the spout, and finally into the container. Once you start eating the popcorn, you realize that not all the kernels popped. You also notice that there are two kinds of unpopped kernels: those that just didn't pop and those that were kicked out of the container before they could get warm enough to pop. In any case, after you are done eating the popped kernels, you cannot resit the temptation to count how many kernels did not pop. Table 1.1 shows the results of 27 popping sessions performed under nearly the same conditions. Each popping session represents a random experiment.[1] The

Trial	Number of Uppopped Kernels
1	32
2	11
3	32
4	9
5	17
6	8
7	7
8	15
9	139
10	110
11	124
12	111
13	67
14	143
15	35
16	52
17	35
18	65
19	44
20	52
21	49
22	18
23	56
24	131
25	55
26	59
27	37

Table 1.1: Number of Unpopped Kernels out of 1/4–cup of popcorn

[1] A *random experiment* is a process or observation, which can be repeated indefinitely under the same conditions, and whose outcomes cannot be predicted with certainty before the experiment is performed.

number of unpopped kernels is a random variable[2] which we obtain from the outcome of each experiment. Denoting the number of unpopped kernels in a given run by X, we may postulate that X follows a Binomial distribution with parameters N and p, where p is the probability that a given kernel will not pop (either because it was kicked out of the container too early, or because it would just not pop) and N is the number of kernels contained in one-quarter cup. We write

$$X \sim \text{binom}(N, p)$$

and have that

$$P(X = k) = \binom{N}{k} p^k (1-p)^{N-k} \quad \text{for } k = 0, 1, 2, \ldots, N,$$

where

$$\binom{N}{k} = \frac{N!}{k!(N-k)!}, \quad k = 0, 1, 2 \ldots, N.$$

This is the underlying probability model that we may postulate for this situation. The probability of a failure to pop for a given kernel, p, and the number of kernels, N, in one–quarter cup are unknown parameters. The challenge before us is to use the data in Table 1.1 on page 6 to estimate the parameter p. Notice that N is also unknown, so we'll also have to estimate N as well; however, the data in Table 1.1 do not give enough information do so. We will therefore have to design a new experiment to obtain data that will allow us to estimate N. This will be done in the next chapter. Before we proceed further, we will will lay out the sampling notions and terminology that are at the foundation of statistical inference.

1.1.2 Sampling: Concepts and Terminology

Suppose we wanted to estimate the number of popcorn kernels in one quarter cup of popcorn. In order to do this we can sample one quarter cup from a bag of popcorn and count the kernels in the quarter cup. Each time we do the sampling we get a value, N_i, for the number of kernels. We postulate that there is a value, μ, which gives the mean value of kernels in one quarter cup of popcorn. It is reasonable to assume that the distribution of each of the N_i, for $i = 1, 2, 3, \ldots$, is normal around μ with certain variance σ^2. That is,

$$N_i \sim \text{normal}(\mu, \sigma^2) \quad \text{for all } i = 1, 2, 3, \ldots,$$

so that each of the N_is has a density function, $f_{_N}$, given by

$$f_{_N}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad \text{for } -\infty < x < \infty.$$

[2] A *random variable* is a numerical outcome of a random experiment whose value cannot be determined with certainty.

Hence, the probability that the number of kernels in one quarter cup of popcorn lies within certain range of values, $a \leqslant N < b$ is

$$P(a \leqslant N < b) = \int_a^b f_{_N}(x)\ \mathrm{d}x.$$

Notice here that we are approximating a discrete random variable, N, by a continuous one. This approximation is justified if we are dealing with large numbers of kernels, so that a few kernels might not make a large relative difference. Table 1.2 shows a few of those numbers. If we also assume that the N_is are independent random variables, then $N_1, N_2, \ldots, N_n$ constitutes a random sample of size n.

Sample	Number of Kernels
1	356
2	368
3	356
4	351
5	339
6	298
7	289
8	352
9	447
10	314
11	332
12	369
13	298
14	327
15	319
16	316
17	341
18	367
19	357
20	334

Table 1.2: Number of Kernels in 1/4–cup of popcorn

Definition 1.1.1 (Random Sample)**.** (See also [HCM04, Definition 5.1.1, p 234]) The random variables, $X_1, X_2, \ldots, X_n$, form a random sample of size n on a random variable X if they are independent and each has the same distribution as that of X. We say that $X_1, X_2, \ldots, X_n$ constitute a random sample from the distribution of X.

Example 1.1.2. The second column of Table 1.2 shows values from a random sample from from the distribution of the number of kernels, N, in one-quarter cup of popcorn kernels.

Given a random sample, $X_1, X_2, \ldots, X_n$, from the distribution of a random variable, X, the sample mean, $\overline{X}_n$, is defined by

$$\overline{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

$\overline{X}_n$ is an example of a statistic.

Definition 1.1.3 (Statistic)**.** (See also [HCM04, Definition 5.1.2, p 235]) A statistic is a function of a random sample. In other words, a statistic is a quantity that is calculated from data contained in a random sample.

Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution of mean μ and variance σ^2. Then the expected value of the sample mean $\overline{X}_n$ is

$$E(\overline{X}_n) = \mu.$$

We say that $\overline{X}_n$ is an **unbiased** estimator for the mean μ.

Example 1.1.4 (Unbiased Estimation of the Variance)**.** Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution of mean μ and variance σ^2. Consider

$$\begin{aligned}
\sum_{k=1}^{n}(X_k - \mu)^2 &= \sum_{k=1}^{n}\left[X_k^2 - 2\mu X_k + \mu^2\right] \\
&= \sum_{k=1}^{n} X_k^2 - 2\mu\sum_{k=1}^{n} X_k + n\mu^2 \\
&= \sum_{k=1}^{n} X_k^2 - 2\mu n\overline{X}_n + n\mu^2.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\sum_{k=1}^{n}(X_k - \overline{X}_n)^2 &= \sum_{k=1}^{n}\left[X_k^2 - 2\overline{X}_n X_k + \overline{X}_n^2\right] \\
&= \sum_{k=1}^{n} X_k^2 - 2\overline{X}_n\sum_{k=1}^{n} X_k + n\overline{X}_n^2 \\
&= \sum_{k=1}^{n} X_k^2 - 2n\overline{X}_n\overline{X}_n + n\overline{X}_n^2 \\
&= \sum_{k=1}^{n} X_k^2 - n\overline{X}_n^2.
\end{aligned}$$

Consequently,

$$\sum_{k=1}^{n}(X_k-\mu)^2-\sum_{k=1}^{n}(X_k-\overline{X}_n)^2 \quad = \quad n\overline{X}_n^2-2\mu n\overline{X}_n+n\mu^2=n(\overline{X}_n-\mu)^2.$$

It then follows that

$$\sum_{k=1}^{n}(X_k-\overline{X}_n)^2=\sum_{k=1}^{n}(X_k-\mu)^2-n(\overline{X}_n-\mu)^2.$$

Taking expectations on both sides, we get

$$\begin{aligned} E\left(\sum_{k=1}^{n}(X_k-\overline{X}_n)^2\right) &= \sum_{k=1}^{n}E\left[(X_k-\mu)^2\right]-nE\left[(\overline{X}_n-\mu)^2\right] \\ &= \sum_{k=1}^{n}\sigma^2-n\mathrm{var}(\overline{X}_n) \\ &= n\sigma^2-n\frac{\sigma^2}{n} \\ &= (n-1)\sigma^2. \end{aligned}$$

Thus, dividing by $n-1$,

$$E\left(\frac{1}{n-1}\sum_{k=1}^{n}(X_k-\overline{X}_n)^2\right)=\sigma^2.$$

Hence, the random variable

$$S_n^2=\frac{1}{n-1}\sum_{k=1}^{n}(X_k-\overline{X}_n)^2,$$

called the **sample variance**, is an unbiased estimator of the variance.

Given a random sample, $X_1, X_2, \ldots, X_n$, from a distribution with mean μ and variance σ^2, and a statistic, $T=T(X_1,X_2,\ldots,X_n)$, based on the random sample, it is of interest to find out what the distribution of the statistic, T, is. This is called the sampling distribution of T. For example, we would like to know what the sampling distribution of the sample mean, $\overline{X}_n$, is. In order to find out what the sampling distribution of a statistic is, we need to know the joint distribution, $F_{(X_1,X_2,\ldots,X_n)}(x_1,x_2,\ldots,x_n)$, of the sample variable $X_1, X_2, \ldots, X_n$ is. Since, the variables $X_1, X_2, \ldots, X_n$ are independently and identically distributed (iid), then we can compute

$$F_{(X_1,X_2,\ldots,X_n)}(x_1,x_2,\ldots,x_n)=F_X(x_1)\cdot F_X(x_2)\cdots F_X(x_n),$$

where F_X is the common distribution. Recall that

$$F_X(x) = \mathrm{P}(X \leqslant x)$$

and

$$F_{(X_1,X_2,\ldots,X_n)}(x_1,x_2,\ldots,x_n) = \mathrm{P}(X_1 \leqslant x_1, X_2 \leqslant x_2, \ldots, X_n \leqslant x_n).$$

If X is a continuous random variable with density $f_X(x)$, then the joint density of the sample is

$$f_{(_1,X_2,\ldots,X_n)}(x_1,x_2,\ldots,x_n) = f_X(x_1)\cdot f_X(x_2)\cdots f_X(x_n).$$

Example 1.1.5. Let $N_1, N_2, \ldots, N_n$ denote a random sample from the experiment consisting of scooping up a quarter-cup of kernels popcorn from bag and counting the number of kernels. Assume that each N_i has a normal(μ, σ^2) distribution. We would like to find the distribution of the sample mean $\overline{N}_n$. We can do this by first computing the moment generating function (mgf), $M_{\overline{N}_n}(t)$, of $\overline{N}_n$:

$$\begin{aligned}
M_{\overline{N}_n}(t) &= E(e^{t\overline{N}_n}) \\
&= E\left(e^{(N_1+N_2+\cdots+N_n)\left(\frac{t}{n}\right)}\right) \\
&= M_{N_1+N_2+\cdots+N_n}\left(\frac{t}{n}\right) \\
&= M_{X_1}\left(\frac{t}{n}\right) M_{N_2}\left(\frac{t}{n}\right)\cdots M_{N_n}\left(\frac{t}{n}\right),
\end{aligned}$$

since the N_is are independent. Thus, since the N_is are also identically distributed,

$$M_{\overline{N}_n}(t) = \left(M_{N_1}\left(\frac{t}{n}\right)\right)^n,$$

where $M_{N_1}\left(\dfrac{t}{n}\right) = e^{\mu t/n + \sigma^2 t^2/2n^2}$, since N_1 has a normal(μ, σ^2) distribution. It then follows that

$$M_{\overline{N}_n}(t) = e^{\mu t + \sigma^2 t^2/2n},$$

which is the mgf of a normal$(\mu, \sigma^2/n)$ distribution. It then follows that $\overline{N}_n$ has a normal distribution with mean

$$E(\overline{N}_n) = \mu$$

and variance

$$\mathrm{var}(\overline{N}_n) = \frac{\sigma^2}{n}.$$

Example 1.1.5 shows that the sample mean, $\overline{X}_n$, for a random sample from a normal(μ, σ^2) distribution follows a normal$(\mu, \sigma^2/n)$. A surprising, and extremely useful, result from the theory of probability, states that for large values of n the sample mean for samples from any distribution are approximately normal$(\mu, \sigma^2/n)$. This is the essence of the Central Limit Theorem:

Theorem 1.1.6 (Central Limit Theorem)**.** [HCM04, Theorem 4.4.1, p 220] Suppose $X_1, X_2, X_3 \ldots$ are independent, identically distributed random variables with $E(X_i) = \mu$ and finite variance $\operatorname{var}(X_i) = \sigma^2$, for all i. Then

$$\lim_{n\to\infty} \mathrm{P}\left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \leqslant z\right) = \mathrm{P}(Z \leqslant z),$$

for all $z \in \mathbb{R}$, where $Z \sim \text{Normal}(0,1)$.

Thus, for large values of n, the distribution function for $\dfrac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$ can be approximated by the standard normal distribution. We write

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} Z \sim \text{Normal}(0,1)$$

and say that $\dfrac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$ **converges in distribution** to Z. In general, we have

Definition 1.1.7 (Convergence in Distribution)**.** A sequence, (Y_n), of random variables is said to converge in distribution to a random variable Y if

$$\lim_{n\to\infty} F_{Y_n}(y) = F_Y(y)$$

for all y where F_Y is continuous. We write

$$Y_n \xrightarrow{D} Y \quad \text{as } n \to \infty.$$

In practice, the Central Limit Theorem is applied to approximate the probabilities

$$\mathrm{P}\left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \leqslant z\right) \approx \mathrm{P}(Z \leqslant z) \quad \text{for larege } \ n,$$

which we could write as

$$F_{\overline{X}_n} \approx F_{\mu + \frac{\sigma}{\sqrt{n}}Z} \quad \text{for larege } \ n;$$

in other words, for large sample sizes, n, the distribution of the sample mean is approximately normal$(\mu, \sigma^2/n)$.

Chapter 2

Estimation

2.1 Estimating the Mean of a Distribution

We saw in the previous section that the sample mean, $\overline{X}_n$, of a random sample, $X_1, X_2, \ldots, X_n$, from a distribution with mean μ is an unbiased estimator for μ; that is, $E(\overline{X}_n) = \mu$. In this section we will see that, as we increase the sample size, n, then the sample means, $\overline{X}_n$, approach μ in probability; that is, for every $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathrm{P}(|\overline{X}_n - \mu| \geqslant \varepsilon) = 0,$$

or

$$\lim_{n\to\infty} \mathrm{P}(|\overline{X}_n - \mu| < \varepsilon) = 1.$$

We then say that $\overline{X}_n$ **converges to μ in probability** and write

$$\overline{X}_n \xrightarrow{\mathrm{P}} \mu \quad \text{as } n \to \infty.$$

Definition 2.1.1 (Convergence in Probability)**.** A sequence, (Y_n), of random variables is said to converge in probability to $b \in \mathbb{R}$, if for every $\varepsilon > 0$

$$\lim_{n\to\infty} \mathrm{P}(|Y_n - b| < \varepsilon) = 1.$$

We write

$$Y_n \xrightarrow{\mathrm{P}} b \quad \text{as } n \to \infty.$$

The fact that $\overline{X}_n$ converges to μ in probability is known as the weak **Law of Large Numbers**. We will prove this fact under the assumption that the distribution being sampled has finite variance, σ^2. Then, the weak Law of Large Numbers will follow from the inequality:

Theorem 2.1.2 (Chebyshev Inequality)**.** Let X be a random variable with mean μ and variance $\mathrm{var}(X)$. Then, for every $\varepsilon > 0$,

$$\mathrm{P}(|X - \mu| \geqslant \varepsilon) \leqslant \frac{\mathrm{var}(X)}{\varepsilon^2}.$$

Proof: We shall prove this inequality for the case in which X is continuous with pdf f_X.

Observe that $\text{var}(X) = E[(X-\mu)^2] = \displaystyle\int_{-\infty}^{\infty} |x-\mu|^2 f_X(x)\ \mathrm{d}x$. Thus,

$$\text{var}(X) \geqslant \int_{A_\varepsilon} |x-\mu|^2 f_X(x)\ \mathrm{d}x,$$

where $A_\varepsilon = \{x \in \mathbb{R} \mid |x-\mu| \geqslant \varepsilon\}$. Consequently,

$$\text{var}(X) \geqslant \varepsilon^2 \int_{A_\varepsilon} f_X(x)\ \mathrm{d}x = \varepsilon^2 \mathrm{P}(A_\varepsilon).$$

we therefore get that

$$\mathrm{P}(A_\varepsilon) \leqslant \frac{\text{var}(X)}{\varepsilon^2},$$

or

$$\mathrm{P}(|X-\mu| \geqslant \varepsilon) \leqslant \frac{\text{var}(X)}{\varepsilon^2}.$$

□

Applying Chebyshev Inequality to the case in which X is the sample mean, $\overline{X}_n$, we get

$$\mathrm{P}(|\overline{X}_n - \mu| \geqslant \varepsilon) \leqslant \frac{\text{var}(\overline{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}.$$

We therefore obtain that

$$\mathrm{P}(|\overline{X}_n - \mu| < \varepsilon) \geqslant 1 - \frac{\sigma^2}{n\varepsilon^2}.$$

Thus, letting $n \to \infty$, we get that, for every $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathrm{P}(|\overline{X}_n - \mu| < \varepsilon) = 1.$$

Later in these notes will we need the fact that a continuous function of a sequence which converges in probability will also converge in probability:

Theorem 2.1.3 (Slutsky's Theorem)**.** Suppose that (Y_n) converges in probability to b as $n \to \infty$ and that g is a function which is continuous at b. Then, $(g(Y_n))$ converges in probability to $g(b)$ as $n \to \infty$.

Proof: Let $\varepsilon > 0$ be given. Since g is continuous at b, there exists $\delta > 0$ such that

$$|y-b| < \delta \Rightarrow |g(y) - g(b)| < \varepsilon.$$

It then follows that the event $A_\delta = \{y \mid |y-b| < \delta\}$ is a subset the event $B_\varepsilon = \{y \mid |g(y) - g(b)| < \varepsilon\}$. Consequently,

$$\mathrm{P}(A_\delta) \leqslant \mathrm{P}(B_\varepsilon).$$

It then follows that

$$\mathrm{P}(|Y_n - b| < \delta) \leqslant \mathrm{P}(|g(Y_n) - g(b)| < \varepsilon) \leqslant 1. \tag{2.1}$$

Now, since $Y_n \xrightarrow{\mathrm{P}} b$ as $n \to \infty$,

$$\lim_{n\to\infty} \mathrm{P}(|Y_n - b| < \delta) = 1.$$

It then follows from Equation (2.1) and the Squeeze or Sandwich Theorem that

$$\lim_{n\to\infty} \mathrm{P}(|g(Y_n) - g(b)| < \varepsilon) = 1.$$

□

Since the sample mean, $\overline{X}_n$, converges in probability to the mean, μ, of sampled distribution, by the weak Law of Large Numbers, we say that $\overline{X}_n$ is a **consistent** estimator for μ.

2.2 Interval Estimate for Proportions

Example 2.2.1 (Estimating Proportions)**.** Let $X_1, X_2, X_3, \ldots$ denote independent identically distributed (iid) Bernoulli(p) random variables. Then the sample mean, $\overline{X}_n$, is an unbiased and consistent estimator for p. Denoting $\overline{X}_n$ by $\widehat{p}_n$, we then have that

$$E(\widehat{p}_n) = p \quad \text{for all } n = 1, 2, 3, \ldots,$$

and

$$\widehat{p}_n \xrightarrow{\mathrm{P}} p \quad \text{as } n \to \infty;$$

that is, for every $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathrm{P}(|\widehat{p}_n - p| < \varepsilon) = 1.$$

By Slutsky's Theorem (Theorem 2.1.3), we also have that

$$\sqrt{\widehat{p}_n(1-\widehat{p}_n)} \xrightarrow{\mathrm{P}} \sqrt{p(1-p)} \quad \text{as } n \to \infty.$$

Thus, the statistic $\sqrt{\widehat{p}_n(1-\widehat{p}_n)}$ is a consistent estimator of the standard deviation $\sigma = \sqrt{p(1-p)}$ of the Bernoulli(p) trials $X_1, X_2, X_3, \ldots$

Now, by the Central Limit Theorem, we have that

$$\lim_{n\to\infty} \mathrm{P}\left(\frac{\widehat{p}_n - p}{\sigma/\sqrt{n}} \leqslant z\right) = \mathrm{P}(Z \leqslant z),$$

where $Z \sim \text{Normal}(0,1)$, for all $z \in \mathbb{R}$. Hence, since $\sqrt{\widehat{p}_n(1-\widehat{p}_n)}$ is a consistent estimator for σ, we have that, for large values of n,

$$\mathrm{P}\left(\frac{\widehat{p}_n - p}{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}/\sqrt{n}} \leqslant z\right) \approx \mathrm{P}(Z \leqslant z),$$

for all $z \in \mathbb{R}$. Similarly, for large values of n,

$$\mathrm{P}\left(\frac{\widehat{p}_n - p}{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}/\sqrt{n}} \leqslant -z\right) \approx \mathrm{P}(Z \leqslant -z).$$

subtracting this from the previous expression we get

$$\mathrm{P}\left(-z < \frac{\widehat{p}_n - p}{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}/\sqrt{n}} \leqslant z\right) \approx \mathrm{P}(-z < Z \leqslant z)$$

for large values of n, or

$$\mathrm{P}\left(-z \leqslant \frac{p - \widehat{p}_n}{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}/\sqrt{n}} < z\right) \approx \mathrm{P}(-z < Z \leqslant z)$$

for large values of n.

Now, suppose that $z > 0$ is such that $\mathrm{P}(-z < Z \leqslant z) \geqslant 0.95$. Then, for that value of z, we get that, approximately, for large values of n,

$$\mathrm{P}\left(\widehat{p}_n - z\frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{\sqrt{n}} \leqslant p < \widehat{p}_n + z\frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{\sqrt{n}}\right) \geqslant 0.95$$

Thus, for large values of n, the intervals

$$\left[\widehat{p}_n - z\frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{\sqrt{n}},\ \widehat{p}_n + z\frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{\sqrt{n}}\right)$$

have the property that the probability that the true proportion p lies in them is at least 95%. For this reason, the interval

$$\left[\widehat{p}_n - z\frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{\sqrt{n}},\ \widehat{p}_n + z\frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{\sqrt{n}}\right)$$

is called the 95% **confidence interval estimate for the proportion** p. To find the value of z that yields the 95% confidence interval for p, observe that

$$\mathrm{P}(-z < Z \leqslant z) = F_{z}(z) - F_{z}(-z) = F_{z}(z) - (1 - F_{z}(z)) = 2F_{z}(z) - 1.$$

Thus, we need to solve for z in the inequality

$$2F_{z}(z) - 1 \geqslant 0.95$$

or

$$F_{z}(z) \geqslant 0.975.$$

This yields $z = 1.96$. We then get that the **approximate** 95% confidence interval estimate for the proportion p is

$$\left[\widehat{p}_n - 1.96\frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{\sqrt{n}},\ \widehat{p}_n + 1.96\frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{\sqrt{n}}\right)$$

Example 2.2.2. In the corn–popping experiment described in Section 1.1.1, out of 356 kernels, 52 fail to pop. In this example, we compute a 95% confidence interval for p, the probability of failure to pop for a given kernel, based on this information. An estimate for p in this case is $\widehat{p}_n = 52/356 \approx 0.146$. An approximate 95% confidence interval estimate for the true proportion of kernels, p, which will not pop is then

$$\left[0.146 - 1.96\frac{\sqrt{0.146(0.854)}}{\sqrt{356}},\ 0.146 + 1.96\frac{\sqrt{0.146(0.854)}}{\sqrt{356}}\right),$$

or about $[0.146-0.037, 0.146+0.037)$, or $[0.109, 0.183)$. Thus, the failure to pop rate is between 10.9% and 18.3% with a 95% confidence level. The confidence level here indicates the probability that the method used to produce the interval estimate from the data will contain the true value of the parameter being estimated.

2.3 Interval Estimates for the Mean

In the previous section we obtained an approximate confidence interval (CI) estimate for the probability that a given kernel will fail to pop. We did this by using the fact that, for large numbers of trials, a binomial distribution can be approximated by a normal distribution (by the Central Limit Theorem). We also used the fact that the sample standard deviation $\sqrt{\widehat{p}_n(1-\widehat{p}_n)}$ is a consistent estimator of the standard deviation $\sigma = \sqrt{p(1-p)}$ of the Bernoulli(p) trials $X_1, X_2, X_3, \ldots$ The consistency condition might not hold in general. However, in the case in which sampling is done from a normal distribution an exact confidence interval estimate may be obtained based on on the sample mean and variance by means of the t–distribution. We present this development here and apply it to the problem of estimating the mean number of popcorn kernels in one quarter cup.

We have already seen that the sample mean, $\overline{X}_n$, of a random sample of size n from a normal(μ, σ^2) follows a normal$(\mu, \sigma^2/n)$ distribution. It then follows that

$$\mathrm{P}\left(\frac{|\overline{X}_n - \mu|}{\sigma/\sqrt{n}}\right) = \mathrm{P}(|Z| \leqslant z) \quad \text{for all} \quad z \in \mathbb{R}, \tag{2.2}$$

where $Z \sim$ normal$(0,1)$. Thus, if we knew σ, then we could obtain the 95% CI for μ by choosing $z = 1.96$ in (2.2). We would then obtain the CI:

$$\left[\overline{X}_n - 1.96\frac{\sigma}{\sqrt{n}}, \overline{X}_n + 1.96\frac{\sigma}{\sqrt{n}}\right].$$

However, σ is generally and unknown parameter. So, we need to resort to a different kind of estimate. The idea is to use the sample variance, S_n^2, to estimate σ^2, where

$$S_n^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \overline{X}_n)^2. \tag{2.3}$$

Thus, instead of considering the normalized sample means

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}},$$

we consider the random variables

$$T_n = \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}}. \tag{2.4}$$

The task that remains then is to determine the sampling distribution of T_n. This was done by William Sealy Gosset in 1908 in an article published in the journal Biometrika under the pseudonym Student [Stu08]. The fact the we can actually determine the distribution of T_n in (2.4) depends on the fact that $X_1, X_2, \ldots, X_n$ is a random sample from a normal distribution and knowledge of the χ^2 distribution.

2.3.1 The χ^2 Distribution

Example 2.3.1 (The Chi–Square Distribution with one degree of freedom)**.** Let $Z \sim \text{Normal}(0, 1)$ and define $X = Z^2$. Give the probability density function (pdf) of X.

