Lecture Notes for Math 448 Statistics

Vladislav Kargin

December 23, 2022

Contents

Chapter 1

Point Estimators

1.1 Basic problem of statistical estimation

Suppose we have a sample of data which was collected by observing a sequence of random experiments. Typically, this is a sequence of numbers $(x_1, x_2, \ldots, x_n)$, where x_i is a real number, but more generally every observation (a datapoint) can be a vector of numerical characteristics. Since experiments results in a random outcome, x_i is a realization of a random variable X_i, so a random sample is the sequence of the random variables

$$X_1, X_2, \ldots, X_n.$$

The main assumption of the mathematical statistics is that this sequence has a cumulative distribution function $F(\vec{x}, \theta)$ where θ is an unknown parameter, which can be any number (or a vector) in a region Θ. The main task is to obtain some information about this parameter.

As an example we can think about X_i as the number of Covid deaths on day i, or GPA of a student i and so on.

In this course, we assume for simplicity that the random variables X_i are i.i.d., independent and identically distributed, that is every datapoint has the same distribution as others and that they are independent of each other. It is a very restrictive requirement, for example, for Covid data it is doubtful that X_{i+1} is independent from X_i, however, this is the simplest setting in which we can develop the statistical theory.

For example, we can look at the sample of X_i, $i = 1, \ldots, n$, where X_i is a lifetime of a smartphone and model X_i as an exponential random variable with mean θ. Potentially, this θ can be any number in $\Theta = (0, \infty)$. Our task is for a specific realization of random variables X_i derive a conclusion about the parameter θ.

Our assumption means that the density of X_1 is

$$f_{X_1}(x_1) = \frac{1}{\theta} e^{x_1/\theta},$$

the density of X_2 is

$$f_{X_2}(x_2) = \frac{1}{\theta} e^{x_2/\theta},$$

and so on.

The joint density of independent datapoints is simply product of the individual densities for each datapoint. In our example,

$$\begin{aligned} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) &= \frac{1}{\theta} e^{x_1/\theta} \times \frac{1}{\theta} e^{x_2/\theta} \times \ldots \times \frac{1}{\theta} e^{x_n/\theta} \\ &= \frac{1}{\theta^n} e^{(\sum_{i=1}^n x_i)/\theta} \end{aligned}$$

In statistics, if we think about this joint density as a function of the model parameter θ, we call it the *likelihood function* and denote it by letter L. So, in our example, we have

$$L(\theta|\vec{x}) = \frac{1}{\theta^n} e^{(\sum_{i=1}^n x_i)/\theta},$$

where we used notation $\vec{x}$ to denote the vector of observed datapoints: $\vec{x} = (x_1, \ldots, x_n)$.

Now, we want to get some information about the parameter θ from the vector $(x_1, \ldots, x_n)$. For example, we could look for a function of $(x_1, \ldots, x_n)$ which would be close to θ. This is called the *point estimation problem* because we can try to find a point (an estimator) which would be close to θ. We will discuss it in the next section.

1.2 An estimator, its bias and variance

One of the main goals in statistics is to guess the value of an unknown parameter θ, given the realization of the data sample. Namely, we are given the realization of random variables $X_1, \ldots, X_n$, and we want to guess θ. Mathematically, this means that we look for a function of the $X_1, \ldots, X_n$, $f(X_1, \ldots, X_n)$, which we call an *estimator*.

A function of the data sample is called a *statistic*, so an estimator is a statistic. It can be any function whatsoever but naturally we want that happen to be a good guess for the true value of the parameter.

Note on notation: If θ is a parameter to be estimated, then $\widehat{\theta}$ denotes its estimator or a value of the estimator for a given sample. More carefully, it is function of the data: $\widehat{\theta} = \widehat{\theta}(X_1, \ldots X_n)$.

Examples of estimators: $\widehat{\theta} = \overline{X} := (X_1 + \ldots + X_n)/n$ or $\widehat{\theta} = X_{(n)} := \max(X_1, \ldots X_n)$. Even very unnatural functions such as $\sin(X_1 \times X_2 \times \ldots \times X_n)$ can be thought as estimators. So how do we distinguish between good and bad estimators?

What do we mean by saying that $\widehat{\theta}$ is a good guess for θ?

Note that $\widehat{\theta} = \widehat{\theta}(X_1, \ldots X_n)$ is random since its value changes from sample to sample. The distribution of this random variable $\widehat{\theta}$ depends on the true value of the parameter θ. One of the things that we can ask from the estimator is that its expected value equal to the true value of the parameter. This is called *unbiasedness*. The second useful property is that when we increase the size of the sample, the estimator converges to the true value of the parameter in the sense of convergence in probability. This is called *consistency*. We will deal with these two concepts one by one.

Bias of an estimator:

Def: $Bias(\hat{\theta}) = \mathbb{E}\widehat{\theta} - \theta$; (The bias of an estimator is its expected value minus the true value of the parameter).

Note that the bias can depend on the true value of the parameter. A good estimator should have zero or at least small bias for all values of the true parameter.

Definition 1.2.1. An estimator $\widehat{\theta} = \widehat{\theta}(X_1, \ldots, X_n)$ is called *unbiased*, if

$$\mathbb{E}\widehat{\theta}(X_1, \ldots, X_n) = \theta,$$

for every $\theta \in \Theta$.

In other words, the estimator $\widehat{\theta}$ is unbiased if its bias is zero for every value of the true parameter $\theta \in \Theta$.

Example 1.2.2. Consider our previous example about the lifetime of smartphones. What is the bias of the following two estimators: $\widehat{\theta} = \overline{X}$ and $\widehat{\theta} = X_1$?

Why does $\overline{X}$ appear to be better than X_1 as an estimator?

The reason is that the variance of $\overline{X}$ decreases as the sample size grows, while the variance of X_1 does not depend on the size of the sample.

Variance of an estimator

Def: $\mathbb{V}\mathrm{ar}(\hat{\theta}) = \mathbb{E}(\widehat{\theta} - \mathbb{E}\widehat{\theta})^2 = \mathbb{E}\widehat{\theta}^2 - (\mathbb{E}\widehat{\theta})^2$;

We want that $\mathbb{V}\mathrm{ar}(\hat{\theta})$ be small for all values of the true parameter θ.

Ideally, both the bias and the variance of the estimator should be small. Sometimes we value unbiasedness more than anything else. We want to make sure that an estimator is unbiased and only after this condition is satisfied we start to look for estimators with low variance among these unbiased estimators.

However, sometimes we can tolerate that an estimator is a bit biased. Moreover, in some cases it is very difficult or even impossible to find an unbiased estimator. In this case, it is useful to define a combined measure of the quality of an estimator.

Def: Mean Squared Error of an extimator is defined as

$$MSE(\widehat{\theta}) = \mathbb{E}\left\{(\widehat{\theta} - \theta)^2\right\}.$$

Theorem 1.2.3 (MSE decomposition)**.**

$$MSE(\widehat{\theta}) = \mathbb{V}\mathrm{ar}(\widehat{\theta}) + [Bias(\widehat{\theta})]^2$$

Proof. By using the linearity of expectation:

$$\begin{aligned}\mathbb{E}\Big[(\widehat{\theta}-\theta)^2\Big] &= \mathbb{E}\Big[(\widehat{\theta}-\mathbb{E}\widehat{\theta}+\mathbb{E}\widehat{\theta}-\theta)^2\Big] \\ &= \mathbb{E}(\widehat{\theta}-\mathbb{E}\widehat{\theta})^2 + 2\mathbb{E}\Big[(\widehat{\theta}-\mathbb{E}\widehat{\theta})(\mathbb{E}\widehat{\theta}-\theta)\Big] + (\mathbb{E}\widehat{\theta}-\theta)^2 \\ &= \mathbb{V}\mathrm{ar}(\widehat{\theta}) + [Bias(\widehat{\theta})]^2 + 2\mathbb{E}\Big[(\widehat{\theta}-\mathbb{E}\widehat{\theta})(\mathbb{E}\widehat{\theta}-\theta)\Big]\}\end{aligned}$$

But in the last term we can take $\mathbb{E}\widehat{\theta}-\theta$ outside of the expectation sign, since it is not random, and we find that this last term is zero:

$$\mathbb{E}\Big[(\widehat{\theta}-\mathbb{E}\widehat{\theta})(\mathbb{E}\widehat{\theta}-\theta)\Big] = (\mathbb{E}\widehat{\theta}-\theta)\mathbb{E}\Big[(\widehat{\theta}-\mathbb{E}\widehat{\theta})\Big] = 0,$$

because $\mathbb{E}\Big[(\widehat{\theta}-\mathbb{E}\widehat{\theta})\Big] = \mathbb{E}\widehat{\theta}-\mathbb{E}\widehat{\theta} = 0$.

□

Example 1.2.4. Suppose that for a certain estimator $\widehat{\theta}$ of θ we know that $\mathbb{E}\widehat{\theta} = a\theta + b$ for some constant $a \neq 0$ and $b \neq 0$.

- What is $Bias(\widehat{\theta})$, in terms of a, b and θ?
- Find a function of $\widehat{\theta}$ that is an unbiased estimator for θ.

Comment: if you find an biased estimator $\hat{\theta}$, you can sometimes easily correct the bias to get an unbiased estimator. However, e.g. if we tried an estimator $\widehat{\theta}$ and found that it has $\mathbb{E}\widehat{\theta} = \sqrt{\theta}$, so we cannot correct the bias by simply taking the square of $\widehat{\theta}$. **The estimator $\tilde{\theta} = \hat{\theta}^2$ will not not unbiased for θ!** If we recall the formula for the second moment of the random variable, then in this particular example we can even compute the bias:

$$\mathbb{E}\hat{\theta}^2 = (\mathbb{E}\hat{\theta})^2 + \mathbb{V}\mathrm{ar}(\hat{\theta}) = \theta + \mathbb{V}\mathrm{ar}(\hat{\theta}),$$

so the bias of the estimator $\hat{\theta}^2$ equals $\mathbb{V}\mathrm{ar}(\hat{\theta})$. In general, it is often quite difficult to find an unbiased estimator.

Let us look at a couple of examples.

Example 1.2.5. The reading on a voltage meter connected to a test circuit is uniformly distributed over the interval $(\theta, \theta+1)$, where θ is the true but unknown voltage of the circuit. Suppose that $Y_1, Y_2, \ldots, Y_n$ denote a random sample of such readings. We are going to try two estimators of θ, $\widehat{\theta} = \overline{Y}$ and $\widehat{\theta} = \min\{Y_1, \ldots, Y_n\}$. First, consider $\widehat{\theta} = \overline{Y}$.

- Calculate the bias of $\overline{Y}$ as an estimator of θ.
- Find an unbiased estimator of θ (based on $\overline{Y}$).
- Find $MSE(\overline{Y})$.