Solution: The pdf of Z is given by

$$f_{_X}(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad \text{for } -\infty < z < \infty.$$

We compute the pdf for X by first determining its cumulative density function (cdf):

$$\begin{aligned} P(X \leq x) &= P(Z^2 \leq x) \quad \text{for } y \geqslant 0 \\ &= P(-\sqrt{x} \leq Z \leq \sqrt{x}) \\ &= P(-\sqrt{x} < Z \leq \sqrt{x}), \quad \text{since Z is continuous.} \end{aligned}$$

Thus,

$$\begin{aligned} P(X \leq x) &= P(Z \leq \sqrt{x}) - P(Z \leq -\sqrt{x}) \\ &= F_{_Z}(\sqrt{x}) - F_{_Z}(-\sqrt{x}) \quad \text{for } x > 0, \end{aligned}$$

since X is continuous.

We then have that the cdf of X is

$$F_{_X}(x) = F_{_Z}(\sqrt{x}) - F_{_Z}(-\sqrt{x}) \quad \text{for } x > 0,$$

from which we get, after differentiation with respect to x,

$$\begin{aligned} f_{_X}(x) &= F'_{_Z}(\sqrt{x}) \cdot \frac{1}{2\sqrt{x}} + F'_{_Z}(-\sqrt{x}) \cdot \frac{1}{2\sqrt{x}} \\ &= f_{_Z}(\sqrt{x}) \frac{1}{2\sqrt{x}} + f_{_Z}(-\sqrt{x}) \frac{1}{2\sqrt{x}} \\ &= \frac{1}{2\sqrt{x}} \left\{ \frac{1}{\sqrt{2\pi}}\, e^{-x/2} + \frac{1}{\sqrt{2\pi}}\, e^{-x/2} \right\} \\ &= \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{x}}\, e^{-x/2} \end{aligned}$$

for $x > 0$. □

Definition 2.3.2. A continuous random variable, X having the pdf

$$f_{_X}(x) = \begin{cases} \dfrac{1}{\sqrt{2\pi}} \cdot \dfrac{1}{\sqrt{x}}\, e^{-x/2} & \text{if } x > 0 \\ \\ 0 & \text{otherwise,} \end{cases}$$

is said to have a Chi–Square distribution with one degree of freedom. We write

$$Y \sim \chi^2(1).$$

Remark 2.3.3. Observe that if $X \sim \chi^2_{_1}$, then its expected value is

$$E(X) = E(Z^2) = 1,$$

since $\text{var}(Z) = E(Z^2) - (E(Z))^2$ and $E(Z) = 0$ and $\text{var}(Z) = 1$. To compute the second moment of X, $E(X^2) = E(Z^4)$, we need to compute the fourth moment of Z. In order to do this, we first compute the mgf of Z is

$$M_{_Z}(t) = e^{t^2/2} \quad \text{for all } t \in \mathbb{R}.$$

Its fourth derivative can be computed to be

$$M^{(4)}_{_Z}(t) = (3 + 6t^2 + t^4)\, e^{t^2/2} \quad \text{for all } t \in \mathbb{R}.$$

Thus,

$$E(Z^4) = M^{(4)}_{_Z}(0) = 3.$$

We then have that the variance of X is

$$\text{var}(X) = E(X^2) - (E(X))^2 = E(Z^4) - 1 = 3 - 1 = 2.$$

Suppose next that we have two independent random variable, X and Y, both of which have a $\chi^2(1)$ distribution. We would like to know the distribution of the sum $X + Y$.

Denote the sum $X+Y$ by W. We would like to compute the pdf f_W. Since X and Y are independent, f_W is given by the convolution of f_X and f_Y; namely,

$$f_{_W}(w) = \int_{-\infty}^{+\infty} f_{_X}(u) f_{_Y}(w-u)\mathrm{d}u,$$

where

$$f_{_X}(x) = \begin{cases} \dfrac{1}{\sqrt{2\pi}}\dfrac{1}{\sqrt{x}}e^{-x/2} & x > 0, \\ \\ 0 & \text{elsewhere}, \end{cases} \qquad f_{_Y}(y) = \begin{cases} \dfrac{1}{\sqrt{2\pi}}\dfrac{1}{\sqrt{y}}e^{-y/2} & y > 0 \\ \\ 0 & \text{otherwise}. \end{cases}$$

We then have that

$$f_{_W}(w) \;=\; \int_0^{\infty} \frac{1}{\sqrt{2\pi}\sqrt{u}} e^{-u/2} f_{_Y}(w-u)\ \mathrm{d}u,$$

since $f_{_X}(u)$ is zero for negative values of u. Similarly, since $f_{_Y}(w-u) = 0$ for $w-u<0$, we get that

$$\begin{aligned} f_{_W}(w) &= \int_0^{w} \frac{1}{\sqrt{2\pi}\sqrt{u}} e^{-u/2} \frac{1}{\sqrt{2\pi}\sqrt{w-u}} e^{-(w-u)/2}\ \mathrm{d}u \\ &= \frac{e^{-w/2}}{2\pi} \int_0^{w} \frac{1}{\sqrt{u}\sqrt{w-u}}\ \mathrm{d}u. \end{aligned}$$

Next, make the change of variables $t = \dfrac{u}{w}$. Then, $\mathrm{d}u = w\mathrm{d}t$ and

$$\begin{aligned} f_{_W}(w) &= \frac{e^{-w/2}}{2\pi} \int_0^{1} \frac{w}{\sqrt{wt}\sqrt{w-wt}}\ \mathrm{d}t \\ &= \frac{e^{-w/2}}{2\pi} \int_0^{1} \frac{1}{\sqrt{t}\sqrt{1-t}}\ \mathrm{d}t. \end{aligned}$$

Making a second change of variables $s = \sqrt{t}$, we get that $t = s^2$ and $\mathrm{d}t = 2s\mathrm{d}s$, so that

$$\begin{aligned} f_{_W}(w) &= \frac{e^{-w/2}}{\pi} \int_0^{1} \frac{1}{\sqrt{1-s^2}}\ \mathrm{d}s \\ &= \frac{e^{-w/2}}{\pi}\, [\arcsin(s)]_0^1 \\ &= \frac{1}{2}\, e^{-w/2} \quad \text{for } w > 0, \end{aligned}$$

and zero otherwise. It then follows that $W = X+Y$ has the pdf of an exponential(2) random variable.

Definition 2.3.4 (χ^2 distribution with n degrees of freedom)**.** Let $X_1, X_2, \ldots, X_n$ be independent, identically distributed random variables with a $\chi^2(1)$ distribution. Then then random variable $X_1 + X_2 + \cdots + X_n$ is said to have a χ^2 distribution with n degrees of freedom. We write

$$X_1 + X_2 + \cdots + X_n \sim \chi^2(n).$$

The calculations preceding Definition 2.3.4 if a random variable, W, has a $\chi^2(2)$ distribution, then its pdf is given by

$$f_W(w) = \begin{cases} \dfrac{1}{2}\, e^{-w/2} & \text{for } w > 0; \\ \\ 0 & \text{for } w \leqslant 0; \end{cases}$$

Our goal in the following set of examples is to come up with the formula for the pdf of a $\chi^2(n)$ random variable.

Example 2.3.5 (Three degrees of freedom)**.** Let $X \sim$ exponential(2) and $Y \sim \chi^2(1)$ be independent random variables and define $W = X + Y$. Give the distribution of W.

Solution: Since X and Y are independent, by Problem 1 in Assignment #3, f_W is the convolution of f_X and f_Y:

$$\begin{aligned} f_W(w) &= f_X * f_Y(w) \\ &= \int_{-\infty}^{\infty} f_X(u) f_Y(w-u)\mathrm{d}u, \end{aligned}$$

where

$$f_X(x) = \begin{cases} \dfrac{1}{2} e^{-x/2} & \text{if } \ x > 0; \\ \\ 0 & \text{otherwise}; \end{cases}$$

and

$$f_Y(y) = \begin{cases} \dfrac{1}{\sqrt{2\pi}} \dfrac{1}{\sqrt{y}}\ e^{-y/2} & \text{if } \ y > 0; \\ \\ 0 & \text{otherwise}. \end{cases}$$

It then follows that, for $w > 0$,

$$\begin{aligned} f_W(w) &= \int_0^{\infty} \frac{1}{2} e^{-u/2} f_Y(w-u)\mathrm{d}u \\ &= \int_0^{w} \frac{1}{2} e^{-u/2} \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{w-u}}\ e^{-(w-u)/2}\mathrm{d}u \\ &= \frac{e^{-w/2}}{2\sqrt{2\pi}} \int_0^{w} \frac{1}{\sqrt{w-u}}\mathrm{d}u. \end{aligned}$$

Making the change of variables $t = u/w$, we get that $u = wt$ and $\mathrm{d}u = w\mathrm{d}t$, so that

$$\begin{aligned} f_{_W}(w) &= \frac{e^{-w/2}}{2\sqrt{2\pi}} \int_0^1 \frac{1}{\sqrt{w - wt}}\, w\mathrm{d}t \\ &= \frac{\sqrt{w}\ e^{-w/2}}{2\sqrt{2\pi}} \int_0^1 \frac{1}{\sqrt{1-t}}\ \mathrm{d}t \\ &= \frac{\sqrt{w}\ e^{-w/2}}{\sqrt{2\pi}} \left[-\sqrt{1-t}\right]_0^1 \\ &= \frac{1}{\sqrt{2\pi}}\ \sqrt{w}\ e^{-w/2}, \end{aligned}$$

for $w > 0$. It then follows that

$$f_{_W}(w) = \begin{cases} \dfrac{1}{\sqrt{2\pi}}\ \sqrt{w}\ e^{-w/2} & \text{if } \ w > 0; \\ 0 & \text{otherwise.} \end{cases}$$

This is the pdf for a $\chi^2(3)$ random variable. □

Example 2.3.6 (Four degrees of freedom)**.** Let $X, Y \sim$ exponential(2) be independent random variables and define $W = X + Y$. Give the distribution of W.

Solution: Since X and Y are independent, $f_{_W}$ is the convolution of $f_{_X}$ and $f_{_Y}$:

$$\begin{aligned} f_{_W}(w) &= f_{_X} * f_{_Y}(w) \\ &= \int_{-\infty}^{\infty} f_{_X}(u) f_{_Y}(w-u)\mathrm{d}u, \end{aligned}$$

where

$$f_{_X}(x) = \begin{cases} \dfrac{1}{2} e^{-x/2} & \text{if } \ x > 0; \\ 0 & \text{otherwise;} \end{cases}$$

and

$$f_{_Y}(y) = \begin{cases} \dfrac{1}{2}\ e^{-y/2} & \text{if } \ y > 0; \\ 0 & \text{otherwise.} \end{cases}$$

It then follows that, for $w > 0$,

$$\begin{aligned} f_{_W}(w) &= \int_0^\infty \frac{1}{2}e^{-u/2} f_{_Y}(w-u)\mathrm{d}u \\ &= \int_0^w \frac{1}{2}e^{-u/2}\frac{1}{2}\ e^{-(w-u)/2}\mathrm{d}u \\ &= \frac{e^{-w/2}}{4}\int_0^w \mathrm{d}u \\ &= \frac{w\ e^{-w/2}}{4}, \end{aligned}$$

for $w > 0$. It then follows that

$$f_{_W}(w) = \begin{cases} \dfrac{1}{4}\ w\ e^{-w/2} & \text{if } \ w > 0; \\ \\ 0 & \text{otherwise.} \end{cases}$$

This is the pdf for a $\chi^2(4)$ random variable. □

We are now ready to derive the general formula for the pdf of a $\chi^2(n)$ random variable.

Example 2.3.7 (n degrees of freedom)**.** In this example we prove that if $W \sim \chi^2(n)$, then the pdf of W is given by

$$f_{_W}(w) = \begin{cases} \dfrac{1}{\Gamma(n/2)\ 2^{n/2}}\ w^{\frac{n}{2}-1}\ e^{-w/2} & \text{if } \ w > 0; \\ \\ 0 & \text{otherwise,} \end{cases} \tag{2.5}$$

where Γ denotes the Gamma function defined by

$$\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\mathrm{d}t \qquad \text{for all real values of } z \text{ except } \ 0, -1, -2, -3, \ldots$$

Proof: We proceed by induction of n. Observe that when $n = 1$ the formula in (2.5) yields, for $w > 0$,

$$f_{_W}(w) = \frac{1}{\Gamma(1/2)\ 2^{1/2}}\ w^{\frac{1}{2}-1}\ e^{-w/2} = \frac{1}{\sqrt{2\pi}}\ \frac{1}{\sqrt{x}}\ e^{-w/2},$$

which is the pdf for a $\chi^(1)$ random variable. Thus, the formula in (2.5) holds true for $n = 1$.

Next, assume that a $\chi^2(n)$ random variable has pdf given (2.5). We will show that if $W \sim \chi^2(n+1)$, then its pdf is given by

$$f_{_W}(w) = \begin{cases} \dfrac{1}{\Gamma((n+1)/2)\ 2^{(n+1)/2}}\ w^{\frac{n-1}{2}}\ e^{-w/2} & \text{if } \ w > 0; \\ \\ 0 & \text{otherwise.} \end{cases} \tag{2.6}$$

By the definition of a $\chi^2(n+1)$ random variable, we have that $W = X + Y$ where $X \sim \chi^2(n)$ and $Y \sim \chi^2(1)$ are independent random variables. It then follows that

$$f_{_W} = f_{_X} * f_{_Y}$$

where

$$f_{_X}(x) = \begin{cases} \dfrac{1}{\Gamma(n/2)\ 2^{n/2}}\ x^{\frac{n}{2}-1}\ e^{-x/2} & \text{if } \ x > 0; \\ \\ 0 & \text{otherwise.} \end{cases}$$

and

$$f_{_Y}(y) = \begin{cases} \dfrac{1}{\sqrt{2\pi}}\dfrac{1}{\sqrt{y}}\ e^{-y/2} & \text{if } \ y > 0; \\ \\ 0 & \text{otherwise.} \end{cases}$$

Consequently, for $w > 0$,

$$\begin{aligned} f_{_W}(w) &= \int_0^w \frac{1}{\Gamma(n/2)\ 2^{n/2}}\ u^{\frac{n}{2}-1}\ e^{-u/2}\ \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{w-u}}\ e^{-(w-u)/2}\mathrm{d}u \\ &= \frac{e^{-w/2}}{\Gamma(n/2)\sqrt{\pi}\ 2^{(n+1)/2}} \int_0^w \frac{u^{\frac{n}{2}-1}}{\sqrt{w-u}}\mathrm{d}u. \end{aligned}$$

Next, make the change of variables $t = u/w$; we then have that $u = wt$, $\mathrm{d}u = w\mathrm{d}t$ and

$$f_{_W}(w) = \frac{w^{\frac{n-1}{2}} e^{-w/2}}{\Gamma(n/2)\sqrt{\pi}\ 2^{(n+1)/2}} \int_0^1 \frac{t^{\frac{n}{2}-1}}{\sqrt{1-t}}\mathrm{d}t.$$

Making a further change of variables $t = z^2$, so that $\mathrm{d}t = 2z\mathrm{d}z$, we obtain that

$$f_{_W}(w) = \frac{2w^{\frac{n-1}{2}} e^{-w/2}}{\Gamma(n/2)\sqrt{\pi}\ 2^{(n+1)/2}} \int_0^1 \frac{z^{n-1}}{\sqrt{1-z^2}}\mathrm{d}z. \tag{2.7}$$

It remains to evaluate the integrals

$$\int_0^1 \frac{z^{n-1}}{\sqrt{1-z^2}}\mathrm{d}z \qquad \text{for } n = 1, 2, 3, \ldots$$

We can evaluate these by making the trigonometric substitution $z = \sin\theta$ so that $\mathrm{d}z = \cos\theta \mathrm{d}\theta$ and

$$\int_0^1 \frac{z^{n-1}}{\sqrt{1-z^2}}\mathrm{d}z = \int_0^{\pi/2} \sin^{n-1}\theta \mathrm{d}\theta.$$

Looking up the last integral in a table of integrals we find that, if n is even and $n \geqslant 4$, then

$$\int_0^{\pi/2} \sin^{n-1}\theta \mathrm{d}\theta = \frac{1\cdot 3\cdot 5\cdots(n-2)}{2\cdot 4\cdot 6\cdots(n-1)},$$

which can be written in terms of the Gamma function as

$$\int_0^{\pi/2} \sin^{n-1}\theta \mathrm{d}\theta = \frac{2^{n-2}\left[\Gamma\left(\frac{n}{2}\right)\right]^2}{\Gamma(n)}. \tag{2.8}$$

Note that this formula also works for $n = 2$.

Similarly, we obtain that for odd n with $n \geqslant 1$ that

$$\int_0^{\pi/2} \sin^{n-1}\theta \mathrm{d}\theta = \frac{\Gamma(n)}{2^{n-1}\left[\Gamma\left(\frac{n+1}{2}\right)\right]^2}\ \frac{\pi}{2}. \tag{2.9}$$

Now, if n is odd and $n \geqslant 1$ we may substitute (2.9) into (2.7) to get

$$\begin{aligned} f_W(w) &= \frac{2w^{\frac{n-1}{2}}e^{-w/2}}{\Gamma(n/2)\sqrt{\pi}\ 2^{(n+1)/2}}\frac{\Gamma(n)}{2^{n-1}\left[\Gamma\left(\frac{n+1}{2}\right)\right]^2}\ \frac{\pi}{2} \\ &= \frac{w^{\frac{n-1}{2}}e^{-w/2}}{\Gamma(n/2)\ 2^{(n+1)/2}}\frac{\Gamma(n)\sqrt{\pi}}{2^{n-1}\left[\Gamma\left(\frac{n+1}{2}\right)\right]^2}. \end{aligned}$$

Now, by Problem 5 in Assignment 1, for odd n,

$$\Gamma\left(\frac{n}{2}\right) = \frac{\Gamma(n)\sqrt{\pi}}{2^{n-1}\ \Gamma\left(\frac{n+1}{2}\right)}.$$

It the follows that

$$f_W(w) = \frac{w^{\frac{n-1}{2}}e^{-w/2}}{\Gamma\left(\frac{n+1}{2}\right)\ 2^{(n+1)/2}}$$

for $w > 0$, which is (2.6) for odd n.

Next, suppose that n is a positive, even integer. In this case we substitute (2.8) into (2.7) and get

$$f_W(w) = \frac{2w^{\frac{n-1}{2}}e^{-w/2}}{\Gamma(n/2)\sqrt{\pi}\ 2^{(n+1)/2}}\ \frac{2^{n-2}\left[\Gamma\left(\frac{n}{2}\right)\right]^2}{\Gamma(n)}$$

or

$$f_W(w) = \frac{w^{\frac{n-1}{2}}e^{-w/2}}{2^{(n+1)/2}}\ \frac{2^{n-1}\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi}\ \Gamma(n)} \tag{2.10}$$

Now, since n is even, $n+1$ is odd, so that by by Problem 5 in Assignment 1 again, we get that

$$\Gamma\left(\frac{n+1}{2}\right) = \frac{\Gamma(n+1)\sqrt{\pi}}{2^n\ \Gamma\left(\frac{n+2}{2}\right)} = \frac{n\Gamma(n)\sqrt{\pi}}{2^n\ \frac{n}{2}\Gamma\left(\frac{n}{2}\right)},$$

from which we get that

$$\frac{2^{n-1}\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi}\ \Gamma(n)} = \frac{1}{\Gamma\left(\frac{n+1}{2}\right)}.$$

Substituting this into (2.10) yields

$$f_W(w) \quad = \quad \frac{w^{\frac{n-1}{2}} e^{-w/2}}{\Gamma\left(\frac{n+1}{2}\right)\ 2^{(n+1)/2}}$$

for $w > 0$, which is (2.6) for even n. This completes inductive step and the proof is now complete. That is, if $W \sim \chi^2(n)$ then the pdf of W is given by

$$f_W(w) = \begin{cases} \dfrac{1}{\Gamma(n/2)\ 2^{n/2}}\ w^{\frac{n}{2}-1}\ e^{-w/2} & \text{if } w > 0; \\ \\ 0 & \text{otherwise,} \end{cases}$$

for $n = 1, 2, 3, \ldots$ □

2.3.2 The t Distribution

In this section we derive a very important distribution in statistics, the Student t distribution, or t distribution for short. We will see that this distribution will come in handy when we complete our discussion of estimating the mean based on a random sample from a normal(μ, σ^2) distribution.

Example 2.3.8 (The t distribution)**.** Let $Z \sim$ normal$(0, 1)$ and $X \sim \chi^2(n-1)$ be independent random variables. Define

$$T = \frac{Z}{\sqrt{X/(n-1)}}.$$

Give the pdf of the random variable T.

Solution: We first compute the cdf, F_T, of T; namely,

$$\begin{aligned} F_T(t) &= P(T \leqslant t) \\ &= P\left(\frac{Z}{\sqrt{X/(n-1)}} \leqslant t\right) \\ &= \iint_R f_{(X,Z)}(x,z)\mathrm{d}x\mathrm{d}z, \end{aligned}$$

where R_t is the region in the xz–plane given by

$$R_t = \{(x,z)\in\mathbb{R}^2 \mid z < t\sqrt{x/(n-1)}, x>0\},$$

and the joint distribution, $f_{(X,Z)}$, of X and Z is given by

$$f_{(X,Z)}(x,z) = f_X(x)\cdot f_Z(z) \quad \text{for} \quad x>0 \text{ and } \quad z\in\mathbb{R},$$

because X and Z are assumed to be independent. Furthermore,

$$f_X(x) = \begin{cases} \dfrac{1}{\Gamma\left(\frac{n-1}{2}\right)\ 2^{(n-1)/2}}\ x^{\frac{n-1}{2}-1}\ e^{-x/2} & \text{if } \ x>0; \\ \\ 0 & \text{otherwise,} \end{cases}$$

and

$$f_Z(z) = \frac{1}{\sqrt{2\pi}}\ e^{-z^2/2}, \quad \text{for } -\infty < z < \infty.$$

We then have that

$$F_T(t) \quad = \quad \int_0^\infty \int_{-\infty}^{t\sqrt{x/(n-1)}} \frac{x^{\frac{n-3}{2}}\ e^{-(x+z^2)/2}}{\Gamma\left(\frac{n-1}{2}\right)\sqrt{\pi}\ 2^{\frac{n}{2}}}\ \mathrm{d}z\mathrm{d}x.$$

Next, make the change of variables

$$\begin{aligned} u &= x \\ v &= \frac{z}{\sqrt{x/(n-1)}}, \end{aligned}$$

so that

$$\begin{aligned} x &= u \\ z &= v\sqrt{u/(n-1)}. \end{aligned}$$

Consequently,

$$F_T(t) \quad = \quad \frac{1}{\Gamma\left(\frac{n-1}{2}\right)\sqrt{\pi}\ 2^{\frac{n}{2}}} \int_{-\infty}^t \int_0^\infty u^{\frac{n-3}{2}}\ e^{-(u+uv^2/(n-1))/2} \left|\frac{\partial(x,z)}{\partial(u,v)}\right| \mathrm{d}u\mathrm{d}v,$$

where the Jacobian of the change of variables is

$$\begin{aligned} \frac{\partial(x,z)}{\partial(u,v)} &= \det\begin{pmatrix} 1 & 0 \\ \\ v/2\sqrt{u}\sqrt{n-1} & u^{1/2}/\sqrt{n-1} \end{pmatrix} \\ \\ &= \frac{u^{1/2}}{\sqrt{n-1}}. \end{aligned}$$

It then follows that

$$F_{_T}(t) \quad = \quad \frac{1}{\Gamma\left(\frac{n-1}{2}\right)\sqrt{(n-1)\pi}\ 2^{\frac{n}{2}}}\int_{-\infty}^{t}\int_{0}^{\infty} u^{\frac{n}{2}-1}\ e^{-(u+uv^2/(n-1))/2}\mathrm{d}u\mathrm{d}v.$$

Next, differentiate with respect to t and apply the Fundamental Theorem of Calculus to get

$$\begin{aligned} f_{_T}(t) &= \frac{1}{\Gamma\left(\frac{n-1}{2}\right)\sqrt{(n-1)\pi}\ 2^{\frac{n}{2}}}\int_{0}^{\infty} u^{\frac{n}{2}-1}\ e^{-(u+ut^2/(n-1))/2}\mathrm{d}u \\ &= \frac{1}{\Gamma\left(\frac{n-1}{2}\right)\sqrt{(n-1)\pi}\ 2^{\frac{n}{2}}}\int_{0}^{\infty} u^{\frac{n}{2}-1}\ e^{-\left(1+\frac{t^2}{n-1}\right)u/2}\mathrm{d}u. \end{aligned}$$

Put $\alpha = \dfrac{n}{2}$ and $\beta = \dfrac{2}{1+\frac{t^2}{n-1}}$. Then,

$$\begin{aligned} f_{_T}(t) &= \frac{1}{\Gamma\left(\frac{n-1}{2}\right)\sqrt{(n-1)\pi}\ 2^{\alpha}}\int_{0}^{\infty} u^{\alpha-1}\ e^{-u/\beta}\mathrm{d}u \\ &= \frac{\Gamma(\alpha)\beta^\alpha}{\Gamma\left(\frac{n-1}{2}\right)\sqrt{(n-1)\pi}\ 2^{\alpha}}\int_{0}^{\infty} \frac{u^{\alpha-1}\ e^{-u/\beta}}{\Gamma(\alpha)\beta^\alpha}\mathrm{d}u, \end{aligned}$$

where

$$f_{_U}(u) = \begin{cases} \dfrac{u^{\alpha-1}\ e^{-u/\beta}}{\Gamma(\alpha)\beta^\alpha} & \text{if } \ u > 0 \\ \\ 0 & \text{if } \ u \leqslant 0 \end{cases}$$

is the pdf of a $\Gamma(\alpha,\beta)$ random variable (see Problem 5 in Assignment #3). We then have that

$$f_{_T}(t) \quad = \quad \frac{\Gamma(\alpha)\beta^\alpha}{\Gamma\left(\frac{n-1}{2}\right)\sqrt{(n-1)\pi}\ 2^{\alpha}} \qquad \text{for } \ t \in \mathbb{R}.$$

Using the definitions of α and β we obtain that

$$f_{_T}(t) \quad = \quad \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)\sqrt{(n-1)\pi}} \cdot \frac{1}{\left(1+\dfrac{t^2}{n-1}\right)^{n/2}} \qquad \text{for } \ t \in \mathbb{R}.$$

This is the pdf of a random variable with a t distribution with $n-1$ degrees of freedom. In general, a random variable, T, is said to have a t distribution with r degrees of freedom, for $r \geqslant 1$, if its pdf is given by

$$f_{_T}(t) \quad = \quad \frac{\Gamma\left(\frac{r+1}{2}\right)}{\Gamma\left(\frac{r}{2}\right)\sqrt{r\pi}} \cdot \frac{1}{\left(1+\dfrac{t^2}{r}\right)^{(r+1)/2}} \qquad \text{for } \ t \in \mathbb{R}.$$

We write $T \sim t(r)$. Thus, in this example we have seen that, if $Z \sim \text{norma}(0,1)$ and $X \sim \chi^2(n-1)$, then

$$\frac{Z}{\sqrt{X/(n-1)}} \sim t(n-1).$$

□

We will see the relevance of this example in the next section when we continue our discussion estimating the mean of a norma distribution.

2.3.3 Sampling from a normal distribution

Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal(μ, σ^2) distribution. Then, the sample mean, $\overline{X}_n$ has a normal$(\mu, \sigma^2/n)$ distribution.

Observe that

$$|\overline{X}_n - \mu| < b \Leftrightarrow \left|\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}\right| < \frac{\sqrt{n}\, b}{\sigma},$$

where

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim \text{normal}(0,1).$$

Thus,

$$\mathrm{P}(|\overline{X}_n - \mu| < b) = \mathrm{P}\left(|Z| < \frac{\sqrt{n}\, b}{\sigma}\right), \tag{2.11}$$

where $Z \sim \text{normal}(0,1)$. Observer that the distribution of the standard normal random variable Z is independent of the parameters μ and σ. Thus, for given values of $z > 0$ we can compute $P(|Z| < z)$. For example, if there is a way of knowing the cdf for Z, either by looking up values in probability tables or suing statistical software packages to compute then, we have that

$$\begin{aligned}
\mathrm{P}(|Z| < z) &= \mathrm{P}(-z < Z < z) \\
&= \mathrm{P}(-z < Z \leqslant z) \\
&= \mathrm{P}(Z \leqslant z) - \mathrm{P}(Z \leqslant -z) \\
&= F_{_Z}(z) - F_{_Z}(-z),
\end{aligned}$$

where $F_{_Z}(-z) = 1 - F_{_Z}(z)$, by the symmetry if the pdf of Z. Consequently,

$$\mathrm{P}(|Z| < z) = 2F_{_Z}(z) - 1 \quad \text{for} \quad z > 0.$$

Suppose that $0 < \alpha < 1$ and let $z_{_{\alpha/2}}$ be the value of z for which $\mathrm{P}(|Z| < z) = 1 - \alpha$. We then have that $z_{_{\alpha/2}}$ satisfies the equation

$$F_{_Z}(z) = 1 - \frac{\alpha}{2}.$$

Thus,

$$z_{\alpha/2} = F_Z^{-1}\left(1 - \frac{\alpha}{2}\right), \tag{2.12}$$

where F_Z^{-1} denotes the inverse of the cdf of Z. Then, setting

$$\frac{\sqrt{n}\, b}{\sigma} = z_{\alpha/2},$$

we see from (2.11) that

$$\mathrm{P}\left(|\overline{X}_n - \mu| < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha,$$

which we can write as

$$\mathrm{P}\left(|\mu - \overline{X}_n| < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha,$$

or

$$\mathrm{P}\left(\overline{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha, \tag{2.13}$$

which says that the probability that the interval

$$\left(\overline{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \tag{2.14}$$

captures the parameter μ is $1-\alpha$. The interval in (2.14) is called the $100(1-\alpha)\%$ confidence interval for the mean, μ, based on the sample mean. Notice that this interval assumes that the variance, σ^2, is known, which is not the case in general. So, in practice it is not very useful (we will see later how to remedy this situation); however, it is a good example to illustrate the concept of a confidence interval.

For a more concrete example, let $\alpha = 0.05$. Then, to find $z_{\alpha/2}$ we may use the NORMINV function in MS Excel, which gives the inverse of the cumulative distribution function of normal random variable. The format for this function is

```
NORMINV(probability,mean,standard_dev)
```

In this case the probability is $1 - \frac{\alpha}{2} = 0.975$, the mean is 0, and the standard deviation is 1. Thus, according to (2.12), $z_{\alpha/2}$ is given by

$$\texttt{NORMINV}(0.975, 0, 1) \approx 1.959963985$$

or about 1.96.

In R, the inverse cdf for a normal random variable is the `qnorm` function whose format is

```
qnorm(probability, mean, standard_deviation).
```

Thus, in R, for $\alpha = 0.05$,

$$z_{\alpha/2} \approx \texttt{qnorm}(\texttt{0.975}, \texttt{0}, \texttt{1}) \approx 1.959964 \approx 1.96.$$

Hence the 95% confidence interval for the mean, μ, of a normal(μ, σ^2) distribution based on the sample mean, $\overline{X}_n$ is

$$\left(\overline{X}_n - 1.96\frac{\sigma}{\sqrt{n}}, \overline{X}_n + 1.96\frac{\sigma}{\sqrt{n}}\right), \tag{2.15}$$

provided that the variance, σ^2, of the distribution is known. Unfortunately, in most situations, σ^2 is an unknown parameter, so the formula for the confidence interval in (2.14) is not useful at all. In order to remedy this situation, in 1908, William Sealy Gosset, writing under the pseudonym of A. Student (see [Stu08]), proposed looking tat the statistic

$$T_n = \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}},$$

where S_n^2 is the sample variance defined by

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

Thus, we are replacing σ in

$$T_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

by the sample standard deviation, S_n, so that we only have one unknown parameter, μ, in the definition of T_n.

In order to find the sampling distribution of T_n, we will first need to determine the distribution of S_n^2, given that sampling is done form a normal(μ, σ^2) distribution. We will find the distribution of S_n^2 by first finding the distribution of the statistic

$$W_n = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2. \tag{2.16}$$

Starting with

$$(X_i - \mu)^2 = [(X_i - \overline{X}_n) + (\overline{X}_n - \mu)]^2,$$

we can derive the identity

$$\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}(X_i - \overline{X}_n)^2 + n(\overline{X}_n - \mu)^2, \tag{2.17}$$

where we have used the fact that

$$\sum_{i=1}^{n}(X_i - \overline{X}_n) = 0. \tag{2.18}$$

Next, dividing the equation in (2.17) by σ^2 and rearranging we obtain that

$$\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 = W_n + \left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}\right)^2, \tag{2.19}$$

where we have used the definition of the random variable W_n in (2.16). Observe that the random variable

$$\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2$$

has a $\chi^2(n)$ distribution since the X_is are iid normal(μ, σ^2) so that

$$\frac{X_i - \mu}{\sigma} \sim \text{normal}(0, 1),$$

and, consequently,

$$\left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2(1).$$

Similarly,

$$\left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}\right)^2 \sim \chi^2(1),$$

since $\overline{X}_n \sim$ normal$(\mu, \sigma^2/n)$. We can then re–write (2.19) as

$$Y = W_n + X, \tag{2.20}$$

where $Y \sim \chi^2(n)$ and $X \sim \chi^2(1)$. If we can prove that W_n and X are independent random variables, we will then be able to conclude that

$$W_n \sim \chi^2(n - 1). \tag{2.21}$$

To see why the assertion in (2.21) is true, if W_n and X are independent, note that from (2.20) we get that the mgf of Y is

$$M_Y(t) = M_{W_n}(t) \cdot M_X(t),$$

by independence of W_n and X. Consequently,

$$\begin{aligned} M_{W_n}(t) &= \frac{M_Y(t)}{M_X(t)} \\ &= \frac{\left(\dfrac{1}{1-2t}\right)^{n/2}}{\left(\dfrac{1}{1-2t}\right)^{1/2}} \\ &= \left(\frac{1}{1-2t}\right)^{(n-1)/2}, \end{aligned}$$

which is the mgf for a $\chi^2(n-1)$ random variable. Thus, in order to prove (2.21), it remains to prove that W_n and $\left(\dfrac{\overline{X}_n-\mu}{\sigma/\sqrt{n}}\right)^2$ are independent random variables.

2.3.4 Distribution of the Sample Variance from a Normal Distribution

In this section we will establish (2.21), which we now write as

$$\frac{(n-1)}{\sigma^2}S_n^2 \sim \chi^2(n-1). \tag{2.22}$$

As pointed out in the previous section, (2.22)will follow from (2.20) if we can prove that

$$\frac{1}{\sigma^2}\sum_{i=1}^n (X_i-\overline{X}_n)^2 \text{ and } \left(\frac{\overline{X}_n-\mu}{\sigma/\sqrt{n}}\right)^2 \text{ are independent.} \tag{2.23}$$

In turn, the claim in (2.23) will follow from the claim

$$\sum_{i=1}^n (X_i-\overline{X}_n)^2 \text{ and } \overline{X}_n \text{ are independent.} \tag{2.24}$$

The justification for the last assertion is given in the following two examples.

Example 2.3.9. Suppose that X and Y are independent independent random variables. Show that X and Y^2 are also independent.

Solution: Compute, for $x\in\mathbb{R}$ and $u\geqslant 0$,

$$\begin{aligned} \mathrm{P}(X\leqslant x, Y^2\leqslant u) &= \mathrm{P}(X\leqslant x, |Y|\leqslant \sqrt{u}) \\ &= \mathrm{P}(X\leqslant x, -\sqrt{u}\leqslant Y\leqslant \sqrt{u}) \\ &= \mathrm{P}(X\leqslant x)\cdot \mathrm{P}(-\sqrt{u}\leqslant Y\leqslant \sqrt{u}), \end{aligned}$$

since X and Y are assumed to be independent. Consequently,

$$\mathrm{P}(X \leqslant x, Y^2 \leqslant u) \quad = \quad \mathrm{P}(X \leqslant x) \cdot \mathrm{P}(Y^2 \leqslant u),$$

which shows that X and Y^2 are independent. □

Example 2.3.10. Let a and b be real numbers with $a \neq 0$. Suppose that X and Y are independent independent random variables. Show that X and $aY + b$ are also independent.

Solution: Compute, for $x \in \mathbb{R}$ and $w \in \mathbb{R}$,

$$\begin{aligned} \mathrm{P}(X \leqslant x, aY + b \leqslant w) \quad &= \quad \mathrm{P}(X \leqslant x, Y \leqslant \tfrac{w-b}{a}) \\ &= \quad \mathrm{P}(X \leqslant x) \cdot \mathrm{P}\left(Y \leqslant \frac{w-b}{a}\right), \end{aligned}$$

since X and Y are assumed to be independent. Consequently,

$$\mathrm{P}(X \leqslant x, aY + b \leqslant w) \quad = \quad \mathrm{P}(X \leqslant x) \cdot \mathrm{P}(aY + b \leqslant w),$$

which shows that X and $aY + b$ are independent. □

Hence, in order to prove (2.22) it suffices to show that the claim in (2.24) is true. To prove this last claim, observe that from (2.18) we get

$$(X_1 - \overline{X}_n) = -\sum_{i=2}^{n} (X_i - \overline{X}_n,$$

so that, squaring on both sides,

$$(X_1 - \overline{X}_n)^2 = \left(\sum_{i=2}^{n} (X_i - \overline{X}_n\right)^2.$$

Hence, the random variable

$$\sum_{i=1}^{n} (X_i - \overline{X}_n)^2 = \left(\sum_{i=2}^{n} (X_i - \overline{X}_n\right)^2 + \sum_{i=2}^{n} (X_i - \overline{X}_n)^2$$

is a function of the random vector

$$(X_2 - \overline{X}_n, X_3 - \overline{X}_n, \ldots, X_n - \overline{X}_n).$$

Consequently, the claim in (2.24) will be proved if we can prove that

$$\overline{X}_n \text{ and } (X_2 - \overline{X}_n, X_3 - \overline{X}_n, \ldots, X_n - \overline{X}_n) \text{ are independent.} \qquad (2.25)$$

The proof of the claim in (2.25) relies on the assumption that the random variables $X_1, X_2, \ldots, X_n$ are iid normal random variables. We illustrate this in the following example for the spacial case in which $n = 2$ and $X_1, X_2 \sim$ normal$(0, 1)$. Observe that, in view of Example 2.3.10, by considering

$$\frac{X_i - \mu}{\sigma} \quad \text{for} \quad i = 1, 2, \ldots, n,$$

we may assume from the outset that $X_1, X_2, \ldots, X_n$ are iid normal$(0, 1)$ random variables.

Example 2.3.11. Let X_1 and X_2 denote independent normal$(0, 1)$ random variables. Define

$$U = \frac{X_1 + X_2}{2} \quad \text{and} \quad V = \frac{X_2 - X_1}{2}.$$

Show that U and V independent random variables.

Solution: Compute the cdf of U and V:

$$\begin{aligned} F_{(U,V)}(u, v) &= \mathrm{P}(U \leqslant u, V \leqslant v) \\ &= \mathrm{P}\left(\frac{X_1 + X_2}{2} \leqslant u, \frac{X_2 - X_1}{2} \leqslant v\right) \\ &= \mathrm{P}\left(X_1 + X_2 \leqslant 2u, X_2 - X_1 \leqslant 2v\right) \\ &= \iint_{R_{u,v}} f_{(X_1,X_2)}(x_1, x_2)\mathrm{d}x_1\mathrm{d}x_2, \end{aligned}$$

where $R_{u,v}$ is the region in $\mathbb{R}^2$ defined by

$$R_{u,v} = \{(x_1, x_2) \in \mathbb{R}^2 \mid x_1 + x_2 \leqslant 2u, x_2 - x_1 \leqslant 2v\},$$

and $f_{(X_1,X_2)}$ is the joint pdf of X_1 and X_2:

$$f_{(X_1,X_2)}(x_1, x_2) = \frac{1}{2\pi} e^{-(x_1^2 + x_2^2)/2} \quad \text{for all} \quad (x_1, x_2) \in \mathbb{R}^2,$$

where we have used the assumption that X_1 and X_2 are independent normal$(0, 1)$ random variables.

Next, make the change of variables

$$r = x_1 + x_2 \quad \text{and} \quad w = x_2 - x_1,$$

so that

$$x_1 = \frac{r - w}{2} \quad \text{and} \quad x_2 = \frac{r + w}{2},$$

and therefore

$$x_1^2 + x_2^2 = \frac{1}{2}(r^2 + w^2).$$

Thus, by the change of variables formula,

$$F_{(U,V)}(u,v) \;=\; \int_{-\infty}^{2u}\int_{-\infty}^{2v} \frac{1}{2\pi} e^{-(r^2+w^2)/4} \left|\frac{\partial(x_1,x_2)}{\partial(r,w)}\right| \,\mathrm{d}w\mathrm{d}r,$$

where

$$\frac{\partial(x_1,x_2)}{\partial(r,w)} \;=\; \det\begin{pmatrix} 1/2 & -1/2 \\ 1/2 & 1/2 \end{pmatrix} = \frac{1}{2}.$$

Thus,

$$F_{(U,V)}(u,v) \;=\; \frac{1}{4\pi}\int_{-\infty}^{2u}\int_{-\infty}^{2v} e^{-r^2/4}\cdot e^{-w^2/4}\,\mathrm{d}w\mathrm{d}r,$$

which we can write as

$$F_{(U,V)}(u,v) \;=\; \frac{1}{2\sqrt{\pi}}\int_{-\infty}^{2u} e^{-r^2/4}\,\mathrm{d}r\cdot\frac{1}{2\sqrt{\pi}}\int_{-\infty}^{2v} e^{-w^2/4}\,\mathrm{d}w.$$

Taking partial derivatives with respect to u and v yields

$$f_{(U,V)}(u,v) \;=\; \frac{1}{\sqrt{\pi}}\,e^{-u^2}\cdot\frac{1}{\sqrt{\pi}}\,e^{-v^2},$$

where we have used the fundamental theorem of calculus and the chain rule. Thus, the joint pdf of U and V is the product of the two marginal pdfs

$$f_U(u) \;=\; \frac{1}{\sqrt{\pi}}\,e^{-u^2} \quad \text{for } -\infty < u < \infty,$$

and

$$f_V(v) \;=\; \frac{1}{\sqrt{\pi}}\,e^{-v^2} \quad \text{for } -\infty < v < \infty.$$

Hence, U and V are independent random variables. □

To prove in general that if $X_1, X_2, \ldots, X_n$ is a random sample from a normal$(0,1)$ distribution, then

$$\overline{X}_n \quad \text{and} \quad (X_2-\overline{X}_n, X_3-\overline{X}_n, \ldots, X_n-\overline{X}_n) \quad \text{are independent,}$$

we may proceed as follows. Denote the random vector

$$(X_2-\overline{X}_n, X_3-\overline{X}_n, \ldots, X_n-\overline{X}_n)$$

by Y, and compute the cdf of $\overline{X}_n$ and Y:

$$\begin{aligned} F_{(\overline{X}_n,Y)}(u, v_2, \quad v_3 \quad, \ldots, v_n) & \\ &= \mathrm{P}(\overline{X}_n \leqslant u, X_2-\overline{X}_n \leqslant v_2, X_3-\overline{X}_n \leqslant v_3, \ldots, X_n-\overline{X}_n \leqslant v_n) \\ &= \iint\cdots\int_{R_{u,v_2,v_3,\ldots,v_n}} f_{(X_1,X_2,\ldots,X_n)}(x_1,x_2,\ldots,x_n)\,\mathrm{d}x_1\mathrm{d}x_2\cdots\mathrm{d}x_n, \end{aligned}$$

where

$$R_{u,v_2,\ldots,v_n} = \{(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n \mid \overline{x} \leqslant u, x_2 - \overline{x} \leqslant v_2, \ldots, x_n - \overline{x} \leqslant v_n\},$$

for $\overline{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n}$, and the joint pdf of $X_1, X_2, \ldots, X_n$ is

$$f_{(X_1,X_2,\ldots,X_n)}(x_1, x_2, \ldots x_n) = \frac{1}{(2\pi)^{n/2}}\, e^{-(\sum_{i=1}^{n} x_i^2)/2} \quad \text{for all } (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n,$$

since $X_1, X_2, \ldots, X_n$ are iid normal$(0, 1)$.

Next, make the change of variables

$$\begin{array}{rcl} y_1 & = & \overline{x} \\ y_2 & = & x_2 - \overline{x} \\ y_3 & = & x_3 - \overline{x} \\ & \vdots & \\ y_n & = & x_n - \overline{x}. \end{array}$$

so that

$$\begin{array}{rcl} x_1 & = & y_1 - \displaystyle\sum_{i=2}^{n} y_i \\ x_2 & = & y_1 + y_2 \\ x_3 & = & y_1 + y_3 \\ & \vdots & \\ x_n & = & y_1 + y_n, \end{array}$$

and therefore

$$\begin{array}{rcl} \displaystyle\sum_{i=1}^{n} x_i^2 & = & \left(y_1 - \displaystyle\sum_{i=2}^{n} y_i\right)^2 + \displaystyle\sum_{i=2}^{n} (y_1 + y_i)^2 \\ & = & n y_1^2 + \left(\displaystyle\sum_{i=2}^{n} y_i\right)^2 + \displaystyle\sum_{i=2}^{n} y_i^2 \\ & = & n y_1^2 + C(y_2, y_3, \ldots, y_n), \end{array}$$

where we have set

$$C(y_2, y_3, \ldots, y_n) = \left(\sum_{i=2}^{n} y_i\right)^2 + \sum_{i=2}^{n} y_i^2.$$

Thus, by the change of variables formula,

$$
\begin{aligned}
F_{(\overline{X}_n,Y)} \quad & (\quad u, v_2, \ldots, v_n) \\
&= \int_{-\infty}^{u} \int_{-\infty}^{v_1} \cdots \int_{-\infty}^{v_n} \frac{e^{-(ny_1^2 + C(y_2,\ldots,y_n)/2}}{(2\pi)^{n/2}} \left| \frac{\partial(x_1, x_2, \ldots, x_n)}{\partial(y_1, y_2, \ldots, y_n)} \right| \, \mathrm{d}y_n \cdots \mathrm{d}y_1,
\end{aligned}
$$

where

$$
\frac{\partial(x_1, x_2, \ldots, x_n)}{\partial(y_1, y_2, \ldots y_n)} = \det \begin{pmatrix} 1 & -1 & -1 & \cdots & -1 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & 1 \end{pmatrix}.
$$

In order to compute this determinant observe that

$$
\begin{array}{rcl}
ny_1 & = & x_1 + x_2 + x_3 + \ldots x_n \\
ny_2 & = & -x_1 + (n-1)x_2 - x_3 - \ldots - x_n \\
ny_3 & = & -x_1 - x_2 + (n-1)x_3 - \ldots - x_n \\
\vdots & & \\
ny_n & = & -x_1 - x_2 - x_3 - \ldots + (n-1)x_n
\end{array},
$$

which can be written in matrix form as

$$
n \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = A \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix},
$$

where A is the $n \times n$ matrix

$$
A = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ -1 & (n-1) & -1 & \cdots & -1 \\ -1 & -1 & (n-1) & \cdots & -1 \\ \vdots & \vdots & \vdots & \vdots & \\ -1 & -1 & -1 & \cdots & (n-1) \end{pmatrix},
$$

whose determinant is

$$
\det A = \det \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 0 & n & 0 & \cdots & 0 \\ 0 & 0 & n & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \\ 0 & 0 & 0 & \cdots & n \end{pmatrix} = n^{n-1}.
$$

Thus, since

$$A\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} = nA^{-1}\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix},$$

it follows that

$$\frac{\partial(x_1, x_2, \ldots, x_n)}{\partial(y_1, y_2, \ldots y_n)} = \det(nA^{-1}) = n^n \cdot \frac{1}{n^{n-1}} = n.$$

Consequently,

$$\begin{aligned} F_{(\overline{X}_n, Y)} & (u, v_2, \ldots, v_n) \\ &= \int_{-\infty}^{u} \int_{-\infty}^{v_1} \cdots \int_{-\infty}^{v_n} \frac{n\ e^{-ny_1^2/2}\ e^{-C(y_2,\ldots,y_n)/2}}{(2\pi)^{n/2}}\ \mathrm{d}y_n \cdots \mathrm{d}y_1, \end{aligned}$$

which can be written as

$$\begin{aligned} F_{(\overline{X}_n, Y)} & (u, v_2, \ldots, v_n) \\ &= \int_{-\infty}^{u} \frac{n\ e^{-ny_1^2/2}}{\sqrt{2\pi}}\ \mathrm{d}y_1 \cdot \int_{-\infty}^{v_1} \cdots \int_{-\infty}^{v_n} \frac{e^{-C(y_2,\ldots,y_n)/2}}{(2\pi)^{(n-1)/2}}\ \mathrm{d}y_n \cdots \mathrm{d}y_2. \end{aligned}$$

Observe that

$$\int_{-\infty}^{u} \frac{n\ e^{-ny_1^2/2}}{\sqrt{2\pi}}\ \mathrm{d}y_1$$

is the cdf of a normal$(0, 1/n)$ random variable, which is the distribution of $\overline{X}_n$. Therefore

$$F_{(\overline{X}_n, Y)}(u, v_2, \ldots, v_n) = F_{\overline{X}_n}(u) \cdot \int_{-\infty}^{v_1} \cdots \int_{-\infty}^{v_n} \frac{e^{-C(y_2,\ldots,y_n)/2}}{(2\pi)^{(n-1)/2}}\ \mathrm{d}y_n \cdots \mathrm{d}y_2,$$

which shows that $\overline{X}_n$ and the random vector

$$Y = (X_2 - \overline{X}_n, X_3 - \overline{X}_n, \ldots, X_n - \overline{X}_n)$$

are independent. Hence we have established (2.22); that is,

$$\frac{(n-1)}{\sigma^2} S_n^2 \sim \chi^2(n-1).$$

2.3.5 The Distribution of T_n

We are now in a position to determine the sampling distribution of the statistic

$$T_n = \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}}, \tag{2.26}$$

where $\overline{X}_n$ and S_n^2 are the sample mean and variance, respectively, based on a random sample of size n taken from a normal(μ, σ^2) distribution.