Solution. It is straightforward to calculate the bias:

$$\begin{aligned} bias(\overline{Y}) = \mathbb{E}(\overline{Y}) - \theta &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(Y_i) - \theta \\ &= \mathbb{E}(Y) - \theta = 1/2, \end{aligned}$$

where we used the fact, that the expectation of a r.v. distributed on $[\theta, \theta+1]$ equals $\theta + 1$.

Then $MSE(\overline{Y}) = bias(\overline{Y})^2 + \mathbb{V}\mathrm{ar}(\overline{Y})$, and since Y_i independent,

$$\mathbb{V}\mathrm{ar}(\overline{Y}) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\mathrm{ar}(Y_i) = \frac{1}{n}\mathbb{V}\mathrm{ar}(Y).$$

In order to calculate the variance of Y, which is uniform on $[\theta, \theta+1]$, note that the variance of the shifted variable is the same, so we can calculate variance of X which is uniform on $[-1/2, 1/2]$,

$$\mathbb{V}\mathrm{ar}(X) = \int_{-1/2}^{1/2} x^2 dx = \Big|_{-1/2}^{1/2} \frac{x^3}{3} = \frac{1}{12}.$$

Altogether $MSE(\overline{Y}) = \frac{1}{4} + \frac{1}{12n}$.

Now we want to use the other estimator, $\widehat{\theta} = \min\{Y_1, \ldots, Y_n\}$.

Example 1.2.6. The sample values $Y_1, Y_2, \ldots, Y_n$ are uniform on $(\theta, \theta+1)$. Consider the estimator $\widehat{\theta} = Y_{(1)} := \min\{Y_1, \ldots, Y_n\}$. Calculate the bias and variance of $\widehat{\theta}$. Can we correct the bias?

Reminder about the distribution of the minimum and the minimum. Recall the notation $Y_{(1)} = \min\{Y_1 \dots Y_n\}$. Then for the CDF of $Y_{(1)}$, we have:

$$\begin{aligned} F_{Y_{(1)}}(y) \equiv \Pr(Y_{(1)} \le y) &= 1 - \Pr(Y_{(1)} > y) \\ &= 1 - \Pr(\text{all } Y_i\text{'s are} > y) \\ &= 1 - [1 - F(y)]^n \end{aligned}$$

And the PDF is

$$f_{Y_{(1)}}(y) = n[1 - F(y)]^{n-1} f(y).$$

Similarly for the maximum we have notation $Y_{(n)} = \max\{Y_1 \dots Y_n\}$. The CDF is

$$\begin{aligned} F_{Y_{(n)}}(y) \equiv \Pr(Y_{(n)} \le y) &= \Pr(Y_i \le y, \text{ for all } i) \\ &= [F(y)]^n \end{aligned}$$

and the PDF is

$$f_{Y_{(n)}}(y) = n[F(y)]^{n-1} f(y).$$

Now let us return to the example.

We want to calculate $\mathbb{E}Y_{(1)}$. It is convenient to define shifted variables $X_i = Y_i - \theta$, since then $\mathbb{E}Y_{(1)} = \mathbb{E}X_{(1)} + \theta$ and it is easier to calculate $\mathbb{E}X_{(1)}$ because X_i are simply uniform random variables on $[0, 1]$. (Of course, the expectation can be calculated without this transformation but the formulas would be more cumbersome.)

Then, since the density and cdf of X_i are $f_X(x) = 1$ and $F_X(x) = x$ supported on $[0, 1]$, then we can use the formulas from above and calculate the pdf of the minimum $X_{(1)}$, $f_{X_{(1)}}(x) = n(1 - x)^{n-1}$. In other words, $X_{(1)}$ has Beta distribution with parameters $\alpha = 1$ and $\beta = n$. By the facts about the Beta distribution, it follows that the expectation is $\mathbb{E}X_{(1)} = \alpha/(\alpha+\beta) = 1/(n + 1)$.

Alternatively, we can simply integrate using the density of $X_{(1)}$, and calculate

$$\mathbb{E}X_{(1)} = n \int_0^1 x(1 - x)^{n-1} dx = \frac{1}{n + 1}.$$

(The integral can be calculated by doing integration by parts or by using a very useful formula for Beta integrals:

$$\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)},$$

where $\Gamma(x)$ is the Gamma function. For integer argument x, $\Gamma(x) = (x-1)!$.)

Hence the bias of $Y_{(1)} = \mathbb{E}Y_{(1)} - \theta = \mathbb{E}X_{(1)} = \frac{1}{n+1}$. Note that in this example the bias $\to 0$ as the sample size increases. In addition, we can easily correct the bias by using $\widehat{\theta} = Y_{(1)} - \frac{1}{n+1}$.

What is the MSE of $\widehat{\theta} = Y_{(1)} - \frac{1}{n+1}$?

Since there is no bias, we only need to calculate $\mathbb{V}\mathrm{ar}(\widehat{\theta}) = \mathbb{V}\mathrm{ar}(Y_{(1)}) = \mathbb{V}\mathrm{ar}(X_{(1)})$.

From the facts about the Beta distribution we have

$$\mathbb{V}\mathrm{ar}(X_{(1)}) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{n}{(n+1)^2(n+2)} \sim \frac{1}{n^2}$$

for large n. (We could also calculate it directly from the density.)

This is the MSE of $\widehat{\theta} = Y_{(1)} - \frac{1}{n+1}$. For large n, it is much smaller than the MSE of the estimator $\overline{Y} - \frac{1}{2}$, which we calculated as $\frac{1}{12n}$.

Example 1.2.7. Calculate the distribution of the minimum for the sample $X_1, \ldots, X_n$ from the exponential distribution with parameter θ. Use the minimum to obtain an unbiased estimate of the parameter θ. What is the variance of this estimator?

Solution First, we calculate the CDF of each observation as

$$F_{X_i}(x) = \int_0^x \frac{1}{\theta} e^{-t/\theta} dt = 1 - e^{-x/\theta}.$$

Then, by using formulas above, we calculate the density of the minimum:

$$\begin{aligned} f_{X_{(1)}}(x) &= n\Big(e^{-x/\theta}\Big)^{n-1} \times \frac{1}{\theta} e^{-x/\theta} \\ &= \frac{n}{\theta} e^{-nx/\theta}. \end{aligned}$$

Hence, the minimum $X_{(1)} = \min\{X_1, \ldots, X_n\}$ is distributed as the exponential random variable with parameter θ/n.

If we set $\widehat{\theta} = nX_{(1)}$, then the expectation of this estimator is θ and it gives an unbiased estimator of θ.

What is its variance?

$$\mathbb{V}\mathrm{ar}\big(nX_{(1)}\big) = n^2\mathbb{V}\mathrm{ar}X_{(1)} = n^2(\theta/n)^2 = \theta^2,$$

so it is not a particularly good estimator of θ. Its variance does not decline as the sample size grows.

Summary: In this section we introduced simple measures that help us to evaluate how good an estimator is, – its bias, variance and mean squared error.

1.3 Consistency

Suppose again that we have a sample $(X_1, \dots X_n)$ from a probability distribution that depends on parameter θ. Note that although we speak about *an* estimator $\widehat{\theta} = \widehat{\theta}(X_1, \dots, X_n)$, in fact the distribution of the estimator depends on n, so it would be more correctly speak about a sequence of random variables $\widehat{\theta}_n$.

Usually, we expect that when the size of the sample becomes larger, that is, n grows, the distribution of the estimator $\widehat{\theta}_n$ become concentrated more and more around the true value of the parameter θ. This is the *minimal requirement* that we can impose on the family of estimators that depend on sample size. If this requirement is not satisfied as in Example 1.2.7 above, then the estimator is not very useful. Technically this property of an estimator is called *consistency* and we are giving its mathematical definition below.

Before that, let us look at some pictures. Plots show a simulation study. A sample $X_1, X_2, \dots$ from the distribution $N(\theta, 1/4)$ was generated with $\theta = 10$ and we computed $\widehat{\theta}_k = (X_1 + \dots X_k)/k$. Figure 1.1 shows a path of $\widehat{\theta}_k$. It suggests that if we get more and more data, $\widehat{\theta}_k$ converges to the true value of θ. In fact, this is a consequence of the strong law of large numbers, which says that this behavior is observed with probability 1.

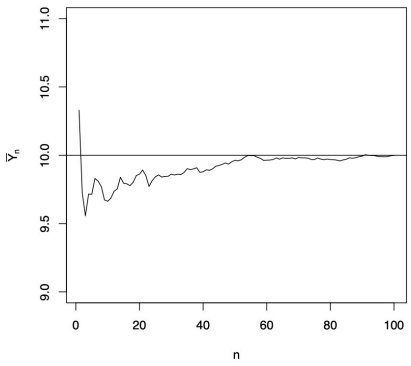


Figure 1.1

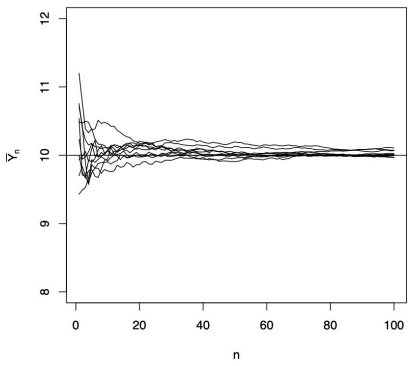


Figure 1.2

What about several different samples? Figure 1.2 shows the situation when the sample $X_1, X_2, \ldots$ was generated 10 times and 10 paths of $\widehat{\theta}_k$ were plotted. This picture suggest that when the sample size grows the distribution of $\widehat{\theta}_k$ around the true value of the parameter θ. Mathematically this is a consequence of the weak law of large numbers.

In order to define the consistency, recall what it means for a sequence of random variables to converge to another random variable.

Definition 1.3.1 (Convergence in probability)**.** A sequence of random variables, $X_1, X_2, \ldots, X_n, \ldots$, is convergent **in probability** to a random variable X if, for any $\epsilon > 0$, as $n \to \infty$

$$\mathbb{P}(|X_n - X| < \epsilon) \to 1,$$

that is,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| < \epsilon) = 1$$

This is denoted either as $X_n \xrightarrow{\mathbb{P}} X$ or as $\text{plim}_{n\to\infty} X_n = X$.

Note that $a_n \equiv \mathbb{P}(|X_n - X| < \epsilon)$ is simply a number (it is not random). Hence, $\{a_1, a_2, \ldots, a_n, \ldots\}$ form a sequence of numbers, and their limit is defined in the usual "calculus" sense.

Definition 1.3.2 (Consistency)**.** An estimator $\widehat{\theta}_n$ is a consistent estimator of θ, if $\widehat{\theta}_n$ converges in probability to θ

$$\widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta.$$

By writing out the definition of the convergence in probability in detail, we see that this definition can be also written as saying that an estimator $\widehat{\theta}_n$ is a consistent estimator of θ, if for any $\varepsilon > 0$, as $n \to \infty$,

$$\mathbb{P}(|\widehat{\theta}_n - \theta| < \varepsilon) \to 1.$$

The consistency of the estimator means that as the sample size goes to infinity, we are become more and more sure that the distance between $\widehat{\theta}_n$ and θ is smaller than any positive ε !

Consistency describes a property of the estimator in the $n \to \infty$ limit. Unlike unbiasedness, it is NOT meant to describe the property of the estimator for a fixed n.

An unbiased estimator can be inconsistent as we can see in Example 1.2.7, and a biased estimator can be consistent (as $Y_{(1)}$ in Example 1.2.6)! Consistency is more important than unbiasedness because it ensures that if collect enough data we will eventually learn the true value of the parameter.

So, how can we tell if the estimator is consistent? One way is to see how MSE changes with n.

Theorem 1.3.3. *If $MSE(\widehat{\theta}_n) \to 0$ as $n \to \infty$, then the estimator $\widehat{\theta}_n$ is consistent.*

Proof. By using Theorem 1.2.3, we note that $MSE(\widehat{\theta}_n) \to 0$ if and only if $bias(\widehat{\theta}_n) \to 0$ and $\mathbb{V}\text{ar}(\widehat{\theta}_n) \to 0$. Fix an $\varepsilon > 0$ and choose n_0 so that $|bias(\widehat{\theta}_n)| < \varepsilon/2$ for all $n > n_0$. Then, by definition of bias, for all $n > n_0$, $|\mathbb{E}\widehat{\theta}_n - \theta| < \varepsilon/2$. Since

$$\begin{aligned} |\widehat{\theta}_n - \theta| &= |\widehat{\theta}_n - \mathbb{E}\widehat{\theta}_n + \mathbb{E}\widehat{\theta}_n - \theta| \leq |\widehat{\theta}_n - \mathbb{E}\widehat{\theta}_n| + |\mathbb{E}\widehat{\theta}_n - \theta| \\ &< |\widehat{\theta}_n - \mathbb{E}\widehat{\theta}_n| + \varepsilon/2, \end{aligned}$$

therefore the event $|\widehat{\theta}_n - \theta| > \varepsilon$ can occur only if $|\widehat{\theta}_n - \mathbb{E}\widehat{\theta}_n| > \varepsilon/2$ occured. Hence, for $n > n_0$, $\mathbb{P}(|\widehat{\theta}_n - \theta| > \varepsilon) \leq \mathbb{P}(|\widehat{\theta}_n - \mathbb{E}\widehat{\theta}_n| > \varepsilon/2)$.

Now apply the Chebyshev inequality,

$$\mathbb{P}(|\widehat{\theta}_n - \mathbb{E}\widehat{\theta}_n| > \varepsilon/2) \leq \frac{\mathbb{V}\mathrm{ar}(\widehat{\theta}_n)}{(\varepsilon/2)^2}$$

By our assumption, the right-hand side can be made arbitrarily small for all sufficiently large n because $\mathbb{V}\mathrm{ar}(\widehat{\theta}_n) \to 0$. We showed that $\mathbb{P}(|\widehat{\theta}_n - \theta| > \varepsilon) \to 0$ for any $\varepsilon > 0$. □

- If $Bias(\widehat{\theta}_n) \to 0$ as $n \to \infty$, then the estimator is called asymptotically unbiased.
- Another way to formulate the theorem is to say that any estimator which is asymptotically unbiased and has its variance converging to 0 as $n \to \infty$ is a consistent estimator.

Example 1.3.4 (Sample mean is a consistent estimator of the population mean). Let $Y_1, Y_2, \ldots$ be a sample from a population with mean μ and variance σ^2.

- Sample mean $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$. Its expectation is μ and its variance is $\mathbb{V}\mathrm{ar}(\overline{Y}_n) = \sigma^2/n \to 0$.
- So $\overline{Y}_n$ is an unbiased and consistent estimator of μ.

Example 1.3.5 (Biased and consistent estimator of the mean). For parameter $\theta = \mu$, consider a modified sample mean $\widehat{\theta}_n = \frac{n}{n-1}\overline{Y}$

- $Bias(\widehat{\theta}) = \mathbb{E}\widehat{\theta}_n - \theta = \frac{n}{n-1}\mu - \mu = \frac{1}{n-1}\mu \to 0$ as $n \to \infty$; ($\widehat{\theta}_n$ is a biased estimator of $\mu \neq 0$ for every n. It is, however, asymptotically unbiased)
- $\mathbb{V}\mathrm{ar}(\widehat{\theta}) = \frac{n^2}{(n-1)^2}\sigma^2/n \to 0$.
- Conclusion: $\widehat{\theta}_n$ is a biased but consistent estimator.

Example 1.3.6.
- $Y_i \sim Unif[\theta, \theta + 1]$
- $\widehat{\theta}_1 = \overline{Y} - 1/2$; $\widehat{\theta}_2 = Y_{(1)} - 1/(n+1)$

- We have shown in Examples 1.2.5 and 1.2.6 that these estimators are both unbiased.
- In addition we showed in these examples that $\mathbb{V}\mathrm{ar}(\widehat{\theta}_1) = 1/(12n)$ and $\mathbb{V}\mathrm{ar}(\widehat{\theta}_2) = n/[(n+1)^2(n+2)]$. Since both variances go to zero as n grows, both estimators are consistent.

Our main tool in establishing consistency of estimators was Theorem . However, it is sometimes cumbersome to calculate MSE of an estimator. There are some other tools to establish consistency of an estimator. We will talk about them later.

Unbiasedness and consistency

- Unbiasedness: concerns expectation; for fixed n
- Consistency:
  - only care about $n \to \infty$;
  - concerns bias and variance (and whether they vanish for large n);
  - however, does not necessarily imply unbiasedness for finite n.

1. Can biased estimator be consistent? Yes!
2. Can unbiased estimator be inconsistent? Yes!

1.4 Some common unbiased estimators

1.4.1 An estimator for the population mean μ

Let $Y_1, Y_2, \ldots, Y_n$ denote a random sample of n independent identically distributed observations from a population with mean μ (that is, in our statistical model one of the parameters is $\mathbb{E}Y_i = \mu$) and variance σ^2 (another parameter is $\mathbb{V}\mathrm{ar}(Y_i) = \sigma^2$). Then the most natural estimator for μ is the *sample mean*:

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

- Expectation and bias of the estimator:
  $\mathbb{E}(\overline{Y}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}Y_i = \mathbb{E}Y_i = \mu$; so $bias(\overline{Y}) = 0$.
- Variance: $\mathbb{V}\text{ar}(\overline{Y}) = \frac{1}{n}\mathbb{V}\text{ar}(Y_i) = \frac{\sigma^2}{n}$;
- $MSE(\overline{Y}) = bias^2 + \mathbb{V}\text{ar} = \frac{\sigma^2}{n}$

These observations shows that the sample mean is an unbiased and consistent estimator for the population mean μ.

Definition 1.4.1. The standard error of an estimator is the square root of its variance $SE(\widehat{\theta}) = \sqrt{\mathbb{V}\text{ar}(\widehat{\theta})}$

Another notation for the variance and the standard error of an estimator $\widehat{\theta}$ is $\sigma^2_{\widehat{\theta}}$ and $\sigma_{\widehat{\theta}}$, respectively.

So the variance of the sample mean $\overline{Y}$ is $\sigma^2_{\overline{Y}} = \sigma^2/n$ and the standard error is $\sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}}$.

1.4.2 An estimator for the population proportion p

Let $Y_1, Y_2, \ldots, Y_n$ denote a random sample of size n from a population with a Bernoulli distribution

$$P(Y_i = 1) = p, \quad P(Y_i = 0) = 1 - p.$$

This is a special case of the situation in the previous section and we can use the same estimator, the sample mean. In this case, the sample mean has a special name, the sample proportion. It is an unbiased and consistent estimator for the parameter p (i.e., for the population proportion).

$$\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

that is, it is the proportion of $Y_i = 1$'s in the sample.

The estimator is unbiased since $\mathbb{E}(\widehat{p}) = \mathbb{E}Y_1 = p$. Its variance and standard error are $\sigma^2_{\widehat{p}} = \frac{pq}{n}$ and $\sigma_{\widehat{p}} = \sqrt{\frac{pq}{n}}$, respectively, with $q = 1 - p$.

1.4.3 An estimator for the difference in population means $\mu_1 - \mu_2$

Suppose we have two samples:

- $\{Y_1^{(1)}, Y_2^{(1)}, \ldots, Y_{n_1}^{(1)}\}$ of size n_1 from Population 1: with mean μ_1 and variance σ_1^2;
- $\{Y_1^{(2)}, Y_2^{(2)}, \ldots, Y_{n_2}^{(2)}\}$ of size n_2 from Population 2: with mean μ_2 and variance σ_2^2;

An unbiased estimator for the difference in population means $\theta = \mu_1 - \mu_2$ (it is the parameter of interest) is the

$$\text{difference in sample means:} \quad \hat{\theta} = \overline{Y}_1 - \overline{Y}_2 = \frac{1}{n_1}\sum_{i=1}^{n_1} Y_i^{(1)} - \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i^{(2)}.$$

- Expectation: $\mathbb{E}(\hat{\theta}) = \mathbb{E}\overline{Y}_1 - \mathbb{E}\overline{Y}_2 = \mathbb{E}Y_1^{(1)} - \mathbb{E}Y_1^{(2)} = \mu_1 - \mu_2$;
- Variance: $\sigma_{\hat{\theta}}^2 = Var(\hat{\theta}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$;
  - Proof: Since the two samples are independent, $\mathbb{V}\text{ar}(\overline{Y}_1 - \overline{Y}_2) = \mathbb{V}\text{ar}(\overline{Y}_1) + \mathbb{V}\text{ar}(\overline{Y}_2)$.
- Standard error: $\sigma_{\hat{\theta}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

It follows that this estimator is unbiased (its expectation equals the estimated parameter) and consistent (because its MSE converges to zero as n_1 and n_2 jointly grow).