We begin by re–writing the expression for T_n in (2.26) as

$$T_n = \frac{\dfrac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}}{\dfrac{S_n}{\sigma}}, \tag{2.27}$$

and observing that

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim \text{normal}(0,1).$$

Furthermore,

$$\frac{S_n}{\sigma} = \sqrt{\frac{S_n^2}{\sigma^2}} = \sqrt{\frac{V_n}{n-1}},$$

where

$$V_n = \frac{n-1}{\sigma^2} S_n^2,$$

which has a $\chi^2(n-1)$ distribution, according to (2.22). It then follows from (2.27) that

$$T_n = \frac{Z_n}{\sqrt{\dfrac{V_n}{n-1}}},$$

where Z_n is a standard normal random variable, and V_n has a χ^2 distribution with $n-1$ degrees of freedom. Furthermore, by (2.23), Z_n and V_n are independent. Consequently, using the result in Example 2.3.8, the statistic T_n defined in (2.26) has a t distribution with $n-1$ degrees of freedom; that is,

$$\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \sim t(n-1). \tag{2.28}$$

Notice that the distribution on the right–hand side of (2.28) is independent of the parameters μ and σ^2; we can can therefore obtain a confidence interval for the mean of of a normal(μ, σ^2) distribution based on the sample mean and variance calculated from a random sample of size n by determining a value $t_{\alpha/2}$ such that

$$\text{P}(|T_n| < t_{\alpha/2}) = 1 - \alpha.$$

We then have that

$$\text{P}\left(\frac{|\overline{X}_n - \mu|}{S_n/\sqrt{n}} < t_{\alpha/2}\right) = 1 - \alpha,$$

or

$$\text{P}\left(|\mu - \overline{X}_n| < t_{\alpha/2}\frac{S_n}{\sqrt{n}}\right) = 1 - \alpha,$$

or

$$\text{P}\left(\overline{X}_n - t_{\alpha/2}\frac{S_n}{\sqrt{n}} < \mu < \overline{X}_n + t_{\alpha/2}\frac{S_n}{\sqrt{n}}\right) = 1 - \alpha.$$

We have therefore obtained a $100(1-\alpha)\%$ confidence interval for the mean of a normal(μ, σ^2) distribution based on the sample mean and variance of a random sample of size n from that distribution; namely,

$$\left(\overline{X}_n - t_{\alpha/2}\frac{S_n}{\sqrt{n}}, \overline{X}_n + t_{\alpha/2}\frac{S_n}{\sqrt{n}}\right). \tag{2.29}$$

To find the value for $z_{\alpha/2}$ in (2.29) we use the fact that the pdf for the t distribution is symmetric about the vertical line at 0 (or even) to obtain that

$$\begin{aligned}
\mathrm{P}(|T_n| < t) &= \mathrm{P}(-t < T_n < t) \\
&= \mathrm{P}(-t < T_n \leqslant t) \\
&= \mathrm{P}(T_n \leqslant t) - \mathrm{P}(T_n \leqslant -t) \\
&= F_{T_n}(t) - F_{T_n}(-t),
\end{aligned}$$

where we have used the fact that T_n is a continuous random variable. Now, by the symmetry if the pdf of T_n $F_{T_n}(-t) = 1 - F_{T_n}(t)$. Thus,

$$\mathrm{P}(|T_n| < t) = 2F_{T_n}(t) - 1 \quad \text{for} \quad t > 0.$$

So, to find $t_{\alpha/2}$ we need to solve

$$F_{T_n}(t) = 1 - \frac{\alpha}{2}.$$

We therefore get that

$$t_{\alpha/2} = F_{T_n}^{-1}\left(1 - \frac{\alpha}{2}\right), \tag{2.30}$$

where $F_{T_n}^{-1}$ denotes the inverse of the cdf of T_n.

Example 2.3.12. Give a 95% confidence interval for the mean of a normal distribution based on the sample mean and variance computed from a sample of size $n = 20$.

> ***Solution:*** In this case, $\alpha = 0.5$ and $T_n \sim t(19)$.
>
> To find $t_{\alpha/2}$ we may use the TINV function in MS Excel, which gives the inverse of the two–tailed cumulative distribution function of random variable with a t distribution. That is, the inverse of the function
>
> $$\mathrm{P}(|T_n| > t) \quad \text{for } t > 0.$$
>
> The format for this function is
>
> ```
> TINV(probability,degrees_freedom)
> ```

In this case the probability of the two tails is $\alpha = 0.05$ and the number of degrees of freedom is 19. Thus, according to (2.30), $t_{\alpha/2}$ is given by

$$\texttt{TINV}(0.05, 19) \approx 2.09,$$

where we have used 0.05 because `TINV` in MS Excel gives two-tailed probability distribution values.

In R, the inverse cdf for a random variable with a t distribution is `qt` function whose format is

```
qt(probability, df).
```

Thus, in R, for $\alpha = 0.05$,

$$t_{\alpha/2} \approx \texttt{qt}(\texttt{0.975}, \texttt{19}) \approx 2.09.$$

Hence the 95% confidence interval for the mean, μ, of a normal(μ, σ^2) distribution based on the sample mean, $\overline{X}_n$, and the sample variance, S_n^2, is

$$\left(\overline{X}_n - 2.09\frac{S_n}{\sqrt{n}}, \overline{X}_n + 2.09\frac{S_n}{\sqrt{n}}\right), \tag{2.31}$$

where $n = 20$. □

Example 2.3.13. Obtain a 95% confidence interval for the average number of popcorn kernels in a 1/4–cup based on the data in Table 1.2 on page 8.

Solution: Assume that the the underlying distribution of the count of kernels in 1/4–cup is normal(μ, σ^2), where μ is the parameter we are trying to estimate and σ^2 is unknown.

The sample mean, $\overline{X}_n$, based on the sample in Table 1.2 on page 8 is $\overline{X}_n \approx 342$. The sample standard deviation is $S_n \approx 35$. Thus, using the formula in (2.31) we get that

$$(326, 358)$$

is a 95% confidence interval for the average number of kernels in one quarter cup. □

Chapter 3

Hypothesis Testing

In the examples of the previous chapters, we have assumed certain underlying distributions which depended on a parameter or more. Based on that assumption, we have obtained estimates for a parameter through calculations made with the values of a random sample; this process yielded statistics which can be used as estimators for the parameters. Assuming an underlying distribution allowed as to determine the sampling distribution for the statistic; in turn, knowledge of the sampling distribution permitted the calculations of probabilities that estimates are within certain range of a parameter.

For instance, the confidence interval estimate for the average number of popcorn kernels in one–quarter cup presented Example 2.3.13 on page 42 relied on the assumption that the number of kernels in one–quarter cup follows a normal(μ, σ^2) distribution. There is nothing sacred about the normality assumption. The assumption was made because a lot is known about the normal distribution and therefore the assumption was a convenient one to use in order to illustrate the concept of a confidence interval. In fact, cursory study of the data in Table 1.2 on page 8 reveals that the shape of the distribution might not be as bell–shaped as one would hope; see the histogram of the data in Figure 3.0.1 on page 44, where N denotes the number of kernels in one–quarter cup. Nevertheless, the hope is that, if a larger sample of one–quarter cups of kernels is collected, then we would expect to see the numbers bunching up around some value close to the true mean count of kernels in one–quarter cup.

The preceding discussion underscores the fact that assumptions, which are made to facilitate the process of parameter estimation, are also there to be questioned. In particular, the following question needs to be addressed: Does the assumed underlying distribution truly reflects what is going on in the real problem being studied? The process of questioning assumptions falls under the realm of *Hypothesis Testing* in statistical inference. In this Chapter we discuss how this can done. We begin with the example of determining whether the number of unpopped kernels in one–quarter cup follows a Poisson distribution.

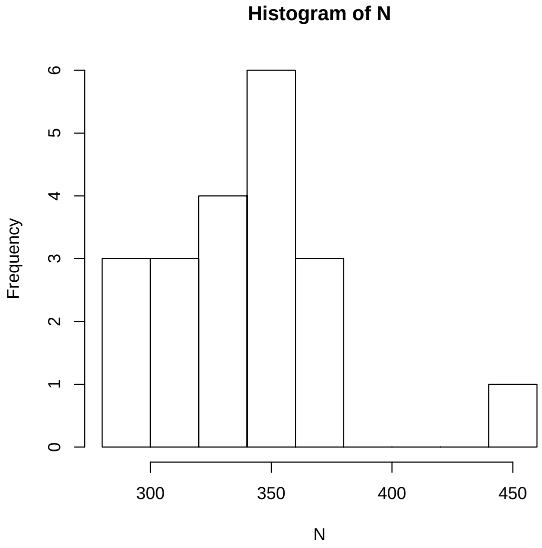

Figure 3.0.1: Histogram of Data in Table 1.2

3.1 Chi–Square Goodness of Fit Test

We begin with the example of determining whether the counts of unpopped kernels in one–quarter cup shown in Table 1.1 on page 6 can be accounted for by a Poisson(λ) distribution, where λ is the average number of unpopped kernels in one quarter cup. A point estimate for λ is then given by the average of the values in the table; namely, 56. Before we proceed any further, I must point out that the reason that I am assuming a Poisson model for the data in Table 1.1 is merely for the sake of illustration of the Chi–Square test that we'll be discussing in this section. However, a motivation for the use of the Poisson model is that a Poisson random variable is a limit of binomial random variables as n tends to infinity under the condition that $np = \lambda$ remains constant (see your solution to Problem 5 in Assignment 2). However, this line of reasoning would be justified if the probability that a given kernel will not pop is small, which is not really justified in this situation since, by the result in Example 2.2.2 on page 17, a 95% confidence interval for p is $(0.109, 0.183)$. In addition, a look at the histogram of the data in Table 1.1, shown in Figure 3.1.2, reveals that the shape of that distribution for the number of unpopped kernels is far from being Poisson. The reason for this is that the right tail of a Poisson distribution should be thinner than that of the distribution shown in Figure 3.1.2.

Moreover, calculation of the sample variance for the data in Table 1.1 yields 1810, which is way too far from the sample mean of 56. Recall that the mean and variance of a Poisson(λ) distribution are both equal to λ.

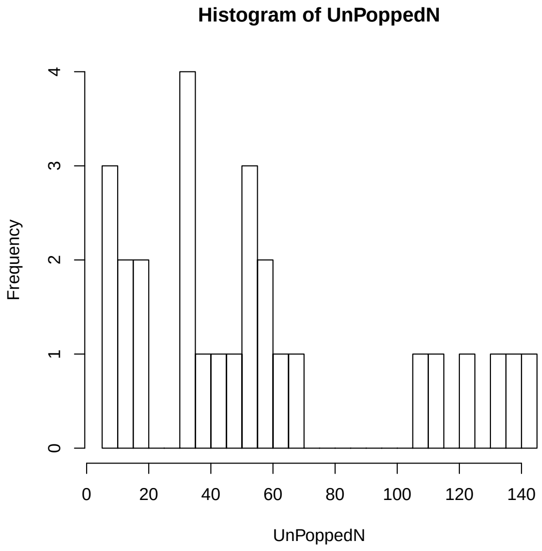

Figure 3.1.2: Histogram of data for unpopped kernels in Table 1.1

Despite all of the objections to the applicability of the Poisson model to the unpopped kernels data listed previously, I will proceed with the Poisson assumption in order to illustrate the Chi–Square Goodness of Fit method, which provides a quantitative way to reject the Poisson model with confidence.

Thus, assuming that the Poisson model is the mechanism behind the observations of unpopped kernels, we may compute the probabilities of observing certain counts of unpopped kernels by using the pmf

$$p_{_X}(k) = \frac{\lambda^k}{k!}\, e^{-\lambda} \quad \text{for } k = 0, 1, 2, \ldots,$$

and 0 elsewhere, where λ is taken to be the estimated value of 56. Hence, the probability that we observe counts between 0 and 50 is

$$\mathrm{P}(0 \leqslant X \leqslant 50) = \sum_{k=0}^{50} p_{_X}(k) \approx 0.2343;$$

between 51 and 55:

$$\mathrm{P}(51 \leqslant X \leqslant 55) = \sum_{k=51}^{55} p_{_X}(k) \approx 0.2479;$$

between 56 and 60:

$$\mathrm{P}(56 \leqslant X \leqslant 60) = \sum_{k=56}^{60} p_{_X}(k) \approx 0.2487;$$

and 61 and above:

$$\mathrm{P}(X \geqslant 61) = \sum_{k=61}^{\infty} p_{X}(k) \approx 0.2691.$$

We have therefore divided the range of possible observations into categories, and the probabilities computed above give the likelihood a given observation (the count of unpopped kernels, in this case) will fall in a given category, assuming that a Poisson model is driving the process. Using these probabilities, we can predict how many observations out of the 27 will fall, on average, in each category. If the probability that a count will fall in category i is p_i, and n is the total number of observations (in this case, $n = 27$), then the predicted number of counts in category i is

$$E_i = np_i.$$

Table 3.1 shows the predicted values in each category as well as the actual (observed) counts.

Category (i)	Counts Range	p_i	Predicted Counts	Observed Counts
1	$0 \leqslant X \leqslant 50$	0.2343	6	14
2	$51 \leqslant X \leqslant 55$	0.2479	7	3
3	$56 \leqslant X \leqslant 60$	0.2487	7	2
4	$X \geqslant 61$	0.2691	7	8

Table 3.1: Counts Predicted by the Poisson Model

The last column in Table 3.1 shows that actual observed counts based on the data in Table 1.1 on page 6. Are the large discrepancies between the observed and predicted counts in the first three categories in Table 3.1 enough evidence for us to dismiss the Poisson hypothesis? One of the goals of this chapter is to answer this question with confidence. We will need to find a way to measure the discrepancy that will allow us to make statements based on probability calculations. A measure of the discrepancy between the values predicted by an assumed probability model and the values that are actually observed in the data was introduced by Karl Pearson in 1900, [Pla83]. In order to motivate the Pearson's statistic, we first present an example involving the multinomial distribution.

3.1.1 The Multinomial Distribution

Consider the general situation of k categories whose counts are given by random variables $X_1, X_1, \ldots, X_k$. Assume that there is a total n of observations so that

$$X_1 + X_2 + \cdots + X_k = n. \tag{3.1}$$

We assume that the probability that a count is going to fall in category i is p_i for $i = 1, 2, \ldots, k$. Assume also that the categories are mutually exclusive and exhaustive so that

$$p_1 + p_2 + \cdots + p_k = 1. \tag{3.2}$$

Then, the distribution of the random vector

$$\mathbf{X} = (X_1, X_2, \ldots, X_k) \tag{3.3}$$

is multinomial so that the joint pmf of the random variables $X_1, X_1, \ldots, X_k$, given that $X_1 + X_2 + \cdots + X_k = n$, is

$$p_{(X_1,X_2,\ldots,X_k)}(n_1, n_2, \ldots, n_k) = \begin{cases} \dfrac{n!}{n_1!n_2!\cdots n_k!} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k} & \text{if } \sum_{i=1}^{k} n_k = n; \\ \\ 0 & \text{otherwise.} \end{cases} \tag{3.4}$$

We first show that each X_i has marginal distribution which is binomial(n, p_i), so that

$$E(X_i) = np_i \quad \text{for all } i = 1, 2, \ldots, k,$$

and

$$\text{var}((X_i)) = np_i(1 - p_i) \quad \text{for all } i = 1, 2, \ldots, k.$$

Note that the $X_1, X_2, \ldots, X_k$ are not independent because of the relation in (3.1). In fact, it can be shown that

$$\text{cov}(X_i, X_j) = -np_jp_j \quad \text{for} \quad i \neq j.$$

We will first establish that the marginal distribution of X_i is binomial. We will show it for X_1 in the following example. The proof for the other variables is similar. In the proof, though, we will need the following extension of the binomial theorem known as the multinomial theorem [CB01, Theorem 4.6.4, p. 181].

Theorem 3.1.1 (Multinomial Theorem)**.** Let $n, n_1, n_2, \ldots, n_k$ denote non–negative integers, and $a_1, a_2, \ldots, a_k$ be real numbers. Then,

$$(a_1 + a_2 + \cdots + a_k)^n = \sum_{n_1+n_2+\cdots+n_k=n} \frac{n!}{n_1!n_2!\cdots n_k!}\, a_1^{n_1} a_2^{n_2} \cdots a_k^{n_k},$$

where the sum is take over all k–tuples of nonnegative integers, $n_1, n_2, \ldots, n_k$ which add up to n.

Remark 3.1.2. Note that when $k = 2$ in Theorem 3.1.1 we recover the binomial theorem,

Example 3.1.3. Let $(X_1, X_2, \ldots, X_k)$ have a multinomil distribution with parameters $n, p_1, p_2, \ldots, p_k$. Then, the marginal distribution of X_1 is binomial(n, p_1).

Solution: The marginal distribution of X_1 has pmf

$$p_{X_1}(n_1) = \sum_{\substack{n_2,n_3,\ldots,n_k \\ n_2+n_3+\ldots+n_k=n-n_1}} \frac{n!}{n_1!n_2!\cdots n_k!} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k},$$

where the summation is taken over all nonnegative, integer values of $n_2, n_3, \ldots, n_k$ which add up to $n - n_1$. We then have that

$$\begin{aligned} p_{X_1}(n_1) &= \frac{p_1^{n_1}}{n_1!} \sum_{\substack{n_2,n_3,\ldots,n_k \\ n_2+n_3+\ldots+n_k=n-n_1}} \frac{n!}{n_2!\cdots n_k!} p_2^{n_2} \cdots p_k^{n_k} \\ &= \frac{p_1^{n_1}}{n_1!} \frac{n!}{(n-n_1)!} \sum_{\substack{n_2,n_3,\ldots,n_k \\ n_2+n_3+\ldots+n_k=n-n_1}} \frac{(n-n_1)!}{n_2!\cdots n_k!} p_2^{n_2} \cdots p_k^{n_k} \\ &= \binom{n}{n_1} p_1^{n_1} (p_2 + p_3 + \cdots + p_k)^{n-n_1}, \end{aligned}$$

where we have applied the multinomial theorem (Theorem 3.1.1). Using (3.2) we then obtain that

$$p_{X_1}(n_1) = \binom{n}{n_1} p_1^{n_1} (1 - p_1)^{n-n_1},$$

which is the pmf of a binomial(n, p_1) distribution. □

3.1.2 The Pearson Chi-Square Statistic

We first consider the example of a multinomial random vector (X_1, X_2) with parameters n, p_1, p_2; in other words, there are only two categories and the counts in each category are binomial(n, p_i) for $i = 1, 2$, with $X_1 + X_2 = n$. We consider the situation when n is very large. In this case, the random variable

$$Z = \frac{X_1 - np_1}{\sqrt{np_1(1-p_1)}}$$

has an approximate normal$(0, 1)$ distribution for large values of n. Consequently, for large values of n,

$$Z^2 = \frac{(X_1 - np_1)^2}{np_1(1-p_1)}$$

has an approximate $\chi^2(1)$ distribution.

Note that we can write

$$\begin{aligned} Z^2 &= \frac{(X_1 - np_1)^2(1-p_1) + (X_1 - np_1)^2 p_1}{np_1(1-p_1)} \\ &= \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_1 - np_1)^2}{n(1-p_1)} \\ &= \frac{(X_1 - np_1)^2}{np_1} + \frac{(n - X_2 - np_1)^2}{n(1-p_1)} \\ &= \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - n(1-p_1))^2}{n(1-p_1)} \\ &= \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2}. \end{aligned}$$

We have therefore proved that, for large values of n, the random variable

$$Q = \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2}$$

has an approximate $\chi^2(1)$ distribution.

The random variable Q is the Pearson Chi–Square statistic for $k = 2$.

Theorem 3.1.4 (Pearson Chi–Square Statistic)**.** Let $(X_1, X_2, \ldots, X_k)$ be a random vector with a multinomial$(n, p_1, \ldots, p_k)$ distribution. The random variable

$$Q = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} \tag{3.5}$$

has an approximate $\chi^2(k-1)$ distribution for large values of n. If the p_is are computed assuming an underlying distribution with c unknown parameters, then the number of degrees of freedom in the chi–square distribution for Q get reduced by c. In other words

$$Q \sim \chi^2(k - c - 1) \quad \text{for large values of } n.$$

Theorem 3.1.4, the proof of which is relegated to Appendix A on page 87 in these notes, forms the basis for the Chi–Square Goodness of Fit Test. Examples of the application of this result will be given in subsequent sections.

3.1.3 Goodness of Fit Test

We now go back to the analysis of the data portrayed in Table 3.1 on page 46. Letting X_1, X_2, X_3, X_4 denote the observed counts in the fourth column of the table, we compute the value of the Pearson Chi–Square statistic according to (3.5) to be

$$\widehat{Q} = \sum_{i=1}^{4} \frac{(X_i - np_i)^2}{np_i} \approx 16.67,$$

where, in this case, $n = 27$ and the p_is are given in the third column of Table 3.1. This is the measure of how far the observed counts are from the counts predicted by the Poisson assumption. How significant is the number 16.67? Is it a big number or not? More importantly, how probable would a value like 16.67, or higher, be if the Poisson assumption is true? The last question is one we could answer approximately by using Pearson's Theorem 3.1.4. Since, $Q \sim \chi^2(2)$ in this case, the answer to the last question is

$$p = \mathrm{P}(Q > 16.67) \approx 0.0002,$$

or 0.02%, less than 1%, which is a very small probability. Thus, the chances of observing the counts in the fourth column of Table 3.1 on page 46, under the assumption that the Poisson hypothesis is true, are very small. The fact that we did observe those counts, and the counts came from observations recorded in Table 1.1 on page 6 suggest that it is highly unlikely that the counts of unpopped kernels in that table follow a Poisson distribution. We are therefore justified in rejecting the Poisson hypothesis on the basis on not enough statistical support provided by the data.

3.2 The Language and Logic of Hypothesis Tests

The argument that we followed in the example presented in the previous section is typical of hypothesis tests.

- **Postulate a Null Hypothesis.** First, we postulated a hypothesis that purports to explain patters observed in data. This hypothesis is the one to be tested against the data. In the example at hand, we want to test whether the counts of unpopped kernels in a one–quarter cup follow a Poisson distribution. The Poisson assumption was used to determine probabilities that observations will fall into one of four categories. We can use these values to formulate a null hypothesis, H_o, in terms of the the predicted probabilities; we write

  $$\mathrm{H}_o: \quad p_1 = 0.2343, \; p_2 = 0.2479, \; p_3 = 0.2487, \; p_4 = 0.2691.$$

  Based on probabilities in H_o, we compute the expected counts in each categories

  $$E_i = np_i \quad \text{for} \quad i = 1, 2, 3, 4.$$

  Remark 3.2.1 (Why were the categories chosen the way we chose them?). Pearson's Theorem 3.1.4 gives an approximate distribution for the Chi–Square statistic in (3.5) for large values of n. A rule of thumb to justify the use the Chi–Square approximation to distribution of the Chi–Square statistic, Q, is to make sure that the expected count in each category is 5 or more. That is why we divided the range of counts in Table 1.1 on page 6 into the four categories shown in Table 3.1 on page 46.

- **Compute a Test Statistic.** In the example of the previous section, we computed the Pearson Chi–Square statistic, $\widehat{Q}$, which measures how far the observed counts, X_1, X_2, X_3, X_4, are from the expected counts, E_1, E_2, E_3, E_4:
$$\widehat{Q} = \sum_{i=1}^{4} \frac{(X_i - E_i)^2}{E_i}.$$
According to Pearson's Theorem, the random variable Q given by (3.5); namely,
$$Q = \sum_{i=1}^{4} \frac{(X_i - np_i)^2}{np_i}$$
has an approximate $\chi^2(4-1-1)$ distribution in this case.

- **Compute or approximate a p–value.** A p–value for a test is the probability that the test statistic will attain the computed value, or more extreme ones, under the assumption that the null hypothesis is true. In the example of the previous section, we used the fact that Q has an approximate $\chi^2(2)$ distribution to compute
$$p\text{–value} = \mathrm{P}(Q \geqslant \widehat{Q}).$$

- **Make a decision.** Either we reject or we don't reject the null hypothesis. The criterion for rejection is some threshold, α with $0 < \alpha < 1$, usually some small probability, say $\alpha = 0.01$ or $\alpha = 0.05$.

  We reject H_o if p–value $< \alpha$; otherwise we don't reject H_o.

  We usually refer to α as a level of significance for the test. If p–value $< \alpha$ we say that we reject H_o at the level of significance α.

  In the example of the previous section
$$p\text{–value} \approx 0.0002 < 0.01;$$
  Thus, we reject the Poisson model as an explanation of the distribution for the counts of unpopped kernels in Table 1.1 on page 6 at the significance level of $\alpha = 1\%$.

Example 3.2.2 (Testing a binomial model)**.** We have seen how to use a chi–square goodness of fit test to determine that the Poisson model for the distribution of counts of unpopped kernels in Table 1.1 on page 6 is not supported by the data in the table. A more appropriate model would be a binomial model. In this case we have two unknown parameters: the mean number of kernels,

n, in one–quarter cup, and the probability, p, that a given kernel will not pop. We have estimated n independently using the data in Table 1.2 on page 8 to be $\widehat{n} = 342$ according to the result in Example 2.3.13 on page 42. In order to estimate p, we may use the average number of unppoped kernels in one–quarter cup from the data in Table 1.1 on page 6 and then divide that number by the estimated value of n to obtain the estimate

$$\widehat{p} = \frac{56}{342} \approx 0.1637.$$

Thus, in this example, we assume that the counts, X, of unpopped kernels in one–quarter cup in Table 1.1 on page 6 follows the distribution

$$X \sim \text{binomial}(\widehat{n}, \widehat{p}).$$

Category (i)	Counts Range	p_i	Predicted Counts	Observed Counts
1	$0 \leqslant X \leqslant 50$	0.2131	6	14
2	$51 \leqslant X \leqslant 55$	0.2652	7	3
3	$56 \leqslant X \leqslant 60$	0.2700	7	2
4	$X \geqslant 61$	0.2517	7	8

Table 3.2: Counts Predicted by the Binomial Model

Table 3.2 shows the probabilities predicted by the binomial hypothesis in each of the categories that we used in the previous example in which we tested the Poisson hypothesis. Observe that the binomial model predicts the same expected counts as the Poisson model. We therefore get the same value for the Pearson Chi–Square statistic, $\widehat{Q} = 16.67$. In this case the approximate, asymptotic distribution of Q is $\chi^2(1)$ because we estimated two parameters, n and p, to compute the p_is. Thus, the p–value in this case is approximated by

$$p\text{–value} \approx 4.45 \times 10^{-5},$$

which is a very small probability. Thus, we reject the binomial hypothesis. Hence the hypothesis that distribution of the counts of unpopped kernels follows a binomial model is not supported by the data. Consequently, the interval estimate for p which we obtained in Example 2.2.2 on page 17 is not justified since that interval was obtained under the assumption of a binomial model. We therefore need to come up with another way to obtain an interval estimate for p.