1.4.4 An estimator for the difference in population proportions $p_1 - p_2$

Suppose we have two samples:

- $\{Y_1^{(1)}, Y_2^{(1)}, \ldots, Y_{n_1}^{(1)}\}$ of size n_1 from Population 1: $P(Y_i^{(1)} = 1) = p_1$ and $P(Y_i^{(1)} = 0) = 1 - p_1$;
- $\{Y_1^{(2)}, Y_2^{(2)}, \ldots, Y_{n_1}^{(2)}\}$ of size n_2 from Population 2: $P(Y_i^{(2)} = 1) = p_2$ and $P(Y_i^{(2)} = 0) = 1 - p_2$;

An unbiased point estimator for the difference in population means $\theta = p_1 - p_2$ (the parameter of interest) is the

difference in sample proportions: $\hat{\theta} = \hat{p}_1 - \hat{p}_2 = \frac{1}{n_1}\sum_{i=1}^{n_1} Y_i^{(1)} - \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i^{(2)}$

- Expectation: $\mathbb{E}(\hat{\theta}) = \mathbb{E}(\hat{p}_1 - \hat{p}_2) = p_1 - p_2 = \theta$;
- Variance: $\sigma^2_{\hat{\theta}} = Var(\hat{\theta}) = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$;
- Standard error: $\sigma_{\hat{\theta}} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$.

where $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$.

Obviously, this estimator is also unbiased and consistent.

1.4.5 An estimator for the variance

Let $Y_1, Y_2, \ldots, Y_n$ denote a sample of size n from a population with mean μ and variance σ^2, then an unbiased estimator for σ^2 is the sample variance:

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

Note that we divide by $n-1$, not by n, as it would seem most natural. If we divided by n the estimator would not be unbiased.

The square root of S^2 is called the sample standard deviation and denoted, as could be expected, S.

Theorem 1.4.2. *S^2 is an unbiased estimator for σ^2.*

Proof.

$$\mathbb{E}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \mathbb{E}\Big[\sum_{i=1}^{n} Y_i^2 - 2\Big(\sum_{i=1}^{n} Y_i\Big)\overline{Y} + n\overline{Y}^2\Big] = \mathbb{E}\Big[\sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2\Big]$$

$$= \sum_{i=1}^{n} \mathbb{E}Y_i^2 - n\mathbb{E}\big(\overline{Y}^2\big)$$

For the first term, we have:

$$\sum_{i=1}^{n} \mathbb{E}Y_i^2 = \sum_{i=1}^{n}[\mu^2 + \sigma^2] = n\mu^2 + n\sigma^2$$

For the second term:

$$\begin{aligned} n\mathbb{E}\left(\overline{Y}^2\right) &= n[\mathbb{V}\mathrm{ar}\overline{Y} + (\mathbb{E}\overline{Y})^2] = n(\sigma^2/n + \mu^2) \\ &= \sigma^2 + n\mu^2 \end{aligned}$$

Therefore,

$$\mathbb{E}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = (n-1)\sigma^2$$

Hence

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

and therefore S^2 is unbiased for σ^2. □

The variance of this estimator is more complicated to compute and we will not do it here. However, it turns out that it goes to zero as $n \to \infty$. In particular this estimator is consistent.

Note that although S^2 is a unbiased estimator for σ^2, the estimator S, that is, the square root of S^2, is NOT an unbiased estimator for σ. However, it turns out that it is still a consistent estimator for σ.

The identity used in the first line of the proof,

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

is an "empirical" analogue of the identity $\mathbb{V}\mathrm{ar}(Y) = \mathbb{E}(Y^2) - (\mathbb{E}Y)^2$.

1.5 The existence of unbiased estimators

We have seen above that several natural parameters have unbiased estimators. So one question is whether it is always possible to find an unbiased estimator for a parameter of interest. Surprisingly, the answer is "no". Here we present an example. It is a little bit artificial but it shows that sometimes it is not simply difficult to find an unbiased estimator.

In this example, each observation is taken from the Bernoulli distribution with parameter p. That is, $X_i = 1$ with probability p and $X_i = 0$ with probability $q = 1 - p$. Of course, we have seen that there is an unbiased estimator for p, namely $\widehat{p} = \overline{X}$. The twist of this example is that we try to estimate $\theta = -\ln p \in \Theta = (0, \infty)$. Suppose, by seeking contradiction, that $\widehat{\theta} = \widehat{\theta}(X_1, \dots X_n)$ is an unbiased estimator of θ and therefore, $\mathbb{E}\widehat{\theta} = \theta = -\ln p$. We will write the expectation by using the basic definition:

$$\mathbb{E}\widehat{\theta}(X_1, \dots, X_n) = \sum_{x_1=0}^{1} \dots \sum_{x_n=0}^{1} \widehat{\theta}(x_1, \dots, x_n)\mathbb{P}(X_1 = x_1, \dots, X_n = x_n).$$

For Bernoulli r.v., we can write $\mathbb{P}(X_i = x_i) = p^{x_i}(1-p)^{1-x_i}$, where x_i can take only two values, 0 and 1. So, by independence of random variables $X_1, \dots, X_n$, we have:

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}.$$

So if $\widehat{\theta}$ is unbiased, then we have:

$$-\ln p = \sum_{x_1=0}^{1} \dots \sum_{x_n=0}^{1} \widehat{\theta}(x_1, \dots, x_n) p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i},$$

and this should be true for every $p \in (0, 1)$ because the estimator is assumed to be unbiased for every $-\ln p \in (0, \infty)$. However this means that the logarithmic function of p equals to a polynomial in p. This is impossible. For example, the limit of the left-hand side for $p \to 0$ is ∞ while the limit the right hand side is finite.

We got a contradiction, so that means that there is no unbiased estimator of $\theta = -\ln p$.

1.6 The error of estimation and the 2-standard-error bound

Definition 1.6.1. The *error of estimation* ε is the distance between an estimator and its target parameter. That is, $\varepsilon = |\widehat{\theta} - \theta|$.

The error of estimation is a *random quantity* that change from sample to sample. We often interested to have a good bound on this quantity that holds with large probability.

Recall that the *standard error of the estimator* $\sigma_{\widehat{\theta}}$ is another name for the standard deviation of the estimator $\hat{\theta}$. That is $\sigma_{\widehat{\theta}} = \sqrt{\mathbb{V}\mathrm{ar}(\widehat{\theta})}$.

By the Chebyshev inequality:

$$\mathbb{P}\{|\widehat{\theta} - \theta| > k\sigma_{\widehat{\theta}}\} \leq \frac{1}{k^2}$$

- For $b = 2 \cdot \sigma_{\widehat{\theta}}$, the RHS of the Chebyshev inequality is $= 25\%$. [This is a bound, the true probability that $|\varepsilon| \geq b$ is smaller, often as small as 5%]

- $2\sigma_{\widehat{\theta}}$ is called the 2-standard-error bound on the error of the estimator. The meaning is that with large probability the error of estimation is smaller than $2\sigma_{\widehat{\theta}}$. (This probability is bounded from above by 75% by Chebyshev inequality and often significantly smaller).

The Central Limit Theorem for sums of independent random variables says that a sum of large number of these variables has a distribution, which is closed to the normal distribution.

Since the estimator for the mean, $\overline{Y}$ is such a sum (only divided by n), it becomes approximately normal when n is large, so if sample size is large, then the estimation error, $|\overline{Y} - \mu|$, is less than 2-standard-error, $2\sigma_{\overline{Y}}$, with probability 95% (instead of 75%).

This observation holds also for the other standard estimators that we considered in the previous section.

Example 1.6.2 (Titanic survivors). In a random sample of 136 Titanic first class passengers that survived the Titanic ship accident, 91 were women. In a random sample of 119 third class survivors, 72 were women. Assume that these are small samples from two large populations of "survivors": first-class survivors and third-class survivors.

What is an unbiased estimate for the difference in proportions of females in these populations? What is the two-standard error bound?

Solution. $\widehat{p}_1 = 66.9\%$; $\widehat{p}_3 = 60.5\%$; $\widehat{p}_1 - \widehat{p}_3 = 6.4\%$

Two standard error bound is:

$$2\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_3(1-\widehat{p}_3)}{n_3}} = 12.1\%$$

Does the data suggest that the total first class and third class survivor populations had approximately the same proportions of females?

Does the data suggest that women from the first and third classes had approximately the same chances to survive?

Solution: Not really. These data say that proportions of women among the first and the third class survivors are approximately the same, meaning that the difference between proportions is within the two-standard error bias. However, this does not really say anything about the chances of survival.

Example 1.6.3 (Titanic survivors II). In a random sample of 95 female passengers in the first class, 91 survived the Titanic ship accident. In a random sample of 145 women in the third class, 72 survived.

What is an unbiased estimate for the difference in proportions of survivors in the populations of the first and the third class female passengers? What is the two-standard error bound?

Solution. $\widehat{p}_1 = 95.8\%$; $\widehat{p}_3 = 49.7\%$; $\widehat{p}_1 - \widehat{p}_3 = 46.1\%$

Two standard error bound is:

$$2\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_3(1-\widehat{p}_3)}{n_3}} = 9.3\%$$

Does the data suggest that female passengers from the third class had lower chances to survive than female passengers from the first class?

Solution: Yes, the difference is far larger than the two standard error bound suggesting, that it is highly unlikely that this happened by chance.

Some additional interesting info about this example: The chances of a man in the first class to survive: 36.9%. The chances of a man in the third class: 13.5%.

Example 1.6.4 (Elementary school IQ). The article "A Longitudinal Study of the Development of Elementary School Children's Private Speech" by

Bivens and Berk (Merrill-Palmer Q., 1990: 443–463) reported on a study of children talking to themselves (private speech).

[The study was motivated by theories in psychology that claim that private speech plays an important role in a child mental development, so one can investigate how private speech is related to IQ (that is, performance on math or verbal tasks) or to changes in IQ. The study found some supporting evidence that task-related private speech is positively related to future changes in IQ. Here we are only interested in IQ data.]

The study included 33 students whose first-grade IQ scores are given here:

082 096 099 102 103 103 106 107 108 108 108 108 109 110 110 111 113 113 113 113 115 115 118 118 119 121 122 122 127 132 136 140 146

(a) Suppose we want an estimate of the average value of IQ for the first graders served by this school. What is an unbiased estimate for this parameter?

[Hint: Sum is 3753.]

Solution. $\widehat{\mu} = \overline{X} = 3753/33 = 113.7273$

(b) Calculate and interpret a point estimate of the population standard deviation σ. [Hint: Sum of squared observations is $432,015$]

Solution.

$$S^2 = \frac{1}{32}\Big(432,015 - 33 \times (113.7273)^2\Big) = 162.3856$$
$$\widehat{\sigma} = S = \sqrt{162.3856} = 12.7431$$

While S^2 is an unbiased estimator of σ^2, the estimator S for σ is biased.

(c) What is the two-standard-error bound on the error of estimation?

Solution.

The two-standard-error bound is $2\sigma_{\widehat{\mu}} = 2\sigma_{\overline{X}} = 2\sigma/\sqrt{n}$. We use the estimate of σ to calculate the bound.

$$2S/\sqrt{n} = 2 \times 12.7431/\sqrt{33} = 4.4366$$

Since the estimate of μ is 113.7273 and the two-error bound for the error

of estimation is 4.4366, the data suggest that this is an above average class, because the nationwide IQ average is around 100.

(d) Calculate a point estimate of the proportion of all such students whose IQ exceeds 100. [Hint: Think of an observation as a "success" if it exceeds 100.]

Solution.

The number of students with IQ above 100 is 30. So the point estimate is $\widehat{p} = 30/33 = 90.91\%$.

Example 1.6.5 (Elementary school IQ II). The data set mentioned in the previous example also includes these third grade verbal IQ observations for males:

117 103 121 112 120 132 113 117 132 149 125 131 136 107 108 113 136 114

(18 observations) and females:

114 102 113 131 124 117 120 90 114 109 102 114 127 127 103

(15 observations)

Let the male values be denoted $X_1, \ldots, X_m$ and the female values $Y_1, \ldots, Y_n$.

(a) Calculate the point estimate for the difference between male and female verbal IQ.

Solution.

$$\overline{X} - \overline{Y} = \frac{2186}{18} - \frac{1707}{15} = 121.4444 - 113.8 = 7.6444$$

(b) What is the standard error of the estimator?

Solution. First we calculate the sample variances S_x^2 and S_y^2 for these two samples.

$$S_x^2 = \frac{1}{17}\Big(268,046 - 18 \times 121.4444^2\Big) = 151.0964$$
$$S_y^2 = \frac{1}{14}\Big(196,039 - 15 \times 113.8^2\Big) = 127.3143$$

Then, we calculate the estimate of the standard error:

$$\widehat{\sigma}_{\widehat{\theta}} = \sqrt{\frac{S_x^2}{m} + \frac{S_y^2}{n}} == \sqrt{\frac{151.0964}{18} + \frac{127.3143}{15}} = 4.1088$$

So we see that the estimate of the difference 7.6444. However, the two-standard-error bound is 8.2176 and the data does not give an evidence that the difference is positive.

Chapter 2

Interval estimators

2.1 Confidence intervals and pivotal quantities

A **point estimator** is a function of data sample that gives a single number that is our "best guess" for the parameter, for examples: $\overline{Y}$ for μ and $\hat{p}$ for p. Often we want to know how far is the estimator from the true value of the parameter. While the standard error of the estimator gives some idea about its quality, in this section, we will talk about a related and more precise concept.

Definition 2.1.1. A **confidence interval** with confidence level $(1-\alpha)$ is a *random* interval $[\widehat{\theta}_L, \widehat{\theta}_U]$ such that

$$\mathbb{P}(\theta \in [\widehat{\theta}_L, \widehat{\theta}_U]) = 1 - \alpha.$$

(Confidence level is sometimes called confidence coefficient.)

The interval is random because both ends of the interval are functions of the data sample, and so they are random. The usually aim to make $1-\alpha$ large like 95% since this is the probability that the confidence interval covers the true value of the parameter, so α should be small, like 5%.

This is the definition of the two-sided confidence interval. The definition of one-sided intervals are similar. For example, for lower one-sided interval we require that

$$\mathbb{P}(\theta \in [\widehat{\theta}_L, \infty]) = 1 - \alpha.$$

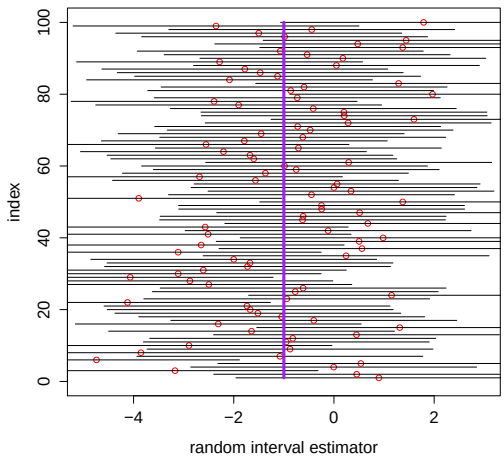


Figure 2.1: Interval estimator for 100 different samples. The confidence interval $(\widehat{\theta} - 1.96\sigma_{\widehat{\theta}},\ \widehat{\theta} + 1.96\sigma_{\widehat{\theta}})$, – centered at the point estimators $\widehat{\theta}$ shown by red circles, – in 95% cases covers the true parameter (shown by the purple line).

Here are examples of confidence intervals: for μ, we could take $(\overline{Y} - 2S, \overline{Y} + 2S)$. Alternatively, we could take $(0.8\overline{Y}, 1.2\overline{Y})$. However, we don't know what is the confidence levels in these examples.

So, here is a question: for a given $(1 - \alpha)$, how can we find the desired confidence interval $(\hat{\theta}_L, \hat{\theta}_U)$? If there are several ways to do it, we would prefer to find an interval, for which the length of the interval is the smallest.

The difficulty is that we do not know true value of the parameter, so for example we cannot use the Chebyshev inequality to build the interval estimate: The interval $(\theta - 2\sigma_{\widehat{\theta}}, \theta - 2\sigma_{\widehat{\theta}})$ is not an interval estimator because we do not know neither θ, no $\sigma_{\widehat{\theta}}$.

We can use $\widehat{\theta}$ instead of θ in this interval and estimate $\sigma_{\widehat{\theta}}$, but why is this OK? This is often fine, as we will see later, but only if the sample size is large.

Here we discuss the pivotal quantity method which in some cases allows us to find confidence intervals exactly.

The pivotal quantity or pivot is a quantity which is a function of both the

sample data and the parameter θ, but whose distribution does not depend on the parameter θ !

Let X denote the data sample. (It is a vector of observations.) Find a function $T(X, \theta)$ (the pivot) so that its distribution does not depend on θ and so it is known.

Use the distribution of T to find a pair of L and U such that

$$\Pr(L \leq T \leq U) = 1 - \alpha$$

Manipulate the inequalities $L \leq T(X, \theta) \leq U$ so that they become

$$L^*(X) \leq \theta \leq U^*(X),$$

so that θ is in the middle!!!

Example 2.1.2 (A sample from a normal distribution). Let $X_1, \ldots, X_n$ be a sample from a normal distribution $\mathcal{N}(\mu, \sigma^2)$, where the parameter σ is known, and we want to find a confidence interval for μ.

Note that $\overline{X} = \frac{1}{n} \sum_{i=1} nX_i$ is a normally distributed random variable with mean μ and variance σ^2/n.

Then the quantity

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

has the normal distribution with mean 0 and variance 1. In particular, its distribution does not depend on μ. (Here T depends not only on the data and parameter μ but also on σ but we assume that σ is known.) By convention, in this important example, the quantity T is denoted Z.

The next task it to look for L and U so that $\Pr(L \leq Z \leq U) = 1 - \alpha$. There are 3 standard ways to do it. One of them to choose L and U so that $\mathbb{P}(Z < L) = \alpha/2$ and $\mathbb{P}(Z > U) = \alpha/2$. By symmetry of the normal distribution $L = -U$ and by convention this U is denoted $z_{\alpha/2}$.

Then we have

$$-z_{\alpha/2} \leq Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}.$$

This inequality can be converted to the desired confidence interval for parameter μ:

$$\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Alternatively, we can look for a "one-sided interval". So, take $L = -\infty$ and look for U such that $\mathbb{P}\{Z > U\} = \alpha$. By definition this U is denoted z_α and it can be found from a table or by using software.

Then, the inequality is

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq z_\alpha$$

and it can be transformed to the *lower confidence bound* on the parameter μ

$$\overline{X} - z_\alpha\frac{\sigma}{\sqrt{n}} \leq \mu.$$

Note the difference from the previous inequality. Here we use the factor z_α before the standard error $\frac{\sigma}{\sqrt{n}}$ while in the previous inequality we used $z_{\alpha/2}$.

Similarly, by using $U = \infty$ and looking for L such that $\mathbb{P}\{Z < L\} = \alpha$, we can derive the *upper confidence bound* on μ:

$$\overline{\mu} < X + z_\alpha\frac{\sigma}{\sqrt{n}}.$$

Example 2.1.3 (Confidence interval for σ^2). Suppose that $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ but now we know μ, and we are interested in deriving a confidence interval for σ^2.

The quantity $T = \sum_{i=1}^{n} \left((X_i - \mu)/\sigma\right)^2$ is known to be distributed according to the χ^2 distribution with n degrees of freedom, $\chi^2(n)$. Therefore, it is a valid pivotal quantity and we can used it to derive a confidence interval for σ^2.

Again, it can be done in three possible ways. One of them is to find the quantities L and U such that $\mathbb{P}(T < L) = \alpha/2$ and $\mathbb{P}(T > U) = \alpha/2$. In this case, the distribution is not symmetric and we need to find 2 really different quantities. The quantity U is $\chi^2_{\alpha/2}(n)$ and the quantity L is $\chi^2_{1-\alpha/2}(n)$. They can be found from a table or by using software.

Then, we get

$$\chi^2_{1-\alpha/2}(n) \leq \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 \leq \chi^2_{\alpha/2}(n),$$

and, by putting σ^2 in the middle,

$$\frac{1}{\chi^2_{\alpha/2}(n)}\sum_{i=1}^{n}(X_i-\mu)^2 \leq \sigma^2 \leq \frac{1}{\chi^2_{1-\alpha/2}(n)}\sum_{i=1}^{n}(X_i-\mu)^2$$

Note that for large n, $\chi^2_{\alpha/2}(n)$ is relatively close to n.

The upper and lower confidence bounds can be found similarly. For example, $(1-\alpha)$ **upper** confidence bound for σ^2 is

$$\sigma^2 \leq \frac{1}{\chi^2_{1-\alpha}(n)}\sum_{i=1}^{n}(X_i-\mu)^2$$

Note that in this inequality we use α instead of $\alpha/2$.

Example 2.1.4. Suppose $X_1, \ldots, X_n$ is a sample from the exponential distribution with mean θ. Find a confidence interval for θ.

Recall that the density of the exponential distribution with mean θ is $(1/\theta)e^{x/\theta}$. Since we are looking for a sample statistic with the distribution that does not depend on θ, it is natural to remove the mean from the distribution of X_i by dividing these random variables by θ. So let $Y_i = X_i/\theta$. Then it is easy to see (and it is easy to check by using one of the methods for calculating the density of Y_i) that Y_i has the exponential distribution with mean 1. Hence it is not depends on the parameter θ.