At this point we need to re–evaluate the model and re–examine the assumptions that went into the choice of the Poisson and binomial distributions as possible explanations for the distribution of counts in Table 1.1 on page 6. An important assumption that goes into the derivations of both models is that of independent trials. In this case, a trial consists of determining whether a given kernel will pop or not. It was mentioned in Section 1.1.1 that a kernel in a hot–air popper might not pop because it gets pushed out of the container because of the popping of kernels in the neighborhood of the given kernel. Thus, the event that a kernel will not pop will depend on whether a nearby kernel popped or not, and no necessarily on some intrinsic property of the kernel. These considerations are not consistent with the independence assumption required by bout the Poisson and the binomial models. Thus, these models are not appropriate for this situation.

How we proceed from this point on will depend on which question we want to answer. If we want to know what the intrinsic probability of not popping for a given kernel is, independent of the popping mechanism that is used, we need to redesign the experiment so that the popping procedure that is used will guarantee the independence of trials required by the binomial or Poisson models. For example, a given number of kernels, n, might be laid out on flat surface in a microwave oven.

If we want to know what the probability of not popping is for the hot–air popper, we need to come up with another way to model the distribution. This process is complicated by the fact that there are two mechanisms at work that prevent a given from popping: an intrinsic mechanism depending on the properties of a given kernel, and the swirling about of the kernels in the container that makes it easy for the popping of a given kernel to cause other kernels to be pushed out before they pop. Both of these mechanisms need to be modeled.

Example 3.2.3 (A test of normality)**.** In this example we test whether the counts of kernels in one–quarter cup shown in Table 1.2 on page 8 can be assumed to come from a normal distribution. We first use the data in the table to estimate μ and σ^2. Based on the calculations in Example 2.3.13 on page 42, we get the following estimates

$$\widehat{\mu} = \overline{X}_n \approx 342,$$

and

$$\widehat{\sigma} = \overline{S}_n \approx 35.$$

We therefore assume that

$$N \sim \text{normal}(\widehat{\mu}, \widehat{\sigma}^2)$$

and use the corresponding pdf to compute the probabilities that the counts will lie in certain ranges.

Table 3.3 on page 54 shows those ranges and their corresponding probabilities. Note that the ranges for the counts were chosen so that the expected count for each category is 5. Table 3.3 shows also the predicted and observed counts from which we get the value for the chi–square statistic, Q, to be $\widehat{Q} = 2/5$. In this case Q has an approximate $\chi^2(1)$ asymptotic distribution, according to

Category (i)	Counts Range	p_i	Predicted Counts	Observed Counts
1	$N \leqslant 319$	0.2555	5	6
2	$319 < N \leqslant 342$	0.2445	5	5
3	$342 < N \leqslant 365$	0.2445	5	5
4	$N > 365$	0.2555	5	4

Table 3.3: Kernels in 1/4–cup Predicted by the Normal Model

Pearson's Theorem, since we estimated two parameters, μ and σ, based on the data. We therefore obtain the approximate p–value

$$p\text{–value} = \mathrm{P}(Q \geqslant \widehat{Q}) \approx 0.5271.$$

Thus, based on the data, we cannot reject the null hypothesis that the counts can be described as following a normal distribution. Hence, we were justified in assuming a normal model when estimating the mean number of kernels in one–quarter cup in Example 2.3.13 on page 42.

3.3 Hypothesis Tests in General

Hypothesis testing is a tool in statistical inference which provides a general framework for rejecting certain hypothesis, known as the **null hypothesis** and denoted by H_o, against an **alternative hypothesis**, denoted by H_1. For instance, in Example 3.2.3 on page 53 we tested the hypothesis that the counts of kernel in a one–quarter cup, shown in Table 1.2 on page 8, follows a normal distribution. In this case, denoting the counts of kernels by N, we may state the null and alternative hypotheses as

$$\mathrm{H}_o: \quad N \text{ is normaly distributed}$$

and

$$\mathrm{H}_1: \quad N \text{ is not normaly distributed.}$$

Here is another example.

Example 3.3.1. We wish to determine whether a given coin is fair or not. Thus, we test the null hypothesis

$$\mathrm{H}_o: \quad p = \frac{1}{2}$$

versus the alternative hypothesis

$$\mathrm{H}_1: \quad p \neq \frac{1}{2},$$

where p denotes the probability that a given toss of the coin will yield a head.

In order to tests the hypotheses in Example 3.3.1, we may perform an experiment which consists of flipping the coin 400 times. If we let Y denote the number of heads that we observe, then H_o may be stated as

$$\mathrm{H}_o\colon \quad Y \sim \text{binomial}(400, 0.5).$$

Notice that this hypothesis completely specifies the distribution of the random variable Y, which is known as a **test statistic**. On the other hand, the hypothesis in the goodness of fit test in Example 3.2.3 on page 53 does not specify a distribution. H_o in Example 3.2.3 simply states that the the count, N, of kernels in one–quarter cup follows a normal distribution, but it does not specify the parameters μ and σ^2.

Definition 3.3.2 (Simple versus Composite Hypotheses)**.** A hypothesis which completely specifies a distribution is said to be a **simple hypothesis**. A hypothesis which is not simple is said to be **composite**.

For example, the alternative hypothesis, $\mathrm{H}_1\colon \; p \neq 0.5$, in Example 3.3.1 is composite since the test statistic, Y, for that test is binomial$(400, p)$ where p is any value between 0 and 1 which is not 0.5. Thus, H_1 is a really a combination of many hypotheses.

The decision to reject or not reject H_o in a hypothesis test is based on a set of observations, $X_1, X_2, \ldots, X_n$; these could be the outcomes of certain experiment performed to test the hypothesis and are, therefore, random variables with certain distribution. Given a set of of observations, $X_1, X_2, \ldots, X_n$, a **test statistic**, $T = T(X_1, X_2, \ldots, X_n)$, may be formed. For instance, in Example 3.3.1 on page 54, the experiment might consist of flipping the coin 400 times and determining the number of heads. If the null hypothesis in that test is true, then the 400 observations are independent Bernoulli(0.5) trials. We can define the test statistic for this test to be

$$T = \sum_{i=1}^{400} X_i$$

so that, in H_o is true,

$$T \sim \text{binomial}(400, 0.5).$$

A test statistic for a hypothesis test may be used to establish a criterion for rejecting H_o. For instance in the coin tossing Example 3.3.1, we can say that we reject the hypothesis that the coin is fair if

$$|T - 200| > c; \tag{3.6}$$

that is, the distance from the statistic T to the mean of the assumed distribution is at least certain **critical** value, c. The condition in (3.6) constitutes a **decision criterion** for rejection of H_o. It the null hypothesis is true and the observed value, $\widehat{T}$, of the test statistic, T, falls within the range specified by the rejection criterion in (3.6), we mistakenly reject H_o when it is in fact true. This is known as a **Type I** error. The probability of committing a Type I error in the coin tossing example is

$$\mathrm{P}\left(|T-200|>c \mid \ \mathrm{H}_o \text{ true}\right).$$

Definition 3.3.3 (Significance Level)**.** The largest probability of making a type I error is denoted by α and is called the **significance level** of the test.

Example 3.3.4. In the coin tossing example (Example 3.3.1), we can set a given significance level, α, as follows. Since the number of tosses is large, $n = 400$, we can use the central limit theorem to get that

$$\begin{aligned}\mathrm{P}\left(|T-200|>c \mid \ \mathrm{H}_o \text{ true}\right) &= \mathrm{P}\left(\frac{|T-200|}{\sqrt{400\cdot(0.5)(1-0.5)}} > \frac{c}{10}\right) \\ &\approx \mathrm{P}\left(|Z| > \frac{c}{10}\right),\end{aligned}$$

where $Z \sim \text{normal}(0,1)$. It then follows that, if we set

$$\frac{c}{10} = z_{\alpha/2},$$

where $z_{\alpha/2}$ is such that $\mathrm{P}(|Z| > z_{\alpha/2}) = \alpha$, we obtain that $c = 10z_{\alpha/2}$. Hence, the rejection criterion

$$|T-200| > 10z_{\alpha/2}$$

yields a test with a significance level α. For example, if $\alpha = 0.05$, then we get that $z_{\alpha/2} = 1.96$ and therefore $c = 19.6 \approx 20$. Hence, the test that rejects H_o if

$$T < 180 \quad \text{or} \quad T > 220$$

has a significance level of $\alpha = 0.05$.

If the null hypothesis, H_o, is in fact false, but the hypothesis test does not yield the rejection of H_o, then a type II error is made. The probability of a type II error is denoted by β.

In general, a hypothesis test is concerned with the question of whether a parameter, θ, from certain underlying distribution is in a certain range or not. Suppose the underlying distribution has pdf or pmf denoted by $f(x \mid \theta)$, where we have explicitly expressed the dependence of the distribution function on the parameter θ; for instance, in Example 3.3.1, the underlying distribution is

$$f(x \mid p) = p^x(1-p)^{1-x}, \quad \text{for} \ \ x = 0 \text{ or } x = 1,$$

and 0 otherwise. The parameter θ in this case is p, the probability of a success in a Bernoulli(p) trial.

In the general setting, the null and alternative hypothesis are statements of the form

$$\mathrm{H}_o\colon \quad \theta \in \Omega_o$$

and

$$\mathrm{H}_1: \quad \theta \in \Omega_1$$

where Ω_o and Ω_1 are complementary subsets of a parameter space $\Omega = \Omega_o \cup \Omega_1$, where $\Omega_o \cap \Omega_1 = \emptyset$. In Example 3.3.1, we have that $\Omega_o = \{0.5\}$ and

$$\Omega_1 = \{p \in [0,1] \mid p \neq 0.5\}.$$

Given a set of observations, $X_1, X_2, \ldots, X_n$, which may be assumed to be iid random variables with distribution $f(x \mid \theta)$, we denote the set all possible values of the n–tuple $(X_1, X_2, \ldots, X_n)$ by $\mathcal{D}$. Consider a statistic

$$T = T(X_1, X_2, \ldots, X_n).$$

A **rejection region**, R, for a test is defined by

$$R = \{(x_1, x_2, \ldots, x_n) \in \mathcal{D} \mid T(x_1, x_2, \ldots, x_n) \in A\}, \tag{3.7}$$

where A is a subset of the real line. For example, in the coin tossing example, we had the rejection region

$$R = \{(x_1, x_2, \ldots, x_n) \in \mathcal{D} \mid |T(x_1, x_2, \ldots, x_n) - E(T)| > c\},$$

or

$$R = \{(x_1, x_2, \ldots, x_n) \in \mathcal{D} \mid |T(x_1, x_2, \ldots, x_n) - np| > c\},$$

since, this this case, $T \sim$ binomial(n,p), where $n = 400$, and p depends on which hypothesis we are assuming to be true. Thus, in this case, the set A in the definition of the rejection region in (3.7) is

$$A = (-\infty, np - c) \cup (np + c, \infty).$$

Given a rejection region, R, for a test of hypotheses

$$\mathrm{H}_o: \quad \theta \in \Omega_o$$

and

$$\mathrm{H}_1: \quad \theta \in \Omega_1,$$

let

$$\mathrm{P}_\theta((x_1, x_1, \ldots, x_n) \in R)$$

denote the probability that the observation values fall in the rejection region under the assumption that the random variables $X_1, X_2, \ldots, X_n$ are iid with distribution $f(x \mid \theta)$. Thus,

$$\max_{\theta \in \Omega_o} \mathrm{P}_\theta((x_1, x_1, \ldots, x_n) \in R)$$

is the largest probability that H_o will be rejected given that H_o; this is the significance level for the test; that is,

$$\alpha = \max_{\theta \in \Omega_o} \mathrm{P}_\theta((x_1, x_1, \ldots, x_n) \in R).$$

In Example 3.3.1 on page 54, $\Omega_o = \{0.5\}$; thus,

$$\alpha = \mathrm{P}_{0.5}\left((x_1, x_2, \ldots, x_n) \in R\right),$$

where

$$R = \{(x_1, x_2, \ldots, x_n) \mid T(x_1, x_2, \ldots, x_n) < np{-}c, \text{ or } T(x_1, x_2, \ldots, x_n) > np{+}c\}.$$

By the same token, for $\theta \in \Omega_1$,

$$\mathrm{P}_\theta((x_1, x_1, \ldots, x_n) \in R)$$

gives the probability of rejecting the null hypothesis when H_o is false. It then follows that the probability of a Type II error, for the case in which $\theta \in \Omega_1$, is

$$\beta(\theta) = 1 - \mathrm{P}_\theta((x_1, x_1, \ldots, x_n) \in R);$$

this is the probability of not rejecting the null hypothesis when H_o is in fact false.

Definition 3.3.5 (Power of a Test)**.** For $\theta \in \Omega_1$, the function

$$\mathrm{P}_\theta\left((x_1, x_1, \ldots, x_n) \in R)\right)$$

is called the **power function** for the test at θ; that is, $\mathrm{P}_\theta\left((x_1, x_1, \ldots, x_n) \in R)\right)$ is the probability of rejecting the null hypothesis when it is in fact false. We will use the notation

$$\gamma(\theta) = \mathrm{P}_\theta\left((x_1, x_1, \ldots, x_n) \in R)\right) \quad \text{for} \quad \theta \in \Omega_1.$$

Example 3.3.6. In Example 3.3.1 on page 54, consider the rejection region

$$R = \{(x_1, x_2, \ldots, x_n) \in \mathcal{D} \mid |T(x_1, x_2, \ldots, x_n) - 200| > 20\}$$

where

$$T(x_1, x_2, \ldots, x_n) = \sum_{j=1}^{n} x_j,$$

for $n = 400$, in the test of $\mathrm{H}_o\colon\ p = 0.5$ against $\mathrm{H}_1\colon\ p \neq 0.5$. The significance level for this test is

$$\alpha = \mathrm{P}_{0.5}\left(|T - 200| > 20\right),$$

where

$$T \sim \text{binomial}(400, 0.5).$$

Thus,

$$\begin{aligned} \alpha &= 1 - \mathrm{P}\left(|T - 200| \leqslant 20\right) \\ &= 1 - \mathrm{P}\left(180 \leqslant T \leqslant 220\right) \\ &= 1 - \mathrm{P}\left(179.5 < T \leqslant 220.5\right), \end{aligned}$$

where we have used the continuity correction, since we are going to be applying the Central Limit Theorem to approximate a discrete distribution; namely, T has an approximate normal$(200, 100)$ distribution in this case, since $n = 400$ is large. We then have that

$$\alpha \quad \approx \quad 1 - \mathrm{P}\left(179.5 < Y \leqslant 220.5\right),$$

where $Y \sim$ normal$(200, 100)$. Consequently,

$$\alpha \quad \approx \quad 1 - (F_{_Y}(220.5) - F_{_Y}(179.5)),$$

where $F_{_Y}$ is cdf of $Y \sim$ normal$(200, 100)$. We then have that

$$\alpha \quad \approx \quad 0.0404.$$

We next compute the power function for this test:

$$\gamma(p) = \mathrm{P}(|T - 200| > 20),$$

where

$$T \sim \text{binomial}(400, p) \quad \text{for} \quad p \neq \frac{1}{2}.$$

We write

$$\begin{aligned} \gamma(p) &= \mathrm{P}(|T - 200| > 20) \\ &= 1 - \mathrm{P}(|T - 200| \leqslant 20) \\ &= 1 - \mathrm{P}(180 \leqslant T \leqslant 220) \\ &= 1 - \mathrm{P}(179.5 < T \leqslant 220.5), \end{aligned}$$

where we have used again the continuity correction, since we are going to be applying the Central Limit Theorem to approximate the distribution of T by that of a normal$(400p, 400p(1-p))$ random variable. We then have that

$$\gamma(p) \quad \approx \quad 1 - \mathrm{P}(179.5 < Y_p \leqslant 220.5),$$

where Y_p denotes a normal$(400p, 400p(1-p))$ random variable. Thus,

$$\gamma(p) \quad \approx \quad 1 - (F_{_{Y_p}}(220.5) - F_{_{Y_p}}(179.5)), \tag{3.8}$$

where $F_{_{Y_p}}$ denotes the cdf of $Y_p \sim$ normal$(400p, 400p(1-p))$.

Table 3.4 on page 60 shows a few values of p and their corresponding approximate values of $\gamma(p)$ according to (3.8). A sketch of the graph of γ as a function of p is shown in Figure 3.3.3 on page 61.

The sketch in Figure 3.3.3 was obtained using the `plot` function in R by typing

```
    plot(p,gammap,type='l',ylab="Power at p")
```

p	$\gamma(p)$
0.10	1.0000
0.20	1.0000
0.30	1.0000
0.40	0.9767
0.43	0.7756
0.44	0.6378
0.45	0.4800
0.46	0.3260
0.47	0.1978
0.48	0.1076
0.49	0.0566
0.50	0.0404
0.51	0.0566
0.52	0.1076
0.53	0.1978
0.54	0.3260
0.55	0.4800
0.56	0.6378
0.57	0.7756
0.60	0.9767
0.70	1.0000
0.80	1.0000
0.90	1.0000

Table 3.4: Table of values of p and $\gamma(p)$

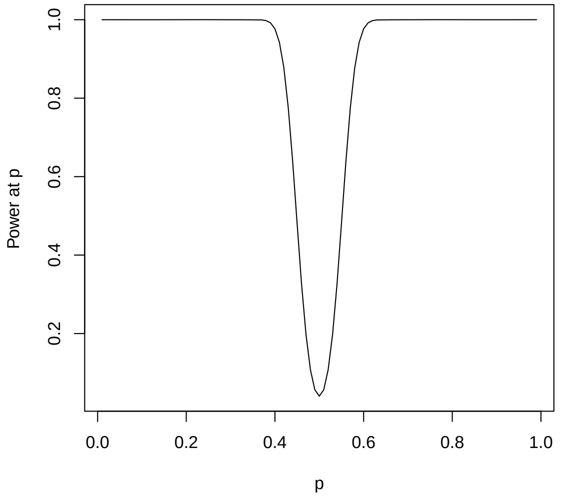

Figure 3.3.3: Sketch of graph of power function for test in Example 3.3.6

where `p` and `gammap` are arrays where values of p and $\gamma(p)$ were stored. These were obtained using the commands:

```
p <- seq(0.01,0.99,by=0.01)
```

and

```
gammap <- 1-(pnorm(220.5,400*p,sqrt(400*p*(1-p)))
                -pnorm(179.5,400*p,sqrt(400*p*(1-p))))
```

Observe that the sketch of the power function in Figure 3.3.3 on page 61 suggests that $\gamma(p)$ tends to 1 as either $p \to 0$ or $p \to 1$, and that $\gamma(p) \to \alpha$ as $p \to 0.5$.

3.4 Likelihood Ratio Test

Likelihood ratio tests provide a general way of obtaining a test statistic, Λ, called a likelihood ratio statistic, and a rejection criterion of the form

$$\Lambda \leqslant c,$$

for some critical value c, for the test of the hypothesis

$$\mathrm{H}_o\colon \quad \theta \in \Omega_o$$

versus the alternative

$$\mathrm{H}_1: \quad \theta \in \Omega_1,$$

based on a random sample, $X_1, X_2, \ldots, X_n$, coming a distribution with distribution function $f(x \mid \theta)$. Here $f(x \mid \theta)$ represents a pdf or a pmf, $\Omega = \Omega_o \cup \Omega_1$ is the parameter space, and $\Omega_o \cap \Omega_1 = \emptyset$.

Before we define the likelihood ratio statistic, Λ, we need to define the concept of a likelihood function.

Definition 3.4.1 (Likelihood Function)**.** Given a random sample, $X_1, X_2, \ldots, X_n$, from a distribution with distribution function $f(x \mid \theta)$, either a pdf or a pmf, where θ is some unknown parameter (either a scalar or a vector parameter), the joint distribution of the sample is given by

$$f(x_1, x_2, \ldots, x_n \mid \theta) = f(x_1 \mid \theta) \cdot f(x_2 \mid \theta) \cdots f(x_n \mid \theta),$$

by the independence condition in the definition of a random sample. If the random variables, $X_1, X_2, \ldots, X_n$, are discrete, $f(x_1, x_2, \ldots, x_n \mid \theta)$ gives the probability of observing the values

$$X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n,$$

under the assumption that the sample is taken from certain distribution with parameter θ. We can also interpret $f(x_1, x_2, \ldots, x_n \mid \theta_o)$ as measuring the likelihood that the parameter θ will take on the value θ_o given that we have observed the values $x_1, x_2, \ldots, x_n$ in the sample. Thus, we call

$$f(x_1, x_2, \ldots, x_n \mid \theta)$$

the **likelihood function** for the parameter θ given the observations

$$X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n,$$

and denote it by $L(\theta \mid x_1, x_2, \ldots, x_n)$; that is,

$$L(\theta \mid x_1, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n \mid \theta).$$

Example 3.4.2 (Likelihood function for independent Bernoulli(p) trials)**.** Let $X_1, X_2, \ldots, X_n$ be a random sample from a Bernoulli(p). Thus, the underlying distribution in this case is

$$f(x \mid p) = p^x (1-p)^{1-x} \quad \text{for } x = 0, 1 \text{ and } 0 \text{ otherwise},$$

where $0 < p < 1$.

We then get that the likelihood function for p, based on the sample observations, is

$$\begin{aligned} L(p \mid x_1, x_2, \ldots, x_n) &= p^{x_1}(1-p)^{1-x_1} \cdot p^{x_2}(1-p)^{1-x_2} \cdots p^{x_n}(1-p)^{1-x_n} \\ &= p^y (1-p)^{n-y} \end{aligned}$$

where $y = \displaystyle\sum_{i=1}^{n} x_i$.

Definition 3.4.3 (Likelihood Ratio Statistic)**.** For a general hypothesis test of

$$\mathrm{H}_o\colon \theta \in \Omega_o$$

against the alternative

$$\mathrm{H}_1\colon \theta \in \Omega_1,$$

based on a random sample, $X_1, X_2, \ldots, X_n$, from a distribution with distribution function $f(x \mid \theta)$, the **likelihood ratio statistic**, $\Lambda(x_1, x_2, \ldots, x_n)$, is defined by

$$\Lambda(x_1, x_2, \ldots, x_n) = \frac{\sup\limits_{\theta\in\Omega_o} L(\theta \mid x_1, x_2, \ldots, x_n)}{\sup\limits_{\theta\in\Omega} L(\theta \mid x_1, x_2, \ldots, x_n)},$$

where $\Omega = \Omega_o \cup \Omega_1$ with $\Omega_o \cap \Omega_1 = \emptyset$.

Example 3.4.4 (Simple hypotheses for Bernoulli(p) trials)**.** Consider the test of

$$\mathrm{H}_o\colon\ p = p_o$$

versus

$$\mathrm{H}_1\colon\ p = p_1,$$

where $p_1 \neq p_o$, based on random sample of size n from a Bernoulli(p) distribution, for $0 < p < 1$. The likelihood ratio statistic for this test is

$$\Lambda(x_1, x_2, \ldots, x_n) = \frac{L(p_o \mid x_1, x_2, \ldots, x_n)}{\max\{L(p_o \mid x_1, x_2, \ldots, x_n), L(p_1 \mid x_1, x_2, \ldots, x_n)\}},$$

since, for this case, $\Omega_o = \{p_o\}$ and $\Omega = \{p_o, p_1\}$; thus,

$$\Lambda(x_1, x_2, \ldots, x_n) = \frac{p_o^y(1-p_o)^{n-y}}{\max\{p_o^y(1-p_o)^{n-y},\ p_1^y(1-p_1)^{n-y}\}},$$

where $y = \sum\limits_{i=1}^{n} x_i$.

Definition 3.4.5 (Likelihood Ratio Test)**.** We can use the likelihood ratio statistic, $\Lambda(x_1, x_2, \ldots, x_n)$, to define the rejection region

$$R = \{(x_1, x_2, \ldots, x_n) \mid \Lambda(x_1, x_2, \ldots, x_n) \leqslant c\},$$

for some critical value c with $0 < c < 1$. This defines a **likelihood ratio test** (LRT) for H_o against H_1.

The rationale for this definition is that, if the likelihood ratio of the sample is very small, the evidence provided by the sample in favor of the null hypothesis is not strong in comparison with the evidence for the alternative. Thus, in this case it makes sense to reject H_o.

Example 3.4.6. Find the likelihood ratio test for

$$\text{H}_o\colon\ p = p_o$$

versus

$$\text{H}_1\colon\ p = p_1,$$

for $p_o \neq p_1$, based on a random sample $X_1, X_2, \ldots, X_n$ from a Bernoulli(p) distribution for $0 < p < 1$.

> ***Solution:*** The rejection region for the likelihood ratio test is given by
>
> $$R\colon \quad \Lambda(x_1, x_2, \ldots, x_n) \leqslant c,$$
>
> for $0 < c < 1$, where
>
> $$\Lambda(x_1, x_2, \ldots, x_n) = \frac{p_o^y(1-p_o)^{n-y}}{\max\{p_o^y(1-p_o)^{n-y},\ p_1^y(1-p_1)^{n-y}\}},$$
>
> with
>
> $$y = \sum_{i=1}^n x_i.$$
>
> Thus, for R to be defined, we must have that
>
> $$\Lambda(x_1, x_2, \ldots, x_n) = \frac{p_o^y(1-p_o)^{n-y}}{p_1^y(1-p_1)^{n-y}};$$
>
> otherwise $\Lambda(x_1, x_2, \ldots, x_n)$ would be 1, and so we wouldn't be able to get the condition $\Lambda(x_1, x_2, \ldots, x_n) \leqslant c$ to hold since $c < 1$. Thus, the LRT for this example will reject H_o if
>
> $$\left(\frac{p_o}{p_1}\right)^y \left(\frac{1-p_o}{1-p_1}\right)^{n-y} \leqslant c,$$
>
> or
>
> $$\left(\frac{1-p_o}{1-p_1}\right)^n \left(\frac{p_o(1-p_1)}{p_1(1-p_o)}\right)^y \leqslant c.$$
>
> Write
>
> $$a = \frac{1-p_o}{1-p_1} \quad \text{and} \quad r = \frac{p_o(1-p_1)}{p_1(1-p_o)}.$$
>
> Then, the LRT rejection region is defined by
>
> $$a^n r^y \leqslant c, \tag{3.9}$$
>
> where
>
> $$y = \sum_{i=1}^n x_i.$$

We consider two cases:

Case 1: $p_1 > p_o$. In this case, $a > 1$ and $r < 1$. Thus, taking the natural logarithm on both sides of (3.9) and solving for y we get that the rejection region for the LRT in this example is equivalent to

$$y \geqslant \frac{\ln c - n \ln a}{\ln r}.$$

In other words, the LRT will reject H_o if

$$Y \geqslant b,$$

where $b = \dfrac{\ln{(c/a^n)}}{\ln r} > 0$, and Y is the statistic

$$Y = \sum_{i=1}^{n} X_i,$$

which counts the number of successes in the sample.