Since we also want to incorporate the information from all of the observations, we are going to use the pivotal quantity $T = \sum_{i=1}^{n}(X_i/\theta)$. By the properties of the exponential distribution T is distributed as a Gamma random variable with parameters n and 1. If $z^{(n)}_{1-\alpha/2}$ and $z^{(n)}_{\alpha/2}$ are critical values for this distribution, that is, if $\mathbb{P}(T > z^{(n)}_{1-\alpha/2}) = 1-\alpha/2$ and $\mathbb{P}(T > z^{(n)}_{\alpha/2}) = \alpha/2$, then $\mathbb{P}(z^{(n)}_{1-\alpha/2} < T \leq z^{(n)}_{\alpha/2}) = 1-\alpha$ and we get a confidence interval for θ:

$$\frac{\sum_{i=1}^{n} X_i}{z^{(n)}_{\alpha/2}} \leq \theta \leq \frac{\sum_{i=1}^{n} X_i}{z^{(n)}_{1-\alpha/2}}.$$

This interval has confidence level $1-\alpha$. It can be shown that the size of the interval decreases when n grows.

Here is a bit less standard example, which shows that the pivot method sometimes gives poor confidence intervals.

Example 2.1.5. Suppose a sample X_1, ..., X_n of random variables distributed according to the exponential distribution with mean θ.

Suppose we want to build a confidence interval for θ with $\alpha = 10\%$.

The quantity $T = nX_{(1)}/\theta$ is pivotal. Indeed, let us write Y to denote $X_{(1)}$. This is a minimum of n i.i.d exponential random variables and it is easy to check that the density of Y is

$$f_Y(y) = \frac{n}{\theta} e^{-ny/\theta}.$$

That is, Y is exponential with the mean θ/n. Similar to the previous example, it is easy to calculate the density of $T = nY/\theta = Y/(\theta/n)$ and check that it is exponential with mean 1. [1]

Now we look for L and U, so that

$$0.90 = \Pr(L \le T \le U) = \int_L^U e^{-t} dt = e^{-L} - e^{-T}.$$

There are infinitely many combinations of L and U which satisfy this. One possibility is to let

$$\Pr(T > U) = e^{-T} = 0.05 \text{ and } \Pr(T < L) = 1 - e^{-L} = 0.05$$

Solutions are $L = 0.051$ and $U = 2.996$ Now we have

$$0.051 \le T = n\frac{X_{(1)}}{\theta} \le 2.996$$

[1] We are making transformation $T = nY/\theta$ that has the inverse transformation $Y = (\theta/n)T$. Let us use notation $y(t)$ for the function $y = (\theta/n)t$. By using the density transformation method, we calculate the density of T as follows :

$$f_T(t)\, dt = f_Y(y(t))\, dy(t) = f_Y(y(t))\, \frac{dy(t)}{dt} dt$$
$$= \frac{n}{\theta} e^{-n(\theta/n)t/\theta} \times \frac{\theta}{n} dt = e^{-t} dt$$

So, T has the exponential density with parameter 1.

We manipulate these two inequalities to put θ in the middle:

$$\frac{nX_{(1)}}{2.996} \leq \theta \leq \frac{nX_{(1)}}{0.051}$$

Remark: The resulting confidence interval is not very good. Indeed, the length of the interval is $nX_{(1)}(\frac{1}{0.051} - \frac{1}{0.2996})$ and $nX_{(1)}$ is always an exponential random variable with parameter θ, so we cannot expect that the length of this confidence interval goes to 0 as n grows.

2.2 Asymptotic confidence intervals

Finding an exact pivotal quantity for a parameter which results in a short confidence interval is difficult! The good news is that for large sample we can easily find an **approximate pivotal quantity**, meaning that a function of data and parameter has a distribution that *converges* to a fixed probability distribution as the size of the sample, n, increases. Most often, this distribution is the standard normal distribution.

By using this approximate pivotal quantity we can obtain the asymptotic confidence intervals.

Definition 2.2.1. An **asymptotic confidence interval** with confidence coefficient $(1-\alpha)$ is a *random* interval $[\widehat{\theta}_L^{(n)}, \widehat{\theta}_U^{(n)}]$ such that

$$\lim_{n\to\infty} \mathbb{P}(\theta \in [\widehat{\theta}_L^{(n)}, \widehat{\theta}_U^{(n)}]) = 1-\alpha,$$

where n is the size of the data sample.

Often, the Central Limit Theorem (CLT) ensures that when the sample size n is large enough an appropriate estimator is approximately normal random variable.

- for $\theta = \mu$, the estimator $\hat{\theta} = \overline{Y}$ is approximately $\sim N(\mu, \frac{\sigma^2}{n})$;
- for $\theta = p$, the estimator $\hat{\theta} = \hat{p}$ is approximately $\sim N(p, \frac{p(1-p)}{n})$;
- for $\theta = \mu_1 - \mu_2$, the estimator $\hat{\theta} = \overline{Y}_1 - \overline{Y}_2$ is approximately $\sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$;

- for $\theta = p_1 - p_2$, the estimator $\hat{\theta} = \hat{p}_1 - \hat{p}_2$ is approximately $\sim N(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2})$;

In general, many estimators are *asymptotically normal*. We will discuss asymptotic normality a bit later. Intuitively, this means that the distribution of an estimator $\hat{\theta}$ is close to the normal distribution, $N(\theta, \sigma^2_{\hat{\theta}})$, where $\sigma^2_{\hat{\theta}} = Var(\hat{\theta})$.

Then we can write an approximate pivotal quantity:

$$Z = \frac{\widehat{\theta} - \theta}{\widehat{\sigma}_{\widehat{\theta}}},$$

Here $\widehat{\sigma}_{\widehat{\theta}}$ is a consistent estimator of $\sigma_{\hat{\theta}}$. By a theorem which is called the Slutsky theorem, the distribution of Z is close to the standard normal distribution.

Then we can proceed as usual and develop a confidence interval from the pivotal quantity. Since Z is only an approximately pivotal quantity, the resulting confidence interval will be only an asymptotic confidence interval, that is, the probability that the interval covers θ equals $1 - \alpha$ only if n is large.

Example 2.2.2 (Two-sided asymptotic confidence interval for a parameter θ). By using our results for the normally distributer variables $X_1, \ldots, X_n$, we write the (appoximate) two-sided interval for θ, based on the point estimator $\widehat{\theta}$ which is assumed to be approximately distributed as $N(\theta, \sigma^2_{\hat{\theta}})$.

The two-sided confidence interval for θ with confidence coefficient $1 - \alpha$ is

$$\left[\hat{\theta} - z_{\alpha/2}\sigma_{\widehat{\theta}}, \quad \hat{\theta} + z_{\alpha/2}\sigma_{\widehat{\theta}}\right]$$

Example 2.2.3 (Upper and lower confidence bounds). The one-sided large sample confidence intervals are as follows:

- The upper bound confidence interval with confidence coefficient $1 - \alpha$ is

$$(-\infty, \quad \hat{\theta} + z_{\alpha}\sigma_{\widehat{\theta}} \quad]$$

- The lower bound confidence interval with confidence coefficient $1 - \alpha$ is

$$[\hat{\theta} - z_{\alpha}\sigma_{\widehat{\theta}} \quad , +\infty)$$

Example 2.2.4. The shopping times of $n = 64$ randomly selected customers at a local supermarket were recorded. The average and variance of the 64 shopping times were 33 minutes and 256 minutes, respectively. Estimate μ, the true average shopping time per customer, with a confidence coefficient of $1 - \alpha = .90$.

Solution. We have $\widehat{\mu} = 33$, and $\widehat{\sigma}_{\widehat{\theta}} = \sqrt{256/64} = 2$. We can get $z_{\alpha/2} = z_{0.05} = qnorm(0.05, lower.tail = F) = 1.644854$, so the CI is

$$(33 - 1.645 \times 2, 33 + 1.645 \times 2) = (29.71, 36.29)$$

We got $z_{0.05}$ by using R function "qnorm". We could also get it using an appropriate table.

Example 2.2.5. Two brands of refrigerators, denoted A and B, are each guaranteed for 1 year. In a random sample of 50 refrigerators of brand A, 12 were observed to fail before the guarantee period ended. An independent random sample of 60 brand B refrigerators also revealed 12 failures during the guarantee period. Estimate the true difference $(p_1 - p_2)$ between proportions of failures during the guarantee period, with the confidence coefficient approximately .98.

Solution.

- $n_A = 50$ and $Y_A = 12$, hence $\widehat{p}_1$=12/50 = 0.24
- $n_B = 60$ and $Y_B = 12$, hence $\widehat{p}_2$=12/60 = 0.2
- Use $\widehat{\theta} = \widehat{p}_1 - \widehat{p}_2 = 0.04$ as the (point) estimator.
- $\widehat{\theta}$ is approximately normal and we have

  $$\widehat{\theta} \pm z_{\alpha/2}\sigma_{\widehat{\theta}}$$

  as the $100(1 - \alpha)\%$ confidence interval.
- Note that

  $$\begin{aligned}\sigma_{\widehat{\theta}} &= \sqrt{\mathbb{V}\text{ar}\widehat{\theta}} \\ &= \sqrt{p_1(1-p_1)/n_A + p_2(1-p_2)/n_B},\end{aligned}$$

  where p_1 and p_2 are unknown but can be approximated by $\widehat{p}_1$ and $\widehat{p}_2$.

So,

$$\widehat{\sigma}_{\widehat{\theta}} = \sqrt{0.24(1-0.24)/50 + 0.2(1-0.2)/60} = 0.0795.$$

We also have $z_{\alpha/2} = z_{0.01} = qnorm(0.01, lower.tail = F)2.326348$, and therefore the confidence interval is

$$(0.04 - 2.326348 \times 0.0795, 0.04 + 2.326348 \times 0.0795) = (-0.1449, 0.2249)$$

We used here the R function "qnorm" to find the value of $z_{0.01}$. The lower.tail option is set to F (FALSE) because we want to find such $z_{0.01}$ such that the *upper tail* of the standard normal distribution is 0.01, that is, we want $\mathbb{P}(Z > z_{0.01}) = 0.01$. Alternatively, we could get the value of $z_{0.01}$ from a table.

For the exam, you are supposed to know how to calculate this confidence interval. However, note that these calculations are already implemented in *R*, although R uses the language of hypothesis testing here, which we will learn later.

In particular, for this example, the confidence interval can be calculated as follows.

```
prop.test(c(12, 12), c(50,60), conf.level = 0.98, correct = F)
```

The first argument is the vector of the number of successes (or failures in our example), and the second argument is the vector of the sample sizes.

```
        2-sample test for equality of proportions without
        continuity correction


data:  c(12, 12) out of c(50, 60)
X-squared = 0.25581, df = 1, p-value = 0.613
alternative hypothesis: two.sided
98 percent confidence interval:
 -0.1448629  0.2248629
sample estimates:
prop 1 prop 2
  0.24   0.20
```

Example 2.2.6. A study was done on 41 first-year medical students to see if their anxiety levels changed during the first semester. One measure used was the level of serum cortisol, which is associated with stress. For each of the 41 students the level was compared during finals at the end of the semester against the level in the first week of classes. The average difference was 2.08 with a standard deviation of 7.88. Find a 95% lower confidence bound for the population mean difference μ. Does the bound suggest that the mean population stress change is necessarily positive?