Case 2: $p_1 < p_o$. In this case, $a < 1$ and $r > 1$. We then get from (3.9) the LRT in this example rejects H_o if

$$y \leqslant \frac{\ln c - n \ln a}{\ln r}.$$

In other words, the LRT will reject H_o if

$$Y \leqslant d,$$

where $d = \dfrac{\ln c - n \ln a}{\ln r}$ can be made to be positive by choosing $n > \dfrac{\ln c}{\ln a}$, and Y is again the number of successes in the sample. □

We next consider the example in which we test

$$\mathrm{H}_o\colon \ p = p_o$$

versus

$$\mathrm{H}_1\colon \ p \neq p_0$$

based on a random sample $X_1, X_2, \ldots, X_n$ from a Bernoulli(p) distribution for $0 < p < 1$. We would like to find the LRT rejection region for this test of hypotheses.

In this case the likelihood ratio statistic is

$$\Lambda(x_1, x_2, \ldots, x_n) = \frac{L(p_o \mid x_1, x_2, \ldots, x_n)}{\sup\limits_{1<p<1} L(p \mid x_1, x_2, \ldots, x_n)}, \tag{3.10}$$

where $L(p \mid x_1, x_2, \ldots, x_n) = p^y(1-p)^{n-y}$ for $y = \displaystyle\sum_{i=1}^n x_i$.

In order to determine the denominator in the likelihood ratio in (3.10), we need to maximize the function $L(p \mid x_1, x_2, \ldots, x_n)$ over $0 < p < 1$. We can do this by maximizing the natural logarithm of the likelihood function,

$$\ell(p) = \ln(L(p \mid x_1, x_2, \ldots, x_n)), \quad 0 < p < 1,$$

since $\ln\colon (0,\infty) \to \mathbb{R}$ is a strictly increasing function. Thus, we need to maximize

$$\ell(p) = y \ln p + (n-y)\ln(1-p) \quad \text{over} \quad 0 < p < 1.$$

In order to do this, we compute the derivatives

$$\ell'(p) = \frac{y}{p} - \frac{n-y}{1-p},$$

and

$$\ell''(p) = -\frac{y}{p^2} - \frac{n-y}{(1-p)^2},$$

and observe that $\ell''(p) < 0$ for all $1 < p < 1$. Thus, a critical point of ℓ; that is, a solution of $\ell'(p) = 0$, will yield a maximum for the function $\ell(p)$.

Solving for p in $\ell'(p) = 0$ yields the critical point

$$\widehat{p} = \frac{1}{n} y,$$

which is the sample proportion of successes. This is an example of a **maximum likelihood estimator** (MLE) for p. In general, we have the following definition.

Definition 3.4.7 (Maximum Likelihood Estimator)**.** Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with distribution function $f(x \mid \theta)$, for θ in some parameter space Ω. A value, $\widehat{\theta}$, of the parameter θ such that

$$L(\widehat{\theta} \mid x_1, x_2, \ldots, x_n) = \sup_{\theta \in \Omega} L(\theta \mid x_1, x_2, \ldots, x_n)$$

is called a **maximum likelihood estimator** for θ, or an MLE for θ.

We therefore have that the likelihood ratio statistic for the test of $\mathrm{H}_o\colon\ p = p_o$ versus $\mathrm{H}_1\colon\ p \neq p_o$, based on a random sample of size n from a Bernoulli(p) distribution, is

$$\Lambda(x_1, x_2, \ldots, x_n) = \frac{p_o^y(1-p_o)^{n-y}}{\widehat{p}^y(1-\widehat{p})^{n-y}},$$

where

$$y = \sum_{i=1}^n x_i$$

and

$$\widehat{p} = \frac{1}{n} y$$

is the MLE for p based on the random sample.

Write the likelihood ratio statistic as

$$\begin{aligned}\Lambda(x_1, x_2, \ldots, x_n) &= \left(\frac{p_o}{\widehat{p}}\right)^y \left(\frac{1-p_o}{1-\widehat{p}}\right)^{n-y} \\ &= \left(\frac{p_o}{\widehat{p}}\right)^y \left(\frac{\dfrac{1}{p_o} - 1}{\dfrac{1}{p_o} - \dfrac{\widehat{p}}{p_o}}\right)^{n-y},\end{aligned}$$

and set $t = \dfrac{\widehat{p}}{p_o}$. Then, Λ can be written as a function of t as follows

$$\Lambda(t) = \frac{1}{t^{np_ot}} \cdot \left(\frac{1-p_o}{1-p_ot}\right)^{n-np_ot}, \quad \text{for } 0 \leqslant t \leqslant \frac{1}{p_o},$$

where we have used the fact that $\widehat{p} = \dfrac{1}{n}y$ so that $y = np_ot$.

We now proceed to sketch the graph of Λ as a function of t for $0 \leqslant t \leqslant \dfrac{1}{p_o}$.

First note that $\Lambda(t)$ attains its maximum value of 1 when $t = 1$; namely, when $\widehat{p} = p_o$. That $t = 1$ is the only value of t at which the maximum for $\Lambda(t)$ is attained can be verified by showing that

$$h(t) = \ln(\Lambda(t)) = -np_ot \ln t - (n - np_ot) \ln\left(\frac{1-p_ot}{1-p_o}\right)$$

attains its maximum solely at $t = 1$. Computing the derivative of h with respect to t we find that

$$\begin{aligned}h'(t) &= -np_o \ln t + np_o \ln\left(\frac{1-p_ot}{1-p_o}\right) \\ &= np_o \ln\left(\frac{1-p_ot}{t(1-p_o)}\right).\end{aligned}$$

Thus, $h'(t) = 0$ if and only if

$$\frac{1-p_ot}{t(1-p_o)} = 1,$$

which implies that $t = 1$ is the only critical point of h. The fact that $t = 1$ yields a maximum for h can be seen by observing that the second derivative of h with respect to t,

$$h''(t) = -\frac{np_o}{t} - \frac{np_o^2}{1-p_ot},$$

is negative for $0 < t < \dfrac{1}{p_o}$.

Observe also that $\lim_{t\to 0^+} h(t) = \ln[(1-p_o)^n]$ and $\lim_{t\to(1/p_o)^-} h(t) = \ln[p_o^n]$, so that

$$\Lambda(0) = (1-p_o)^n \quad \text{and} \quad \Lambda(1/p_o) = p_o^n.$$

Putting all the information about the graph for $\Lambda(t)$ together we obtain a sketch as the one shown in Figure 3.4.4,where we have sketched the case $p_o = 1/4$ and $n = 20$ for $0 \leqslant t \leqslant 4$. The sketch in Figure 3.4.4 suggests that, given any positive value of c such that $c < 1$ and $c > \max\{p_o^n, (1-p_o)^n\}$, there exist positive values t_1 and t_2 such that $0 < t_1 < 1 < t_2 < 1/p_o$ and

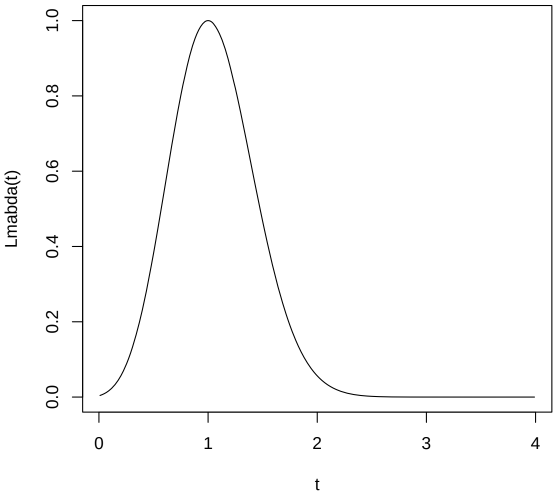

Figure 3.4.4: Sketch of graph of $\Lambda(t)$ for $p_o = 1/4$, $n = 20$, and $0 \leqslant t \leqslant 4$

$$\Lambda(t) = c \quad \text{for} \ \ t = t_1, t_2.$$

Furthermore,

$$\Lambda(t) \leqslant c \quad \text{for} \ \ t \leqslant t_1 \text{ or } t \geqslant t_2.$$

Thus, the LRT rejection region for the test of $\text{H}_o\colon\ p = p_o$ versus $\text{H}_1\colon\ p \neq p_o$ is equivalent to

$$\frac{\widehat{p}}{p_o} \leqslant t_1 \text{ or } \frac{\widehat{p}}{p_o} \geqslant t_2,$$

which we could rephrase in terms of $Y = \sum_{i=1}^{n} X_i$ as

$$R\colon \quad Y \leqslant t_1 n p_o \text{ or } Y \geqslant t_2 n p_o,$$

for some t_1 and t_2 with $0 < t_1 < 1 < t_2$. This rejection region can also be phrased as

$$R\colon \quad Y < np_o - b \text{ or } Y > np_o + b,$$

for some $b > 0$. The value of b will then be determined by the significance level that we want to impose on the test.

Example 3.4.8 (Likelihood ratio test based on a sample from a normal distribution)**.** We wish to test the hypothesis

$$\text{H}_o\colon \quad \mu = \mu_o,\ \sigma^2 > 0$$

versus the alternative

$$\text{H}_1\colon \quad \mu \neq \mu_o,\ \sigma^2 > 0,$$

based on a random sample, $X_1, X_2, \ldots, X_n$, from a normal(μ, σ^2) distribution. Observe that both H_o and H_1 are composite hypotheses.

The likelihood function in this case is

$$\begin{aligned} L(\mu, \sigma \mid x_1, x_2, \ldots, x_n) &= \frac{e^{-(x_1-\mu)^2/2\sigma^2}}{\sqrt{2\pi}\ \sigma} \cdot \frac{e^{-(x_2-\mu)^2/2\sigma^2}}{\sqrt{2\pi}\ \sigma} \cdots \frac{e^{-(x_n-\mu)^2/2\sigma^2}}{\sqrt{2\pi}\ \sigma} \\ &= \frac{e^{-\sum_{i=1}^n (x_i-\mu)^2/2\sigma^2}}{(2\pi)^{n/2}\ \sigma^n}. \end{aligned}$$

The likelihood ratio statistic is

$$\Lambda(x_1, x_2, \ldots, x_n) = \frac{\sup\limits_{\sigma>0} L(\mu_o, \sigma \mid x_1, x_2, \ldots, x_n)}{L(\widehat{\mu}, \widehat{\sigma} \mid x_1, x_2, \ldots, x_n)}, \tag{3.11}$$

where $\widehat{\mu}$ is the MLE for μ and $\widehat{\sigma}^2$ is the MLE for σ^2. To find these MLEs, we need to maximize the natural logarithm of the likelihood function:

$$\ell(\mu, \sigma \mid x_1, x_2, \ldots, x_n) = -\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2 - n \ln \sigma - \frac{n}{2}\ln(2\pi).$$

We therefore need to look at the first partial derivatives

$$\begin{aligned} \frac{\partial \ell}{\partial \mu}(\mu, \sigma) &= \frac{1}{\sigma^2}\sum_{i=1}^n (x_i - \mu) = \frac{n}{\sigma^2}(\overline{x} - \mu) \\ \frac{\partial \ell}{\partial \sigma}(\mu, \sigma) &= \frac{1}{\sigma^3}\sum_{i=1}^n (x_i - \mu)^2 - \frac{n}{\sigma}, \end{aligned}$$

where $\overline{x} = \dfrac{1}{n}\displaystyle\sum_{i=1}^n x_i$, and the second partial derivatives

$$\frac{\partial^2 \ell}{\partial \mu^2}(\mu, \sigma) = -\frac{n}{\sigma^2},$$

$$\frac{\partial^2 \ell}{\partial\sigma\partial\mu}(\mu,\sigma) = \frac{\partial^2 \ell}{\partial\mu\partial\sigma}(\mu,\sigma) = -\frac{2n}{\sigma^3}(\overline{x}-\mu),$$

and

$$\frac{\partial^2 \ell}{\partial\sigma^2}(\mu,\sigma) = -\frac{3}{\sigma^4}\sum_{i=1}^n (x_i-\mu)^2 + \frac{n}{\sigma^2},$$

The critical points of $\ell(\mu,\sigma)$ are solutions to the system

$$\left\{\begin{array}{rcl} \dfrac{\partial \ell}{\partial\mu}(\mu,\sigma) & = & 0 \\[2ex] \dfrac{\partial \ell}{\partial\sigma}(\mu,\sigma) & = & 0, \end{array}\right.$$

which yields

$$\begin{array}{rcl} \widehat{\mu} & = & \overline{x}, \\[2ex] \widehat{\sigma}^2 & = & \dfrac{1}{n}\displaystyle\sum_{i=1}^n (x_i-\overline{x})^2. \end{array}$$

To see that $\ell(\mu,\sigma)$ is maximized at these values, look at the Hessian matrix,

$$\begin{pmatrix} \dfrac{\partial^2 \ell}{\partial\mu^2}(\mu,\sigma) & \dfrac{\partial^2 \ell}{\partial\sigma\partial\mu}(\mu,\sigma) \\[3ex] \dfrac{\partial^2 \ell}{\partial\mu^{\partial}\sigma}(\mu,\sigma) & \dfrac{\partial^2 \ell}{\partial\sigma^2}(\mu,\sigma) \end{pmatrix},$$

at $(\widehat{\mu},\widehat{\sigma})$ to get

$$\begin{pmatrix} -\dfrac{n}{\widehat{\sigma}^2} & 0 \\[3ex] 0 & -\dfrac{2n}{\widehat{\sigma}^2} \end{pmatrix},$$

which has negative eigenvalues. It then follows that $\ell(\mu,\sigma)$ is maximized at $(\widehat{\mu},\widehat{\sigma})$. Hence, $\overline{x}$ is the MLE for μ and

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (x_i-\overline{x})^2$$

is the MLE for σ^2. Observe that $\widehat{\sigma}^2$ is not the sample variance, S_n^2. In fact,

$$\widehat{\sigma}^2 = \frac{n-1}{n}\, S_n^2,$$

so that

$$E(\widehat{\sigma}^2) = \frac{n-1}{n}\sigma^2,$$

and so $\widehat{\sigma}^2$ is not an unbiased estimator of σ^2. It is, however, the maximum likelihood estimator of σ^2.

We then have that the denominator in the likelihood ratio in (3.11) is

$$L(\widehat{\mu},\widehat{\sigma} \mid x_1, x_2, \ldots, x_n) = \frac{e^{-\sum_{i=1}^n (x_i-\overline{x})^2/2\widehat{\sigma}^2}}{(2\pi)^{n/2}\ \widehat{\sigma}^n} = \frac{e^{-n/2}}{(2\pi)^{n/2}\ \widehat{\sigma}^n}.$$

To compute the numerator in (3.11), we need to maximize

$$\begin{aligned} \ell(\sigma) &= \ln(L(\mu_o,\sigma \mid x_1, x_2, \ldots, x_n)) \\ &= -\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu_o)^2 - n\ln\sigma - \frac{n}{2}\ln(2\pi). \end{aligned}$$

Taking derivatives of ℓ we obtain

$$\ell'(\sigma) = \frac{1}{\sigma^3}\sum_{i=1}^n (x_i-\mu_o)^2 - \frac{n}{\sigma}$$

and

$$\ell''(\sigma) = -\frac{3}{\sigma^4}\sum_{i=1}^n (x_i-\mu_o)^2 + \frac{n}{\sigma^2}.$$

Thus, a critical point of $\ell(\sigma)$ is the value, $\overline{\sigma}$, of σ given by

$$\overline{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (x_i-\mu_o)^2.$$

Note that

$$\ell''(\overline{\sigma}) = -\frac{2n}{\overline{\sigma}^2} < 0,$$

so that $\ell(\sigma)$ is maximized when $\sigma = \overline{\sigma}$. We then have that

$$\sup_{\sigma>0} L(\mu_o,\sigma \mid x_1, x_2, \ldots x_n) = L(\mu_o,\overline{\sigma} \mid x_1, x_2, \ldots x_n),$$

where

$$\overline{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (x_i-\mu_o)^2.$$

Observe that

$$\begin{aligned} \sum_{i=1}^n (x_i-\mu_o)^2 &= \sum_{i=1}^n (x_i-\overline{x}+\overline{x}-\mu_o)^2 \\ &= \sum_{i=1}^n (x_i-\overline{x})^2 + \sum_{i=1}^n (\overline{x}-\mu_o)^2, \end{aligned}$$

since

$$\sum_{i=1}^{n} 2(x_i - \overline{x})(\overline{x} - \mu_o) = 2(\overline{x} - \mu_o)\sum_{i=1}^{n}(x_i - \overline{x}) = 0.$$

We then have that

$$\overline{\sigma}^2 = \widehat{\sigma}^2 + (\overline{x} - \mu_o)^2. \tag{3.12}$$

Consequently,

$$\sup_{\sigma>0} L(\mu_o, \sigma \mid x_1, x_2, \ldots x_n) = \frac{e^{-\sum_{i=1}^{n}(x_i-\mu_o)^2/2\overline{\sigma}^2}}{(2\pi)^{n/2}\ \overline{\sigma}^n} = \frac{e^{-n/2}}{(2\pi)^{n/2}\ \overline{\sigma}^n}.$$

Thus, the likelihood ratio statistic in (3.11) is

$$\Lambda(x_1, x_2, \ldots, x_n) = \frac{\sup\limits_{\sigma>0} L(\mu_o, \sigma \mid x_1, x_2, \ldots, x_n)}{L(\widehat{\mu}, \widehat{\sigma} \mid x_1, x_2, \ldots, x_n)} = \frac{\widehat{\sigma}^n}{\overline{\sigma}^n}.$$

Hence, an LRT will reject H$_o$ is

$$\frac{\widehat{\sigma}^n}{\overline{\sigma}^n} \leqslant c,$$

for some c with $0 < c < 1$, or

$$\frac{\widehat{\sigma}^2}{\overline{\sigma}^2} \leqslant c^{2/n},$$

or

$$\frac{\overline{\sigma}^2}{\widehat{\sigma}^2} \geqslant \frac{1}{c^{2/n}},$$

where $\dfrac{1}{c^{2/n}} > 1$. In view of (3.12), we see that and LRT will reject H$_o$ if

$$\frac{(\overline{x} - \mu_o)^2}{\widehat{\sigma}^2} \geqslant \frac{1}{c^{2/n}} - 1 \equiv k,$$

where $k > 0$, and $\widehat{\sigma}^2$ is the MLE for σ^2. Writing $\dfrac{n-1}{n}S_n^2$ for $\widehat{\sigma}^2$ we see that an LRT will reject H$_o$ if

$$\frac{|\overline{x} - \mu_o|}{S_n/\sqrt{n}} \geqslant \sqrt{(n-1)k} \equiv b,$$

where $b > 0$. Hence, the LRT can be based in the test statistic

$$T_n = \frac{\overline{X}_n - \mu_o}{S_n/\sqrt{n}}.$$

Note that T_n has a $t(n-1)$ distribution if H$_o$ is true. We then see that if $t_{\alpha/2,n-1}$ is such that

$$\mathrm{P}(|T| \geqslant t_{\alpha/2,n-1}) = \alpha, \quad \text{for } \ T \sim t(n-1),$$

then the LRT of $\mathrm{H}_o\colon\ \mu = \mu_o$ versus $\mathrm{H}_1\colon\ \mu \neq \mu_o$ which rejects H_o if

$$\frac{|\overline{X}_n - \mu_o|}{S_n/\sqrt{n}} \geqslant t_{\alpha/2,n-1},$$

has significance level α.

Observe also that the set of values of μ_o which do not get rejected by this test is the open interval

$$\left(\overline{X}_n - t_{\alpha/2,n-1}\frac{S_n}{\sqrt{n}}, \overline{X}_n + t_{\alpha/2,n-1}\frac{S_n}{\sqrt{n}}\right),$$

which is a $100(1-\alpha)\%$ confidence interval for the mean, μ, of a normal(μ,σ^2) distribution based on a random sample of size n from that distribution. This provides another interpretation of a confidence interval based on a hypothesis test.

3.5 The Neyman–Pearson Lemma

Consider a test of a simple hypothesis

$$\mathrm{H}_o\colon\ \theta = \theta_o$$

versus the alternative

$$\mathrm{H}_1\colon\ \theta = \theta_1$$

based on a random sample of size n from a distribution with distribution function $f(x \mid \theta)$. The likelihood ratio statistic in this case is

$$\Lambda(x_1, x_2, \ldots, x_n) = \frac{L(\theta_o \mid x_1, x_2, \ldots, x_n)}{L(\theta_1 \mid x_1, x_2, \ldots, x_n)}. \tag{3.13}$$

The rejection region for the LRT is

$$R = \{(x_1, x_2, \ldots, x_n) \mid \Lambda(x_1, x_2, \ldots, x_n) \leqslant c\} \tag{3.14}$$

for some $0 < c < 1$.

If the significance level of the test is α, then

$$\alpha = \int_R f(x_1, x_2, \ldots, x_n \mid \theta_o)\ \mathrm{d}x_1\mathrm{d}x_2\cdots\mathrm{d}x_n,$$

for the case in which $f(x \mid \theta)$ is a pdf. Thus,

$$\alpha = \int_R L(\theta_o \mid x_1, x_2, \ldots, x_n)\ \mathrm{d}x_1\mathrm{d}x_2\cdots\mathrm{d}x_n, \tag{3.15}$$

It then follows that the power of the LRT is

$$\gamma(\theta_1) = \int_R L(\theta_1 \mid x_1, x_2, \ldots, x_n)\ \mathrm{d}x_1\mathrm{d}x_2\cdots\mathrm{d}x_n; \tag{3.16}$$

that is, the probability of reject H_o when H_1 is true.

Consider next another test with rejection region $\widetilde{R}$ and significance level α. We then have that

$$\alpha = \int_{\widetilde{R}} L(\theta_o \mid x_1, x_2, \ldots, x_n) \,\mathrm{d}x_1 \mathrm{d}x_2 \cdots \mathrm{d}x_n, \tag{3.17}$$

and the power of the new test is

$$\widetilde{\gamma}(\theta_1) = \int_{\widetilde{R}} L(\theta_1 \mid x_1, x_2, \ldots, x_n) \,\mathrm{d}x_1 \mathrm{d}x_2 \cdots \mathrm{d}x_n. \tag{3.18}$$

The Neyman–Pearson Lemma states that

$$\widetilde{\gamma}(\theta_1) \leqslant \gamma(\theta_1); \tag{3.19}$$

in other words, out of all the tests of the simple hypothesis $\mathrm{H}_o\colon\ \theta = \theta_o$ versus $\mathrm{H}_1\colon\ \theta = \theta_1$, the LRT yields the largest possible power. Consequently, the LRT gives the smallest probability of making a Type II error our of the tests of significance level α.

The proof of the Neyman–Pearson Lemma is straight forward. First observe that

$$R = (R \cap \widetilde{R}) \cup (R \cap \widetilde{R}^c), \tag{3.20}$$

where $\widetilde{R}^c$ denotes the complement of $\widetilde{R}$. It then follows from (3.15) that

$$\alpha = \int_{R\cap\widetilde{R}} L(\theta_o \mid \mathbf{x}) \,\mathrm{d}\mathbf{x} + \int_{R\cap\widetilde{R}^c} L(\theta_o \mid \mathbf{x}) \,\mathrm{d}\mathbf{x}, \tag{3.21}$$

where we have abbreviated the vector $(x_1, x_2, \ldots, x_n)$ by $\mathbf{x}$, and the volume element $\mathrm{d}x_1\mathrm{d}x_2 \cdots \mathrm{d}x_n$ by $\mathrm{d}\mathbf{x}$. Similarly, using

$$\widetilde{R} = (\widetilde{R} \cap R) \cup (\widetilde{R} \cap R^c), \tag{3.22}$$

and (3.17) we get that

$$\alpha = \int_{\widetilde{R}\cap R} L(\theta_o \mid \mathbf{x}) \,\mathrm{d}\mathbf{x} + \int_{\widetilde{R}\cap R^c} L(\theta_o \mid \mathbf{x}) \,\mathrm{d}\mathbf{x}. \tag{3.23}$$

Combining (3.21) and (3.23) we then get that

$$\int_{R\cap\widetilde{R}^c} L(\theta_o \mid \mathbf{x}) \,\mathrm{d}\mathbf{x} = \int_{\widetilde{R}\cap R^c} L(\theta_o \mid \mathbf{x}) \,\mathrm{d}\mathbf{x}. \tag{3.24}$$

To prove (3.19), use (3.18) and (3.22) to get

$$\widetilde{\gamma}(\theta_1) = \int_{\widetilde{R}\cap R} L(\theta_1 \mid \mathbf{x}) \,\mathrm{d}\mathbf{x} + \int_{\widetilde{R}\cap R^c} L(\theta_1 \mid \mathbf{x}) \,\mathrm{d}\mathbf{x} \tag{3.25}$$

Similarly, using (3.16) and (3.20), we get that

$$\gamma(\theta_1) = \int_{R\cap\widetilde{R}} L(\theta_1 \mid \mathbf{x}) \,\mathrm{d}\mathbf{x} + \int_{R\cap\widetilde{R}^c} L(\theta_1 \mid \mathbf{x}) \,\mathrm{d}\mathbf{x}. \tag{3.26}$$

Next, subtract (3.25) from (3.26) to get

$$\gamma(\theta_1) - \widetilde{\gamma}(\theta_1) = \int_{R\cap\widetilde{R}^c} L(\theta_1 \mid \mathbf{x})\, \mathrm{d}\mathbf{x} - \int_{\widetilde{R}\cap R^c} L(\theta_1 \mid \mathbf{x})\, \mathrm{d}\mathbf{x}, \tag{3.27}$$

and observe that

$$cL(\theta_1 \mid \mathbf{x}) \geqslant L(\theta_o \mid \mathbf{x}) \quad \text{on } R \cap \widetilde{R}^c \tag{3.28}$$

and

$$cL(\theta_1 \mid \mathbf{x}) \leqslant L(\theta_o \mid \mathbf{x}) \quad \text{on } \widetilde{R} \cap R^c, \tag{3.29}$$

where we have used (3.13) and (3.14). Multiplying the inequality in (3.29) by -1 we get that

$$-cL(\theta_1 \mid \mathbf{x}) \geqslant -L(\theta_o \mid \mathbf{x}) \quad \text{on } \widetilde{R} \cap R^c. \tag{3.30}$$

It then follows from (3.27), (3.28) and (3.30)

$$\gamma(\theta_1) - \widetilde{\gamma}(\theta_1) \geqslant \frac{1}{c}\left(\int_{R\cap\widetilde{R}^c} L(\theta_o \mid \mathbf{x})\, \mathrm{d}\mathbf{x} - \int_{\widetilde{R}\cap R^c} L(\theta_o \mid \mathbf{x})\, \mathrm{d}\mathbf{x}\right) = 0 \tag{3.31}$$

where we have used (3.24). The inequality in (3.19) now follows from (3.31). Thus, we have proved the Neymann–Pearson Lemma.