Example 2.2.7. A random sample of 539 households from a mid-western city was selected, and it was determined that 133 of these households owned at least one firearm ("The Social Determinants of Gun Ownership: Self-Protection in an Urban Environment," Criminology, 1997: 629–640). Using a 95% confidence level, calculate a lower confidence bound for the proportion of all households in this city that own at least one firearm.

2.3 How to determine the sample size

The sample size dilemma

- We need to collect samples to make inference about the population parameter. Question: how large n should be? 30? 40? 50?
- On the one hand, the more data you have, the more accurate is your estimator $\widehat{\theta}$ for θ
- On the other hand, collecting samples is NOT free: it costs money, time, personnel..

Conclusion: We want minimal sample size which allows us to achieve given precision with a given level of confidence.

The formulas for confidence intervals provide an easy way to find the required size of the sample.

$$\text{Precision} \ + \ \text{Confidence level} \ \rightarrow \ \text{Required minimal sample size.}$$

Rather than give a bunch of formulas we illustrate the method in examples.

Example 2.3.1. The reaction of an individual to a stimulus in a psychological experiment may take one of two forms, A or B. If an experimenter wishes to estimate the probability p that a person will react in manner A, how many people must be included in the experiment? Assume that the experimenter will be satisfied if the error of estimation is less than .04 with probability equal to .90. Assume also that he expects p to lie somewhere in the neighborhood of .6.

- This is an estimating p in a binomial distribution problem. n is to be found.
- Although n is yet to be found, let's assume that it is large enough, in which case $\widehat{\theta} = \widehat{p}$ is approximately normal, and then $Z \equiv \frac{\widehat{p}-p}{\sqrt{p(1-p)/n}} \approx \frac{\widehat{p}-p}{\sqrt{\tilde{p}(1-\tilde{p})/n}}$ is approximately standard normal and hence, $\mathbb{P}(-z_{0.05} \leq \frac{\widehat{p}-p}{\sqrt{\tilde{p}(1-\tilde{p})/n}} \leq z_{0.05}) = 0.9$. This is to say that with probability 0.9,
$$|\widehat{p} - p| < z_{0.05}\sqrt{p(1-p)/n}$$
- If we want this error to be smaller than 0.04, only need to have $z_{0.05}\sqrt{\tilde{p}(1-\tilde{p})/n} \leq 0.04$. Solve n and we have $n \geq 406$.
- The prior information $\tilde{p} = 0.6$ is used to approximate the standard error

Example 2.3.2. Telephone pollsters often interview between 1000 and 1500 individuals regarding their opinions on various issues. A survey question asks if a person believes that the performance of their athletics teams has a positive impact on the perceived prestige of the institutions. The goal of the survey is to see if there is a difference between the opinions of men and women on this issue. Suppose that you design the survey and wish to estimate the difference in a pair of proportions, correct to within .02, with probability .9. How many interviewees should be included in each sample?

1. What is the standard error of $\widehat{\theta}$? $\sigma_{\widehat{\theta}} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_1}}$
2. $\widehat{\theta} = \widehat{p}_1 - \widehat{p}_2$ is approximately normal with mean $\theta = p_1 - p_2$ and variance $\sigma^2_{\widehat{\theta}}$;
3. Hence, with probability 0.9, $|\widehat{\theta} - \theta| < z_{0.05}\sigma_{\widehat{\theta}}$
4. Take $n_1 = n_2 = n$. For p_1 and p_2, since we have no prior information, we just replace them by 0.5, as the most conservative guess.
5. If we want this error to be smaller than 0.02, only need to solve
$$z_{0.05}\sqrt{\frac{1/4}{n} + \frac{1/4}{n}} < 0.02$$
6. $$n \geq \frac{1}{2}\left(\frac{z_{0.05}}{0.02}\right)^2 = \frac{1}{2}\left(\frac{1.645}{0.02}\right)^2 = 3382.5$$

So we should take $n = 3383$.

Example 2.3.3. A state wildlife service wants to estimate the mean number of days that each licensed hunter actually hunts during a given season, with a bound on the error of estimation equal to 2 hunting days. If data collected in earlier surveys have shown σ to be approximately equal to 10, how many hunters must be included in the survey?

- The client is not sophisticated and does not formulate explicitly what is the level of confidence required. In this case, it is typical to set the confidence level at 95% and use the 2-standard-error bound. If we want the error of estimation to be less than 2, then
$$2 > 2\sigma_{\widehat{\theta}} = 2\frac{\sigma}{\sqrt{n}} = 2\frac{10}{\sqrt{n}} \Rightarrow n > 100.$$

2.4 Small-sample confidence intervals

2.4.1 Small sample CIs for for μ and $\mu_1 - \mu_2$

Suppose, the parameter of interest is the population mean μ and we have a sample $X_1, \ldots, X_n$. When the sample size is large, the Central Limit Theorem ensures that $\overline{X}$ is approximately normal with distribution $N(\mu, \frac{\sigma^2}{n})$.

In addition, the parameter σ^2 can be consistently estimated by the sample variance. Thus, the quantity

$$Z = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

is approximately pivotal with the asymptotic distribution $N(0,1)$. Based on Z, we can find

- 2-sided confidence interval for μ: $[\overline{X} - z_{\alpha/2}\frac{S}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}]$
- 1-sided lower confidence interval for μ: $[\overline{X} - z_{\alpha}\frac{S}{\sqrt{n}}, \infty]$
- 1-sided upper confidence interval for μ: $[-\infty, \overline{X} + z_{\alpha}\frac{S}{\sqrt{n}}]$;

Now suppose that the sample size n is small, say less than 30. Then the quantity Z may have a distribution which is very different from the standard normal distribution. **Using the normal distribution in the case of small samples leads to erroneous intervals!**

What can we do? In general, there is no universal answer. The answer depends on the distribution of the data points X_i.

If the data happen to be normally distributed and σ^2 is known, then

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \tag{2.1}$$

is pivotal and normal. However, in most cases this fact cannot be used to construct a confidence interval for μ since σ^2 is not known.

We can try to use the sample variance S^2 instead of σ^2, however then

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \tag{2.2}$$

is pivotal but its distribution is not normal. The reason is that S^2 is a random quantity and dividing the normal random variable $\overline{X} - \mu$ by a random quantity instead of a deterministic coefficient breaks the normality of the quotient.

What is the distribution of T? Let us recall first some facts from the probability theory.

Definition 2.4.1. If $X_1, \ldots, X_n$ are i.i.d and distribute as standard normal r.v.s ($\sim N(0,1)$), then the random variable $Y \equiv X_1^2 + \cdots + X_n^2$ has the χ^2 *distribution* with n degrees of freedom.

Definition 2.4.2. Suppose random variables $Z \sim N(0,1)$ and $Y \sim \chi^2(n)$ are *independent*. Then the random variable

$$T \equiv \frac{Z}{\sqrt{\frac{Y}{n}}} \tag{2.3}$$

is distributed according to the *t-distribution* with n degrees of freedom, denoted as

$$T \sim t(n)$$

.

This distribution is also often called Student's t-distribution. It was discovered by William Gosset who worked for Guinness brewery and wrote his papers under the pen name Student.

For large n, – say for $n > 30$, the $t(n)$- distribution is very close to the standard normal distribution $N(0,1)$. And in general, the t distribution has many properties that are similar to the normal distribution. For example, its PDF is symmetric with respect to 0. However, t distribution has heavier tails: When n is small, a r.v. with the $t(n)$-distribution takes large values with much larger probability than a standard normal random variable.

Now, note that T in the formula (2.2) has quite similar expression as in (2.3). However, is it true that $\overline{X}$ and S^2 (2.2) are independent? And is S^2 has χ^2-distribution with n degrees of freedom. The answer is almost positive by the following remarkable theorem which we will state without proof.

Theorem 2.4.3 (Joint distribution of sample mean and sample variance)**.** *Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. Then the sample variance S^2 is independent of the sample mean $\overline{X}$ and*

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

has the χ^2 distribution with $n-1$ degrees of freedom.

Note that the random variable has χ^2 distribution with one less degree of freedom than if we would add together n independent standard normal random variables. The reason for this reduction is that the random variables $X_i - \overline{Y}$ are not independent. Intuitively they can be expressed in terms of $n-1$ independent normal random variables, hence the reduction in the degree of freedom. The most surprising in this theorem is the independence of $\overline{X}$ and S^2.

By using this theorem, we can show that

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1),$$

that is T has a t-distribution with degrees freedom $df = n - 1$.

Based on T,

- 2-sided confidence interval for μ: $[\overline{X} - t_{\alpha/2}\frac{S}{\sqrt{n}}, \overline{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}]$
- 1-sided lower confidence interval for μ: $[\overline{X} - t_{\alpha}\frac{S}{\sqrt{n}}, \infty]$
- 1-sided upper confidence interval for μ: $[-\infty, \overline{X} + t_{\alpha}\frac{S}{\sqrt{n}}]$;

t_α can be found using statistical software or the t-table (Table 5, look for subscript α with $df = n - 1$).

Note that the statistic in both large sample and small sample cases is the same:

$$T \equiv \frac{\overline{X} - \mu_Y}{S/\sqrt{n}}$$

The only difference is that in the case of a large sample, it is distributed as normal random variable, and in the case of a small sample it is distributed as a t-random variable.

Remark 1: the small sample confidence intervals based on t distribution are longer than asymptotic confidence intervals based on the standard normal distribution.

Remark 2: the small sample confidence intervals based on t distribution are valid only if the data are normally distributed.

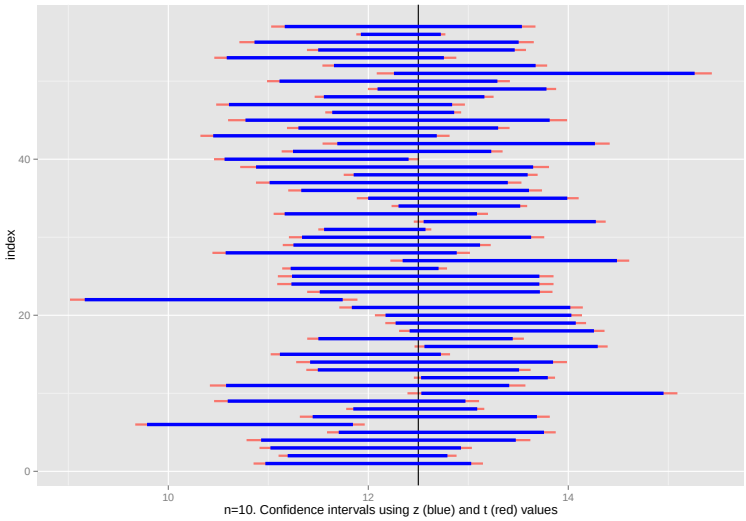


Figure 2.2: Comparison of z and t confidence intervals for $n = 10$.