The Neyman–Pearson Lemma applies only to tests of simple hypotheses. For instance, in Example 3.4.6 of page 64 dealing with the test of $\mathrm{H}_o\colon\ p = p_o$ versus $\mathrm{H}_1\colon\ p = p_1$, for $p_1 > p_o$, based on a random sample $X_1, X_2, \ldots, X_n$ from a Bernoulli(p) distribution for $0 < p < 1$, we saw that the LRT rejects the null hypothesis at some significance level, α, is

$$Y = \sum_{i=1}^{n} X_i \geqslant b, \tag{3.32}$$

for some $b > 0$ determined by α. By the Neyman–Pearson Lemma, this is the most powerful test at that significance level; that is, the test with the smallest probability of a Type II error. Recall that the value of b yielding a significance level α may be obtained, for large sample sizes, n, by applying the Central Limit Theorem. In fact, assuming that the null hypothesis is true, the test statistic Y in (3.32) is binomial(n, p_o). We then have that

$$\begin{aligned}
\alpha &= \mathrm{P}(Y \geqslant b) \\
&= \mathrm{P}\left(\frac{Y - np_o}{\sqrt{np_o(1-p_o)}} \geqslant \frac{b - np_o}{\sqrt{np_o(1-p_o)}}\right) \\
&\approx \mathrm{P}\left(Z \geqslant \frac{b - np_o}{\sqrt{np_o(1-p_o)}}\right),
\end{aligned}$$

where $Z \sim \text{normal}(0,1)$. Thus, if z_α is such that $\mathrm{P}(Z \geqslant z_\alpha) = \alpha$, then

$$b = np_o + z_\alpha\ \sqrt{np_o(1-p_o)} \tag{3.33}$$

in (3.32) gives the most powerful test at the significance level of α. Observe that this value of b depends only on p_o and n; it does not depend on p_1.

Now consider the test of $\mathrm{H}_o\colon\ p = p_o$ versus $\mathrm{H}_1\colon\ p > p_o$. Since, the alternative hypothesis is not simple, we cannot apply the Neyman–Pearson Lemma directly. However, by the previous considerations, the test that rejects $\mathrm{H}_o\colon\ p = p_o$ if

$$Y \geqslant b,$$

where b is given by (3.33) for large n is the most powerful test at level α for every $p_1 > p_o$; i.e., for every possible value in the alternative hypothesis $\mathrm{H}_1\colon\ p > p_o$. We then say that the LRT is the **uniformly most powerful test** (UMP) at level α in this case.

Definition 3.5.1 (Uniformly most powerful test)**.** A test of a simple hypothesis $\mathrm{H}_o\colon\ \theta = \theta_o$ against a composite alternative hypothesis $\mathrm{H}_1\colon\ \theta \in \Omega_1$ is said to be **uniformly most powerful test** (UMP) at a level α, if it is most powerful at that level for every simple alternative $\theta = \theta_1$ in Ω_1.

Chapter 4

Evaluating Estimators

Given a random sample, $X_1, X_2, \ldots, X_n$, from a distribution with distribution function $f(x \mid \theta)$, we have seen that there might be more than one statistic,

$$T = T(X_1, X_2, \ldots, X_n),$$

that can be used to estimate the parameter θ. For example, if $X_1, X_2, \ldots, X_n$ is a random sample from a normal(μ, σ^2) distribution, then the sample variance,

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2,$$

and the maximum likelihood estimator,

$$\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2,$$

are both estimators for the variance σ^2. The sample variance, S_n^2, is unbiased, while the MLE is not.

As another example, consider a random sample, $X_1, X_2, \ldots, X_n$, from a Poisson distribution with parameter λ. Then, the sample mean, $\overline{X}_n$ and the then the sample variance, S_n^2. are both unbiased estimators for λ.

Given two estimators for a given parameter, θ, is there a way to evaluate the two estimators in such a way that we can tell which of the two is the better one? In this chapter we explore one way to measure how good an estimator is, the mean squared error or MSE. We will then see how to use that measure to compare one estimator to others.

4.1 Mean Squared Error

Given a random sample, $X_1, X_2, \ldots, X_n$, from a distribution with distribution function $f(x \mid \theta)$, and an estimator, $W = W(X_1, X_2, \ldots, X_n)$, for the parameter

θ, we define the **mean squared error** (MSE) of W to be the expected value of $(W-\theta)^2$. We denote this expected value by $E_\theta\left[(W-\theta)^2\right]$ and compute it, for the case in which $f(x \mid \theta)$ is a pdf, using the formula

$$E_\theta\left[(W-\theta)^2\right] = \int_{\mathbb{R}^n} (W-\theta)^2 f(x_1, x_2, \ldots, x_n \mid \theta)\ \mathrm{d}x_1 \mathrm{d}x_2 \cdots \mathrm{d}x_n,$$

where $f(x_1, x_2, \ldots, x_n \mid \theta)$ is the joint distribution of the sample. The subscript, θ, in the expectation symbol for expectation, E, reminds us that we are using $f(x_1, x_2, \ldots, x_n \mid \theta)$. By the same token, the expectation of the W is written $E_\theta(W)$. We also write

$$\mathrm{MSE}(W) = E_\theta\left[(W-\theta)^2\right].$$

Observe that, since expectation is a linear operation,

$$\begin{aligned}
\mathrm{MSE}(W) &= E_\theta\left[((W - E_\theta(W)) + (E_\theta(W) - \theta))^2\right] \\
&= E_\theta\left[((W - E_\theta(W)))^2 + 2(W - E_\theta(W))(E_\theta(W) - \theta) + (E_\theta(W) - \theta)^2\right] \\
&= E_\theta\left[(W - E_\theta(W))^2\right] + E_\theta\left[(E_\theta(W) - \theta)^2\right]
\end{aligned}$$

since

$$\begin{aligned}
E_\theta\left[2(W - E_\theta(W))(E_\theta(W) - \theta)\right] &= 2(E_\theta(W) - \theta)\ E_\theta\left[(W - E_\theta(W)\right] \\
&= 2(E_\theta(W) - \theta)\ \left[E_\theta(W) - E_\theta(W)\right] \\
&= 0.
\end{aligned}$$

We then have that

$$\mathrm{MSE}(W) = \mathrm{var}_\theta(W) + [E_\theta(W) - \theta]^2;$$

that is, the mean square error of W is the sum of the variance of W and the quantity $[E_\theta(W) - \theta]^2$. The expression $E_\theta(W) - \theta$ is called the **bias** of the estimator W and is denoted by $\mathrm{bias}_\theta(W)$; that is,

$$\mathrm{bias}_\theta(W) = E_\theta(W) - \theta.$$

We then have that the mean square error of an estimator is

$$\mathrm{MSE}(W) = \mathrm{var}_\theta(W) + [\mathrm{bias}_\theta(W)]^2.$$

Thus, if the estimator, W, is unbiased, then $E_\theta(W) = \theta$, so that

$$\mathrm{MSE}(W) = \mathrm{var}_\theta(W), \quad \text{for an unbiased estimator.}$$

Example 4.1.1. Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal(μ, σ^2) distribution. Then, the sample mean, $\overline{X}_n$, and the sample variance, S_n^2, are unbiased estimators of the μ and σ^2, respectively. It then follows that

$$\text{MSE}(\overline{X}_n) = \text{var}(\overline{X}_n) = \frac{\sigma^2}{n}$$

and

$$\text{MSE}(S_n^2) = \text{var}(S_n^2) = \frac{2\sigma^4}{n-1},$$

where we have used the fact that

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

and therefore

$$\frac{(n-1)^2}{\sigma^4}\text{var}(S_n^2) = 2(n-1).$$

Example 4.1.2 (Comparing the sample variance and the MLE in a sample from a norma distribution)**.** Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal(μ, σ^2) distribution. The MLE for σ^2 is the estimator

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

Since $\widehat{\sigma}^2 = \dfrac{n-1}{n}\ S_n^2$, and S_n^2 is an unbiased estimator for θ it follows that

$$E(\widehat{\sigma}^2) = \frac{n-1}{n}\ \sigma^2 = \sigma^2 - \frac{\sigma^2}{n}.$$

It then follows that the bias of $\widehat{\sigma}^2$ is

$$\text{bias}(\widehat{\sigma}^2) = E(\widehat{\sigma}^2) - \sigma^2 = -\frac{\sigma^2}{n},$$

which shows that, on average, $\widehat{\sigma}^2$ underestimates σ^2.

Next, we compute the variance of $\widehat{\sigma}^2$. In order to do this, we used the fact that

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

so that

$$\text{var}\left(\frac{n-1}{\sigma^2} S_n^2\right) = 2(n-1).$$

It then follows from $\widehat{\sigma}^2 = \dfrac{n-1}{n}\ S_n^2$ that

$$\frac{n^2}{\sigma^4}\text{var}(\widehat{\sigma}^2) = 2(n-1),$$

so that

$$\mathrm{var}(\widehat{\sigma}^2) = \frac{2(n-1)\sigma^4}{n^2}.$$

It the follows that the mean squared error of $\widehat{\sigma}^2$ is

$$\begin{aligned}\mathrm{MSE}(\widehat{\sigma}^2) &= \mathrm{var}(\widehat{\sigma}^2) + \mathrm{bias}(\widehat{\sigma}^2)\\ &= \frac{2(n-1)\sigma^4}{n^2} + \frac{\sigma^4}{n^2}\\ &= \frac{2n-1}{n^2}\,\sigma^4.\end{aligned}$$

Comparing the value of $\mathrm{MSE}(\widehat{\sigma}^2)$ to

$$\mathrm{MSE}(S_n^2) = \frac{2\sigma^4}{n-1},$$

we see that

$$\mathrm{MSE}(\widehat{\sigma}^2) < \mathrm{MSE}(S_n^2).$$

Hence, the MLE for σ^2 has a smaller mean squared error than the unbiased estimator S_n^2. Thus, $\widehat{\sigma}^2$ is a more precise estimator than S_n^2; however, S_n^2 is more accurate than $\widehat{\sigma}^2$.

4.2 Crámer–Rao Theorem

If $W = W(X_1, X_2, \ldots, X_n)$ is an unbiased estimator for θ, where $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with distribution function $f(x \mid \theta)$, we saw in the previous section that the mean squared error of W is given by

$$\mathrm{MSE}(W) = \mathrm{var}_\theta(W).$$

The question we would like to answer in this section is the following: Out of all unbiased estimators of θ based on the random sample $X_1, X_2, \ldots, X_n$, is there one with the smallest possible variance, and consequently the smallest possible MSE?

We will provide a partial answer to the question posed above. The answer is based on a lower bound for the variance of a statistic, W, based on a random sample from a distribution with distribution function $f(x \mid \theta)$. The lower bound was discovered independently by Rao and Crámer around the middle of the twentieth century. The idea is to show that

$$\mathrm{var}_\theta(W) \geqslant b(\theta, n)$$

for all estimators, W, based on the sample, for a function b of the parameter θ. The Crámer–Rao inequality can be derived as a consequence of the Cauchy–Schwarz inequality: For any statistics, W_1 and W_2, based on the sample,

$$[\mathrm{cov}(W_1, W_2)]^2 \leqslant \mathrm{var}(W_1) \cdot \mathrm{var}(W_2). \tag{4.1}$$

The proof of (4.1) is very straightforward. Define a function $h\colon \mathbb{R} \to \mathbb{R}$ by

$$h(t) = \mathrm{var}(W_1 + tW_2) \quad \text{for all} \;\; t \in \mathbb{R},$$

and observe that $h(t) \geqslant 0$ for all $t \in \mathbb{R}$. By the properties of the expectation operator and the definition of variance,

$$\begin{aligned} h(t) &= E\left[(W_1 + tW_2)^2\right] - \left[E(W_1 + tW_2)\right]^2 \\ &= E\left[W_1^2 + 2tW_1W_2 + t^2W_2^2\right] - \left[E(W_1) + tE(W_2)\right]^2 \\ &= E\left[W_1^2\right] + 2tE\left[W_1W_2\right] + t^2E\left[W_2^2\right] - \left[E(W_1)\right]^2 - 2tE(W_1)E(W_2) - t^2\left[E(W_2)\right]^2 \\ &= \mathrm{var}(W_1) + 2\ \mathrm{cov}(W_1, W_2)\ t + \mathrm{var}(W_2)\ t^2, \end{aligned}$$

where we have used the definition of covariance

$$\mathrm{cov}(W_1, W_2) = E(W_1W_2) - E(W_1)E(W_2). \tag{4.2}$$

It then follows that $h(t)$ is quadratic polynomial which is never negative. Consequently, the discriminant,

$$[2\ \mathrm{cov}(W_1, W_2)]^2 - 4\ \mathrm{var}(W_2)\mathrm{var}(W_1),$$

is at most 0; that is,

$$4\ \left[\mathrm{cov}(W_1, W_2)\right]^2 - 4\ \mathrm{var}(W_2)\mathrm{var}(W_1) \leqslant 0,$$

from which the Cauchy–Schwarz inequality in (4.1) follows.

To obtain the Crámmer–Rao lower bound, we will apply the Cauchy–Schwarz inequality (4.1) to the case $W_1 = W$ and

$$W_2 = \frac{\partial}{\partial\theta}\left[\ln\left(L(\theta \mid X_1, X_2, \ldots, X_n)\right)\right].$$

In other words, W_1 is the estimator in question and W_2 is the partial derivative with respect to the parameter θ of the natural logarithm of the likelihood function.

In order to compute $\mathrm{cov}(W_1, W_2)$, we will need the expected value of W_2. Note that

$$W_2 = \frac{1}{L(\theta \mid X_1, X_2, \ldots X_n)}\frac{\partial}{\partial\theta}\left[L(\theta \mid X_1, X_2, \ldots, X_n)\right],$$

so that

$$E_\theta(W_2) = \int_{\mathbb{R}^n} \frac{1}{L(\theta \mid \mathbf{x})}\frac{\partial}{\partial\theta}\left[L(\theta \mid \mathbf{x})\right]\ L(\theta \mid \mathbf{x})\ \mathrm{d}\mathbf{x},$$

where we have denoted the vector $(x_1, x_2, \ldots, x_n)$ by $\mathbf{x}$ and the volume element, $\mathrm{d}x_1\mathrm{d}x_2\cdots\mathrm{d}x_n$, by $\mathrm{d}\mathbf{x}$. We then have that

$$E_\theta(W_2) = \int_{\mathbb{R}^n} \frac{\partial}{\partial\theta}\left[L(\theta \mid \mathbf{x})\right]\ \mathrm{d}\mathbf{x}.$$

Assuming that the order of differentiation and integration can be changed, we then have that

$$E_\theta(W_2) = \frac{\partial}{\partial\theta}\left[\int_{\mathbb{R}^n} L(\theta \mid (x))\ \mathrm{d}\mathbf{x}\right] = \frac{\partial}{\partial\theta}(1) = 0,$$

so that W_2 has expected value 0. It then follows from (4.2) that

$$\begin{aligned} \mathrm{cov}(W_1, W_2) &= E_\theta(W_1 W_2) \\ &= \int_{\mathbb{R}^n} W(\mathbf{x})\ \frac{1}{L(\theta \mid \mathbf{x})}\frac{\partial}{\partial\theta}\left[L(\theta \mid \mathbf{x})\right]\ L(\theta \mid \mathbf{x})\ \mathrm{d}\mathbf{x} \\ &= \int_{\mathbb{R}^n} W(\mathbf{x})\ \frac{\partial}{\partial\theta}\left[L(\theta \mid \mathbf{x})\right]\ \mathrm{d}\mathbf{x} \\ &= \int_{\mathbb{R}^n} \frac{\partial}{\partial\theta}\left[W(\mathbf{x})\ L(\theta \mid \mathbf{x})\right]\ \mathrm{d}\mathbf{x}. \end{aligned}$$

Thus, if the order of differentiation and integration can be interchanged, we have that

$$\begin{aligned} \mathrm{cov}(W_1, W_2) &= \frac{\partial}{\partial\theta}\left[\int_{\mathbb{R}^n} W(\mathbf{x})\ L(\theta \mid \mathbf{x})\ \mathrm{d}\mathbf{x}\right] \\ &= \frac{\partial}{\partial\theta}\left[E_\theta(W)\right]. \end{aligned}$$

Thus, if we set

$$g(\theta) = E_\theta(W)$$

for all θ in the parameter range, we see that

$$\mathrm{cov}_\theta(W_1, W_2) = g'(\theta).$$

In particular, if W is an unbiased estimator, $\mathrm{cov}_\theta(W, W_2) = 1$, where

$$W_2 = \frac{\partial}{\partial\theta}\left[\ln\left(L(\theta \mid X_1, X_2, \ldots, X_n)\right)\right].$$

Applying the Cauchy–Schwarz inequality in (4.1) we then have that

$$[g'(\theta)]^2 \leqslant \mathrm{var}(W)\cdot\mathrm{var}(W_2). \tag{4.3}$$

In order to compute $\mathrm{var}(W_2)$, observe that

$$\begin{aligned} W_2 &= \frac{\partial}{\partial\theta}\left(\sum_{i=1}^{n}\ln(f(X_i \mid \theta)\right) \\ &= \sum_{i=1}^{n}\frac{\partial}{\partial\theta}\left(\ln(f(X_i \mid \theta)\right). \end{aligned}$$

Thus, since $X_1, X_2, \ldots, X_n$ are iid random variable with distribution function $f(x \mid \theta)$,

$$\begin{aligned} \operatorname{var}(W_2) &= \sum_{i=1}^{n} \operatorname{var}\left(\frac{\partial}{\partial\theta}\left(\ln(f(X_i \mid \theta)\right)\right) \\ &= n \cdot \operatorname{var}\left(\frac{\partial}{\partial\theta}\left(\ln(f(X_i \mid \theta)\right)\right) \end{aligned}$$

The variance of the random variable $\dfrac{\partial}{\partial\theta}\left[\ln\left(f(X \mid \theta)\right)\right]$ is called the **Fisher Information** and is denoted by $I(\theta)$. We then have that

$$\operatorname{var}(W_2) = nI(\theta).$$

We then obtain from (4.3) that

$$[g'(\theta)]^2 \leqslant nI(\theta)\ \operatorname{var}(W),$$

which yields the Crámer-Rao lower bound

$$\operatorname{var}(W) \geqslant \frac{[g'(\theta)]^2}{nI(\theta)}, \tag{4.4}$$

where

$$I(\theta) = \operatorname{var}\left(\frac{\partial}{\partial\theta}\left[\ln\left(f(X \mid \theta)\right)\right]\right)$$

is the Fisher information. For the case in which W is unbiased we obtain from (4.4) that

$$\operatorname{var}(W) \geqslant \frac{1}{nI(\theta)}. \tag{4.5}$$

Example 4.2.1. Let $X_1, X_2, \ldots, X_n$ be a random sample from a Poisson(λ) distribution. Then,

$$f(X, \lambda) = \frac{\lambda^X}{X!}\, e^{-\lambda},$$

so that

$$\ln(f(X, \lambda)) = X \ln \lambda - \lambda - \ln(X!)$$

and

$$\frac{\partial}{\partial\lambda}[\ln(f(X, \lambda))] = \frac{X}{\lambda} - 1.$$

Then the Fisher information is

$$I(\lambda) = \frac{1}{\lambda^2}\ \operatorname{var}(X) = \frac{1}{\lambda^2} \cdot \lambda = \frac{1}{\lambda}.$$

Thus, the Crámer–Rao lower bound for unbiased estimators is obtained from (4.5) to be

$$\frac{1}{nI(\lambda)} = \frac{\lambda}{n}.$$

Observe that if $W = \overline{X}_n$, the sample mean, then W is unbiased and

$$\text{var}(W) = \frac{\lambda}{n}.$$

Thus, in this case, the lower bound for the variance of unbiased estimators is attained at the sample mean. We say that $\overline{X}_n$ is an **efficient** estimator.

Definition 4.2.2 (Efficient Estimator)**.** An unbiased estimator, W, of a parameter, θ, is said to be **efficient** if its variance is the lower bound in the Crámer–Rao inequality; that is, if

$$\text{var}(W) = \frac{1}{nI(\theta)},$$

where $I(\theta)$ is the Fisher information,

$$I(\theta) = \text{var}\left(\frac{\partial}{\partial\theta}(\ln(f(X \mid \theta))\right).$$

For any unbiased estimator, W, of θ, we define the efficiency of W, denoted $\text{eff}_\theta(W)$, to be

$$\text{eff}_\theta(W) = \frac{1/(nI(\theta))}{\text{var}_\theta(W)}.$$

Thus, by the Crámer–Rao inequality (4.5),

$$\text{eff}_\theta(W) \leqslant 1$$

for all unbiased estimators, W, of θ. Furthermore, $\text{eff}_\theta(W) = 1$ if and only if W is efficient.

Next, we turn to the question of computing the Fisher information, $I(\theta)$. First, observe that

$$I(\theta) = \text{var}\left(\frac{\partial}{\partial\theta}(\ln(f(X \mid \theta))\right) = E_\theta\left[\left(\frac{\partial}{\partial\theta}(\ln(f(X \mid \theta))\right)^2\right], \tag{4.6}$$

since

$$E_\theta\left[\frac{\partial}{\partial\theta}(\ln(f(X \mid \theta))\right] = 0. \tag{4.7}$$

To see why (4.7) is true, observe that

$$\begin{aligned} E_\theta\left[\frac{\partial}{\partial\theta}(\ln(f(X \mid \theta))\right] &= \int_{-\infty}^{\infty} \frac{\partial}{\partial\theta}(\ln(f(x \mid \theta)) \; f(x \mid \theta) \, \mathrm{d}x \\ &= \int_{-\infty}^{\infty} \frac{1}{f(x \mid \theta)} \frac{\partial}{\partial\theta}(f(x \mid \theta)) \; f(x \mid \theta) \, \mathrm{d}x \\ &= \int_{-\infty}^{\infty} \frac{\partial}{\partial\theta}(f(x \mid \theta)) \, \mathrm{d}x \\ &= \frac{\partial}{\partial\theta}\left(\int_{-\infty}^{\infty} f(x \mid \theta) \, \mathrm{d}x\right), \end{aligned}$$

assuming that the order of differentiation and integration can be interchanged. The identity in (4.7) now follows from the fact that

$$\int_{-\infty}^{\infty} f(x \mid \theta)\ \mathrm{d}x = 1.$$

Next, differentiate (4.7) with respect to θ one more time to obtain that

$$\frac{\partial}{\partial\theta} E_{\theta}\left[\frac{\partial}{\partial\theta}\left(\ln(f(X \mid \theta)\right)\right] = 0, \qquad (4.8)$$

where, assuming that the order of differentiation and integration can be interchanged,

$$\begin{aligned}
\frac{\partial}{\partial\theta} E_{\theta}\left[\frac{\partial}{\partial\theta}\left(\ln(f(x \mid \theta)\right)\right] &= \frac{\partial}{\partial\theta}\left[\int_{-\infty}^{\infty}\frac{\partial}{\partial\theta}\left(\ln(f(x \mid \theta)\right)\ f(x \mid \theta)\ \mathrm{d}x\right] \\
&= \int_{-\infty}^{\infty}\frac{\partial^2}{\partial\theta^2}\left(\ln(f(x \mid \theta)\right)\ f(x \mid \theta)\ \mathrm{d}x \\
&\qquad + \int_{-\infty}^{\infty}\frac{\partial}{\partial\theta}\left(\ln(f(x \mid \theta)\right)\ \frac{\partial}{\partial\theta} f(x \mid \theta)\ \mathrm{d}x \\
&= E_{\theta}\left[\frac{\partial^2}{\partial\theta^2}\left(\ln(f(x \mid \theta))\right)\right] \\
&\qquad + \int_{-\infty}^{\infty}\frac{1}{f(x \mid \theta)}\left[\frac{\partial}{\partial\theta} f(x \mid \theta)\right]^2\ \mathrm{d}x,
\end{aligned}$$

where

$$\begin{aligned}
\int_{-\infty}^{\infty}\frac{1}{f(x \mid \theta)}\left[\frac{\partial}{\partial\theta} f(x \mid \theta)\right]^2\ \mathrm{d}x &= \int_{-\infty}^{\infty}\left[\frac{1}{f(x \mid \theta)}\frac{\partial}{\partial\theta} f(x \mid \theta)\right]^2\ f(x \mid \theta)\ \mathrm{d}x \\
&= \int_{-\infty}^{\infty}\left[\frac{\partial}{\partial\theta}\ln(f(x \mid \theta))\right]^2\ f(x \mid \theta)\ \mathrm{d}x \\
&= E_{\theta}\left[\left(\frac{\partial}{\partial\theta}\ln(f(x \mid \theta))\right)^2\right].
\end{aligned}$$

It then follows from (4.6) that

$$\int_{-\infty}^{\infty}\frac{1}{f(x \mid \theta)}\left[\frac{\partial}{\partial\theta} f(x \mid \theta)\right]^2\ \mathrm{d}x = I(\theta).$$

Consequently,

$$\frac{\partial}{\partial\theta} E_{\theta}\left[\frac{\partial}{\partial\theta}\left(\ln(f(x \mid \theta)\right)\right] = E_{\theta}\left[\frac{\partial^2}{\partial\theta^2}\left(\ln(f(x \mid \theta))\right)\right] + I(\theta)$$

In view of (4.7) we therefore have that

$$I(\theta) = -E_{\theta}\left[\frac{\partial^2}{\partial\theta^2}\left(\ln(f(x \mid \theta))\right)\right], \tag{4.9}$$

which gives another formula for computing the Fisher information.

Appendix A

Pearson Chi–Square Statistic

The goal of this appendix is to prove the first part of Theorem 3.1.4 on page 49; namely: assume that

$$(X_1, X_2, \ldots, X_k)$$

is a random vector with a multinomial$(n, p_1, p_2, \ldots, p_k)$ distribution, and define

$$Q = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}. \tag{A.1}$$

Then, for large values of n, Q has an approximate $\chi^2(k-1)$ distribution. The idea for the proof presented here comes from Exercise 3 on page 60 in [Fer02].

We saw is Section 3.1.2 that the result in Theorem 3.1.4 is true for $k = 2$. We begin the discussion in this appendix with the case $k = 3$. It is hoped that the main features of the proof of the general case will be seen in this simple case.

Consider the random vector $\mathbf{U} = (U_1, U_2, U_3)$ with a multinomial$(1, p_1, p_2, p_3)$ distribution. In other words, each component function, U_i, is a Bernoulli(p_i) random variable for $i = 1, 2, 3$, and the distribution of $\mathbf{U}$ is conditioned on

$$U_1 + U_2 + U_3 = 1.$$

We then have that

$$E(U_j) = p_j \quad \text{for} \quad j = 1, 2, 3;$$

$$\text{var}(U_j) = p_j(1 - p_j) \quad \text{for} \quad j = 1, 2, 3;$$

and

$$\text{cov}(U_i, U_j) = -p_i p_j \quad \text{for} \quad i \neq j.$$

Suppose now that we have a sequence of independent multinomial$(1, p_1, p_2, p_3)$ random vectors

$$\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_n, \ldots$$

We then have that the random vector

$$\mathbf{X}_n = (X_1, X_2, X_3) = \sum_{i=1}^{n} \mathbf{U}_i$$

has a multinomial(n, p_1, p_2, p_3) distribution.

We now try to get an expression for the Pearson chi–square statistic in (A.1), for $k = 3$, in terms of the bivariate random vector

$$\mathbf{W}_n = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}.$$

The expected value of the random vector $\mathbf{W}_n$ is

$$E(\mathbf{W}_n) = \begin{pmatrix} np_1 \\ np_2 \end{pmatrix},$$

and its covariance matrix is

$$\mathbf{C}_{W_n} = n \begin{pmatrix} p_1(1-p_1) & -p_1p_2 \\ -p_1p_2 & p_2(1-p_2) \end{pmatrix},$$

or

$$\mathbf{C}_W = n\mathbf{C}_{(U_1,U_2)},$$

where $\mathbf{C}_{(U_1,U_2)}$ is the covariance matrix for the bivariate random vector $\begin{pmatrix} U_1 \\ U_2 \end{pmatrix}$, for $(U_1, U_2, U_3) \sim$ multinomial$(1, p_1, p_2, p_3)$. Note that the determinant of the matrix $\mathbf{C}_{(U_1,U_2)}$ is $p_1p_2p_3$, which is different from 0 since we are assuming that

$$0 < p_i < 1 \quad \text{for } i = 1, 2, 3.$$

It then follows that $\mathbf{C}_{(U_1,U_2)}$ is invertible with inverse

$$\mathbf{C}^{-1}_{(U_1,U_2)} = \begin{pmatrix} \dfrac{1}{p_1} + \dfrac{1}{p_3} & \dfrac{1}{p_3} \\ \\ \dfrac{1}{p_3} & \dfrac{1}{p_2} + \dfrac{1}{p_3} \end{pmatrix}.$$

Consequently,

$$n\ \mathbf{C}^{-1}_{W_n} = \begin{pmatrix} \dfrac{1}{p_1} + \dfrac{1}{p_3} & \dfrac{1}{p_3} \\ \\ \dfrac{1}{p_3} & \dfrac{1}{p_2} + \dfrac{1}{p_3} \end{pmatrix}.$$

Note also that

$$(\mathbf{W}_n - E(\mathbf{W}_n))^T \mathbf{C}^{-1}_{W_n} (\mathbf{W}_n - E(\mathbf{W}_n)),$$

where $(\mathbf{W}_n - E(\mathbf{W}_n))^T$ is the transpose of the column vector $(\mathbf{W}_n - E(\mathbf{W}_n))$, is equal to

$$n^{-1}(X_1 - np_1, X_2 - np_2) \begin{pmatrix} \dfrac{1}{p_1} + \dfrac{1}{p_3} & \dfrac{1}{p_3} \\ \dfrac{1}{p_3} & \dfrac{1}{p_2} + \dfrac{1}{p_3} \end{pmatrix} \cdot \begin{pmatrix} X_1 - np_1 \\ X_2 - np_2 \end{pmatrix},$$

which is equal to

$$n^{-1}\left(\frac{1}{p_1} + \frac{1}{p_3}\right)(X_1 - np_1)^2$$
$$+\frac{n^{-1}}{p_3}(X_1 - np_1)(X_2 - np_2) + \frac{n^{-1}}{p_3}(X_2 - np_2)(X_1 - np_1)$$
$$n^{-1}\left(\frac{1}{p_2} + \frac{1}{p_3}\right)(X_2 - np_2)^2.$$

Note that

$$\begin{aligned} (X_1 - np_1)(X_2 - np_2) &= (X_1 - np_1)(n - X_1 - X_3 - np_2) \\ &= (X_1 - np_1)(n(1 - p_2) - X_1 - X_3) \\ &= (X_1 - np_1)(n(p_1 + p_3) - X_1 - X_3) \\ &= -(X_1 - np_1)(X_1 - np_1 + X_3 - np_3) \\ &= -(X_1 - np_1)^2 - (X_1 - np_1)(X_3 - np_3). \end{aligned}$$

Similarly, we obtain that

$$(X_2 - np_2)(X_1 - np_1) = -(X_2 - np_2)^2 - (X_2 - np_2)(X_3 - np_3).$$

We then have that

$$(\mathbf{W}_n - E(\mathbf{W}_n))^T \mathbf{C}_{W_n}^{-1} (\mathbf{W}_n - E(\mathbf{W}_n))$$

is equal to

$$n^{-1}\left(\frac{1}{p_1} + \frac{1}{p_3} - \frac{1}{p_3}\right)(X_1 - np_1)^2$$
$$-\frac{n^{-1}}{p_3}(X_1 - np_1)(X_3 - np_3) - \frac{n^{-1}}{p_3}(X_2 - np_2)(X_3 - np_3)$$
$$n^{-1}\left(\frac{1}{p_2} + \frac{1}{p_3} - \frac{1}{p_3}\right)(X_2 - np_2)^2,$$

or

$$\frac{1}{np_1}(X_1 - np_1)^2$$

$$-\frac{1}{np_3}(X_3 - np_3)[(X_1 - np_1) + (X_2 - np_2)]$$

$$\frac{1}{np_2}(X_2 - np_2)^2,$$

where

$$\begin{aligned}(X_3 - np_3)[(X_1 - np_1) + (X_2 - np_2)] &= (X_3 - np_3)[X_1 + X_2 - n(p_1 + p_2)] \\ &= (X_3 - np_3)[n - X_3 - n(1 - p_3)] \\ &= -(X_3 - np_3)^2.\end{aligned}$$

We have therefore shown that

$$(\mathbf{W}_n - E(\mathbf{W}_n))^T \mathbf{C}_{W_n}^{-1} (\mathbf{W}_n - E(\mathbf{W}_n))$$

is equal to

$$\frac{1}{np_1}(X_1 - np_1)^2 + \frac{1}{np_3}(X_3 - np_3)^2 + \frac{1}{np_2}(X_2 - np_2)^2;$$

that is,

$$(\mathbf{W}_n - E(\mathbf{W}_n))^T \mathbf{C}_{W_n}^{-1} (\mathbf{W}_n - E(\mathbf{W}_n)) = \sum_{j=1}^{3} \frac{(X_j - np_j)^2}{np_j},$$

which is the Pearson Chi–Square statistic for the case $k = 3$. Consequently,

$$Q = (\mathbf{W}_n - E(\mathbf{W}_n))^T \mathbf{C}_{W_n}^{-1} (\mathbf{W}_n - E(\mathbf{W}_n)).$$

Our goal now is to show that, as $n \to \infty$,

$$(\mathbf{W}_n - E(\mathbf{W}_n))^T \mathbf{C}_{W_n}^{-1} (\mathbf{W}_n - E(\mathbf{W}_n))$$

tends in distribution to a $\chi^2(2)$ random variable.

Observe that the matrix $\mathbf{C}_{W_n}^{-1}$ is symmetric and positive definite. Therefore, it has a square root, $\mathbf{C}_{W_n}^{-1/2}$, which is also symmetric and positive definite. Consequently,

$$Q = (\mathbf{W}_n - E(\mathbf{W}_n))^T \left(\mathbf{C}_{W_n}^{-1/2}\right)^T \mathbf{C}_{W_n}^{-1/2} (\mathbf{W}_n - E(\mathbf{W}_n)),$$

which we can write as

$$Q = (\mathbf{C}_{W_n}^{-1/2}(\mathbf{W}_n - E(\mathbf{W}_n)))^T \mathbf{C}_{W_n}^{-1/2} (\mathbf{W}_n - E(\mathbf{W}_n)).$$

Next, put

$$\mathbf{Z}_n = \mathbf{C}_{W_n}^{-1/2}(\mathbf{W}_n - E(\mathbf{W}_n))$$

and apply the multivariate central limit theorem (see, for instance, [Fer02, Theorem 5, p. 26]) to obtain that

$$\mathbf{Z}_n \xrightarrow{D} \mathbf{Z} \sim \text{normal}(0, I) \quad \text{as } n \to \infty;$$

that is, the bivariate random vectors, $\mathbf{Z}_n = \mathbf{C}_{W_n}^{-1/2}(\mathbf{W}_n - E(\mathbf{W}_n))$, converge in distribution to the bivariate random vector $\mathbf{Z}$ with mean $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and covariance matrix

$$I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

In other words

$$\mathbf{Z} = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix},$$

where Z_1 and Z_2 are independent normal$(0,1)$ random variables. Consequently,

$$(\mathbf{C}_{W_n}^{-1/2}(\mathbf{W}_n - E(\mathbf{W}_n))^T \mathbf{C}_{W_n}^{-1/2}(\mathbf{W}_n - E(\mathbf{W}_n)) \xrightarrow{D} \mathbf{Z}^T\mathbf{Z} = Z_1^2 + Z_2^2$$

as $n \to \infty$; that is,

$$Q = \sum_{i=1}^{3} \frac{(X_i - np_i)^2}{np_i}$$

converges in distribution to a $\chi^2(2)$ random variable as $n \to \infty$, which we wanted to prove.

The proof of the general case is analogous. Begin with the multivariate random vector

$$(X_1, X_2, \ldots, X_k) \sim \text{multinomial}(n, p_1, p_2, \ldots, p_k).$$

Define the random vector

$$\mathbf{W}_n = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_{k-1} \end{pmatrix}$$

with covariance matrix $\mathbf{C}_{W_n}$. Verify that

$$Q = \sum_{j=1}^{k} \frac{(X_j - np_j)^2}{np_j} = (\mathbf{C}_{W_n}^{-1/2}(\mathbf{W}_n - E(\mathbf{W}_n)))^T \mathbf{C}_{W_n}^{-1/2}(\mathbf{W}_n - E(\mathbf{W}_n)).$$

Next, put

$$\mathbf{Z}_n = \mathbf{C}_{W_n}^{-1/2}(\mathbf{W}_n - E(\mathbf{W}_n))$$

and apply the multivariate central limit theorem to obtain that

$$\mathbf{Z}_n \xrightarrow{D} \mathbf{Z} \sim \text{normal}(0, I) \quad \text{as } n \to \infty,$$

where I is the $(k-1) \times (k-1)$ identity matrix, so that

$$\mathbf{Z} = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_{k-1} \end{pmatrix},$$

where $Z_1, Z_2, \ldots, Z_{k-1}$ are independent normal$(0,1)$ random variables. Consequently,

$$(\mathbf{C}_{W_n}^{-1/2}(\mathbf{W}_n - E(\mathbf{W}_n))^T \mathbf{C}_{W_n}^{-1/2}(\mathbf{W}_n - E(\mathbf{W}_n)) \xrightarrow{D} \mathbf{Z}^T\mathbf{Z} = \sum_{j=1}^{k-1} Z_j^2 \sim \chi^2(k-1)$$

as $n \to \infty$. This proves that

$$Q = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}$$

converges in distribution to a $\chi^2(k-1)$ random variable as $n \to \infty$.

Appendix B

The Variance of the Sample Variance

The main goal of this appendix is to compute the variance of the sample variance based on a sample from and arbitrary distribution; i.e.,

$$\text{var}(S_n^2),$$

where

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

We will come up with a formula based on the second and fourth central moments of the underlying distribution. More precisely, we will prove that

$$\text{var}(S_n^2) = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\mu_2^2\right), \tag{B.1}$$

where μ_2 denotes the second central moment, or variance, of the distribution and μ_4 is the fourth central moment.

In general, we define the first central moment, μ_1, of the distribution of X to be

$$\mu_1 = E(X),$$

the mean on the distribution. The second central moment of X, μ_2, is

$$\mu_2 = E\left[(X - E(X))^2\right];$$

in other words, μ_2 is the variance of the distribution. Similarly, for any $k \geqslant 2$, the kth central moment, μ_k, of X is

$$\mu_k = E\left[(X - E(X))^k\right].$$

First observe that, for each i and j

$$\begin{aligned}(X_i - X_j)^2 &= (X_i - \overline{X}_n + \overline{X}_n - X_j)^2 \\ &= (X_i - \overline{X}_n)^2 - 2(X_i - \overline{X}_n)(\overline{X}_n - X_j) + (\overline{X}_n - X_j)^2,\end{aligned}$$

so that

$$\sum_i \sum_j (X_i - X_j)^2 = \sum_i \sum_j (X_i - \overline{X}_n)^2 + \sum_i \sum_j (X_j - \overline{X}_n)^2, \tag{B.2}$$

since

$$\sum_i \sum_j (X_i - \overline{X}_n)(\overline{X}_n - X_j) = \sum_i (X_i - \overline{X}_n) \sum_j (\overline{X}_n - X_j) = 0.$$

It then follows from (B.2) that

$$\begin{aligned}\sum_i \sum_j (X_i - X_j)^2 &= n \sum_i (X_i - \overline{X}_n)^2 + n \sum_j (X_j - \overline{X}_n)^2 \\ &= n(n-1)S_n^2 + n(n-1)S_n^2\end{aligned}$$

from which we obtain another formula for the sample variance:

$$S_n^2 = \frac{1}{2n(n-1)} \sum_i \sum_j (X_i - X_j)^2,$$

which we can also write as

$$S_n^2 = \frac{1}{n(n-1)} \sum\sum_{i<j} (X_i - X_j)^2. \tag{B.3}$$

In order to compute $\text{var}(S_n^2)$ we will need to compute the expectation of $(S_n^2)^2$, where, according to the formula in (B.3),

$$(S_n^2)^2 = \frac{1}{n^2(n-1)^2} \sum\sum_{i<j} \sum\sum_{k<\ell} (X_i - X_j)^2 (X_k - X_\ell)^2.$$

It then follows by the linearity of the expectation operator that

$$E\left[(S_n^2)^2\right] = \frac{1}{n^2(n-1)^2} \sum\sum_{i<j} \sum\sum_{k<\ell} E\left[(X_i - X_j)^2 (X_k - X_\ell)^2\right]. \tag{B.4}$$

We will then to compute the expectations $E\left[(X_i - X_j)^2(X_k - X_\ell)^2\right]$ for all possible values of i, j, k, ℓ ranging from 1 to n such that $i < j$ and $k < \ell$. There are $\left[\dfrac{n(n-1)}{2}\right]^2$ of those terms contributing to the expectation of $(S_n^2)^2$ in (B.4). Out of those terms, $\dfrac{n(n-1)}{2}$, or $\dbinom{n}{2}$, are of the form

$$E\left[(X_i - X_j)^4\right], \tag{B.5}$$

where $i < j$. We compute the expectations in (B.5) as follows

$$\begin{aligned} E\left[(X_i - X_j)^4\right] &= E\left[(X_i - \mu_1 + \mu_1 - X_j)^4\right] \\ &= E\left[(X_i - \mu_1)^4 + 4(X_i - \mu_1)^3(\mu_1 - X_j) \right. \\ &\quad +6(X_i - \mu_1)^2(\mu_1 - X_j)^2 \\ &\quad\quad \left. +4(X_i - \mu_1)(\mu_1 - X_j)^3 + (\mu_1 - X_j)^4\right] \\ &= \mu_4 + 6\mu_2 \cdot \mu_2 + \mu_4, \end{aligned}$$

where we have used the independence of the X_is and the definition of the central moments. We then have that

$$E\left[(X_i - X_j)^4\right] = 2\mu_4 + 6\mu_2^2, \quad \text{for } i \neq j. \tag{B.6}$$

For the rest for the expectations, $E\left[(X_i - X_j)^2(X_k - X_\ell)^2\right]$, in (B.4) there are two possibilities

(i) $i \neq k$ and $j \neq \ell$, or

(ii) either $i = k$, or $j = \ell$, but not both simultaneously.

In case (i) we obtain, by the independence of the X_is and the definition of the central moments, that

$$E\left[(X_i - X_j)^2(X_k - X_\ell)^2\right] = E\left[(X_i - X_j)^2\right] \cdot E\left[(X_k - X_\ell)^2\right], \tag{B.7}$$

where

$$\begin{aligned} E\left[(X_i - X_j)^2\right] &= E\left[(X_i - \mu_1 + \mu_1 - X_j)^2\right] \\ &= E\left[(X_i - \mu_1)^2\right] + E\left[(X_j - \mu_1)^2\right] \end{aligned}$$

since

$$E\left[(X_i - \mu_1)(\mu_1 - X_j)\right] = E(X_i - \mu_1) \cdot E(\mu_1 - X_j) = 0.$$

Consequently,

$$E\left[(X_i - X_j)^2\right] = 2\mu_2.$$

Similarly,

$$E\left[(X_k - X_\ell)^2\right] = 2\mu_2.$$

We then have from (B.7) that

$$E\left[(X_i - X_j)^2(X_k - X_\ell)^2\right] = 4\mu_2^2, \quad \text{for } i \neq j \neq k \neq \ell. \tag{B.8}$$

There are

$$4!\binom{n}{4}$$

of making the choices for $i \neq j \neq k \neq \ell$. Since we are only interested in those choices with $i < j$ and $k < \ell$ we get a total of

$$6\binom{n}{4}$$

choices in case (i). Consequently the number of choices in case (ii) is

$$\binom{n}{2}^2 - \binom{n}{2} - 6\binom{n}{4}.$$

One of the expectations in case (ii) is of the form

$$E\left[(X_i - X_j)^2 (X_i - X_\ell)^2\right],$$

where $j \neq \ell$. In this case we have

$$\begin{aligned}
E\left[(X_i - X_j)^2(X_i - X_\ell)^2\right] &= E\left[(X_i - \mu_1 + \mu_1 - X_j)^2(X_i - \mu_1 + \mu_1 - X_\ell)^2\right] \\
&= E\big[\big((X_i-\mu_1)^2 + 2(X_i-\mu_1)(\mu_1 - X_j) + (\mu_1 - X_j)^2\big) \\
&\qquad \big((X_i-\mu_1)^2 + 2(X_i-\mu_1)(\mu_1 - X_\ell) + (\mu_1 - X_\ell)^2\big)\big] \\
&= E\big[(X_i-\mu_1)^4 + 2(X_i-\mu_1)^3(\mu_1 - X_\ell) \\
&\qquad +(X_i-\mu_1)^2(X_\ell-\mu_1)^2 + 2(X_i-\mu_1)^3(\mu_1 - X_j) \\
&\qquad +4(X_i-\mu_1)^2(\mu_1 - X_j)(\mu_1 - X_\ell) \\
&\qquad +2(X_i-\mu_1)(\mu_1 - X_j)(X_\ell - \mu_1)^2 \\
&\qquad +(X_i-\mu_1)^2(X_j-\mu_1)^2 \\
&\qquad +2(X_i-\mu_1)(X_j-\mu_1)^2(\mu_1 - X_\ell) \\
&\qquad +(X_j-\mu_1)^2(X_\ell-\mu_1)^2\big].
\end{aligned}$$

Next, use the linearity of the expectation operator, the independence of X_i, X_j and X_ℓ, and the definition of the central moments to get

$$\begin{aligned}
E\left[(X_i - X_j)^2(X_i - X_\ell)^2\right] &= \mu_4 + \mu_2 \cdot \mu_2 + \mu_2 \cdot \mu_2 + \mu_2 \cdot \mu_2 \\
&= \mu_4 + 3\mu_2^2.
\end{aligned}$$

We obtain the same value for all the other expectations in case (ii); i.e.,

$$E\left[(X_i - X_j)^2(X_i - X_\ell)^2\right] = \mu_4 + 3\mu_2^2, \quad \text{for } i \neq j \neq \ell. \tag{B.9}$$

It follows from (B.4) and the values of the possible expectations, $E\left[(X_i - X_j)^2(X_k - X_\ell)^2\right]$, we have computed in equations (B.6), (B.8) and (B.9), that

$$\begin{aligned}
E\left[n^2(n-1)^2(S_n^2)^2\right] &= \sum\sum_{i<j}\sum\sum_{k<\ell} E\left[(X_i - X_j)^2(X_k - X_\ell)^2\right] \\
&= \binom{n}{2}(2\mu_4 + 6\mu_2^2) + 6\binom{n}{4}(4\mu_2^2) \\
&\qquad + \left(\binom{n}{2}^2 - \binom{n}{2} - 6\binom{n}{4}\right)(\mu_4 + 3\mu_2^2).
\end{aligned}$$

Noting that $\displaystyle\binom{n}{2} = \frac{n(n-1)}{2}$, the above expression simplifies to

$$\begin{aligned}
E\left[n^2(n-1)^2(S_n^2)^2\right] &= n(n-1)(\mu_4 + 3\mu_2^2) + n(n-1)(n-2)(n-3)\mu_2^2 \\
&\qquad + n(n-1)(n-2)(\mu_4 + 3\mu_2^2).
\end{aligned}$$

Thus, dividing by $n(n-1)$ on both sides of the previous equation we then obtain that

$$\begin{aligned} E\left[n(n-1)(S_n^2)^2\right] &= \mu_4 + 3\mu_2^2 + (n-2)(n-3)\mu_2^2 + (n-2)(\mu_4 + 3\mu_2^2) \\ &= (n-1)\mu_4 + (n^2-2n+3)\mu_2^2. \end{aligned}$$

Dividing by $n-1$ we then have that

$$nE\left[(S_n^2)^2\right] = \mu_4 + \frac{n^2-2n+3}{n-1}\mu_2^2,$$

from which we obtain that

$$E\left[(S_n^2)^2\right] = \frac{1}{n}\left(\mu_4 + \frac{n^2-2n+3}{n-1}\mu_2^2\right).$$

Thus,

$$\begin{aligned} \operatorname{var}(S_n^2) &= E\left[(S_n^2)^2\right] - \left[E(S_n^2)\right]^2 \\ &= \frac{1}{n}\left(\mu_4 + \frac{n^2-2n+3}{n-1}\mu_2^2\right) - (\mu_2)^2, \end{aligned}$$

since S_n^2 is an unbiased estimator of μ_2. Simplifying we then obtain that

$$\operatorname{var}(S_n^2) = \frac{1}{n}\mu_4 + \frac{3-n}{n(n-1)}\mu_2^2,$$

which yields the equation in (B.1).

Bibliography

[CB01] G. Casella and R. L. Berger. *Statistical Inference.* Duxbury Press, second edition, 2001.

[Fer02] T. S. Ferguson. *A Course in Large Sample Theory.* Chapman & Hall/CRC, 2002.

[HCM04] R. V. Hogg, A. Craig, and J. W. McKean. *Introduction to Mathematical Statistics.* Prentice Hall, 2004.

[Pla83] R. L. Plackett. Karl pearson and the chi–square test. *International Statistical Review*, 51(1):59–72, April 1983.

[Stu08] Student. The probable error of a mean. *Biometrika*, 6(1):1–25, March 1908.