Figure 2.2 compares the z and t confidence intervals in a sample with n observations.

- Red: using the $t_{\alpha/2}$ value – the correct one
- Blue: using the $z_{\alpha/2}$ value – the incorrect one
  - Fail to take the uncertainty of S^2 into consideration
  - Fail to deliver the promised coverage probability
  - Shorter than the red one (for small α, $z_{\alpha/2} < t_{\alpha/2}$)

Example 2.4.4. The carapace lengths of ten lobsters examined in a study of the infestation of the *Thenus orientalis* lobster by two types of barnacles, *Octolasmis tridens* and *O. lowei*, are given in the following table. Find a 95% confidence interval for the mean carapace length (in millimeters, mm) of *T.orientalis* lobsters caught in the seas in the vicinity of Singapore.

Lobster Field Number	*A061*	*A062*	*A066*	*A070*	*A067*	*A069*	*A064*	*A068*	*A065*	*A063*
Carapace Length (mm)	78	66	65	63	60	60	58	56	52	50

This is a small sample estimation problem for μ. Below are calculations done in R.

```
> x = c(78,66,65,63,60,60,58,56,52,50)
> x
 [1] 78 66 65 63 60 60 58 56 52 50
> mean(x) # sample mean
[1] 60.8
> sum((x-60.8)^2)/(10-1)  # sample variance
[1] 63.51111
> sqrt(sum((x-60.8)^2)/(10-1))  # sample standard deviation
[1] 7.969386
> qt(0.975,9) # 0.025 percentage point
              # for t distribution with df=9
[1] 2.262157
> qt(0.975,9)*sqrt(sum((x-60.8)^2)/(10-1))/sqrt(10)
[1] 5.700955
```

Answer: 60.8 ± 5.700955. Note that we have divided by $\sqrt{9}$ when we estimated S and then again by $\sqrt{10}$ when we estimated $\sigma_{\overline{X}}$. Do not forget the second division.

Alternatively, one can use the "sd" function to calculate the sample standard deviation:

```
>mean(x)-qt(0.975,9)*sd(x)/sqrt(10)
>mean(x)+qt(0.975,9)*sd(x)/sqrt(10)
```

produce the required interval.

Finally, one can also get the confidence interval in a simpler ways by using function "t-test":

```
t.test(x)
```

produces the following output:

```
        One Sample t-test

data:  x
t = 24.126, df = 9, p-value = 1.727e-09
alternative hypothesis: true mean is not equal to 0
```

```
95 percent confidence interval:
 55.09904 66.50096
sample estimates:
mean of x
    60.8
```

Exercise 2.4.5. The reaction time (RT) to a stimulus is the interval of time commencing with stimulus presentation and ending with the first discernible movement of a certain type. The article "Relationship of Reaction Time and Movement Time in a Gross Motor Skill" (Percept. Motor Skills, 1973: 453–454) reports that the sample average RT for 16 experienced swimmers to a pistol start was .214 s and the sample standard deviation was .036 s.

Making any necessary assumptions, derive a 90% CI for true average RT for all experienced swimmers.

Two sample t-test

We have samples $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots Y_{n_2}$, with observations distributed according to $N(\mu_1, \sigma_1)$ and, $N(\mu_1, \sigma_1)$ respectively. We assume that n_1 and n_2 are small and want to find C.I. for $\mu_1 - \mu_2$.

Here we consider only the most simple case when it is assumed that $\sigma_1 = \sigma_2 = \sigma$. Then we can define the pooled-sample estimator for the common variance σ^2,

$$\begin{aligned} S_p^2 &\equiv \frac{\sum_{i=1}^{n_1}(X - \overline{X})^2 + \sum_{i=1}^{n_2}(Y_i - \overline{Y})^2}{n_1 + n_2 - 2} \\ &= \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \end{aligned}$$

In this case,

$$T = \frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{\widehat{\sigma}_{\overline{Y}_1 - \overline{Y}_2}} = \frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2),$$

that is, T has the t-distribution with $df = n_1 + n_2 - 2$;

So in this simple case, we have the following confidence interval for $\mu_1 - \mu_2$

$$\left(\overline{X} - \overline{Y} - t_{\alpha/2}^{(n_1+n_2-2)} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \overline{X} - \overline{Y} + t_{\alpha/2}^{(n_1+n_2-2)} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right),$$

Similarly, the lower bound confidence interval for $\mu_1 - \mu_2$ is

$$\left(\overline{X} - \overline{Y} - t_\alpha S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \infty\right)$$

and the upper bound confidence interval for $\mu_1 - \mu_2$ is

$$\left(-\infty, \overline{X} - \overline{Y} + t_\alpha S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$$

In more general case, the formulas are more complicated, and one has rely on software.

Example 2.4.6. To reach maximum efficiency in performing an assembly operation in a manufacturing plant, new employees require approximately a 1-month training period. A new method of training was suggested, and a test was conducted to compare the new method with the standard procedure. Two groups of nine new employees each were trained for a period of 3 weeks, one group using the new method and the other following the standard training procedure. The length of time (in minutes) required for each employee to assemble the device was recorded at the end of the 3-week period. The resulting measurements are as shown in Table 8.3 (see the book). Estimate the true mean difference $(\mu_1 - \mu_2)$ with confidence coefficient .95. Assume that the assembly times are approximately normally distributed, that the variances of the assembly times are approximately equal for the two methods, and that the samples are independent.

2.4.2 Small sample CIs for population variance σ^2

Population variance σ^2 quantifies the amount of variability in the population. We have already shown that if we observe the data sample $(X_1, \ldots, X_n)$, then σ^2 can be estimated by an unbiased point estimator

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

How do we get a confidence interval for σ^2?

If the sample is large then S^2 is approximately normally distributed. It is not difficult to derive a formula for $\mathbb{V}\mathrm{ar}(S^2)$ and develop an estimator for this

variance. Then the standard method for large sample confidence intervals works. In practice, however, we are usually interested in confidence intervals for σ^2 when we have a small data sample.

So assume that the data sample is small. In this case we must restrict ourself to the situation when the data is normally distributed.

Assume that all sample data points $X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$.

The pivotal quantity (see Theorem 2.4.3) is

$$T = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

We need to find L and U so that

$$\mathbb{P}(L \leq \frac{(n-1)S^2}{\sigma^2} \leq U) = 1 - \alpha$$

A usual choice is $L = \chi^2_{1-(\alpha/2)}$ and $U = \chi^2_{\alpha/2}$, both corresponding to $(n-1)$ d.f.

Hence,

$$\mathbb{P}\left(\chi^2_{1-(\alpha/2)}(n-1) \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{\alpha/2}(n-1)\right) = 1 - \alpha$$

$$\Leftrightarrow \mathbb{P}\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}\right) = 1 - \alpha$$

- Suppose we want a one-sided bound for σ^2, say lower bound. We will want to make use of the pivotal quantity:

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

- Note that if $b = \chi^2_{\alpha}$ for $(n-1)$ d.f., then

$$\mathbb{P}\Big(\frac{(n-1)S^2}{\sigma^2} \leq b\Big) = 1 - \alpha$$

- Then we have, with $100(1-\alpha)\%$ probability,

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

Hence $\frac{(n-1)S^2}{\chi^2_{\alpha,n-1}}$ is a $(1-\alpha)$ confidence lower bound for σ^2.

Similarly, $\frac{(n-1)S^2}{\chi^2_{1-\alpha;n-1}}$ is a $(1-\alpha)$ confidence upper bound for σ^2.

What if we want to build a confidence interval for the standard deviation $\sigma = \sqrt{\sigma^2}$, instead of σ^2? This is simple:

- Since we know that
$$\mathbb{P}\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right) = 1-\alpha$$
- This is equivalent to
$$\mathbb{P}\left(\sqrt{\frac{(n-1)S^2}{\chi^2_{\alpha/2}}} \leq \sigma \leq \sqrt{\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}}\right) = 1-\alpha$$
- Therefore we can easily obtain a C.I. for σ.

Example 2.4.7. Suppose that you wished to describe the variability of the carapace lengths of this population of lobsters. Find a 90% confidence interval for the population variance σ^2.

Lobster Field Number	*A061*	*A062*	*A066*	*A070*	*A067*	*A069*	*A064*	*A068*	*A065*	*A063*
Carapace Length (mm)	78	66	65	63	60	60	58	56	52	50

```
> x = c(78,66,65,63,60,60,58,56,52,50)
> x
 [1] 78 66 65 63 60 60 58 56 52 50
> sum(( x-mean(x) )^2) # the numerator of sample variance and the CI LB and UB
[1] 571.6
> # and then we calculate the denominators of the CI LB and UB
> qchisq(0.05,9)
[1] 3.325113
> qchisq(0.95,9)
[1] 16.91898
```

The answer is $(571.6/16.91898, 571.6/3.325113)$. Note that $\chi^2_{0.95,9} = 3.325113 =$ `qchisq(0.05,9)` and $\chi^2_{0.05,9} = 16.91898 =$ `qchisq(0.95,9)`.

Both can also be obtained from Table 6.

Example 2.4.8. An optical firm purchases glass to be ground into lenses. As it is important that the various pieces of glass have nearly the same index of refraction, the firm is interested in controlling the variability. A simple random sample of size $n = 20$ measurements yields $S^2 = (1.2)10^{-4}$. From previous experience, it is known that the normal distribution is a reasonable model for the population of these measurements. Find a 95% CI for σ.

Chapter 3

Advanced properties of point estimators

3.1 More about consistency of estimators

We have seen in Chapter 1, that it is often difficult or even impossible to find an *unbiased* estimator of a parameter θ. What about consistent estimators? Can we find consistent estimators? We will see later that the answer is positive and there are some useful methods to find a consistent estimator. What we are going to do in this section is to study how one can prove that an estimator is consistent.

Recall that the consistency of an estimator $\widehat{\theta}$ means that for every $\theta \in \Theta$ the estimator converges in probability to the true value of the parameter. We have shown that $\widehat{\theta}$ is consistent if its Mean Square Error converges to zero for every $\theta \in \Theta$ as $n \to \infty$. In fact, due to the MSE decomposition theorem, it is enough to show that the bias and the variance of the estimator converge to zero.

In practice, however, it is often difficult to calculate the variance of the estimator. So, it is a good news that in some cases consistency can be established without actually calculating the variance.

Since consistency is all about convergence in probability, here are some properties of this mode of convergence of random variables.

Theorem 3.1.1. *Suppose that* $\widehat{\theta}_n \xrightarrow{p} \theta$ *and* $\widehat{\theta}'_n \xrightarrow{p} \theta'$*, then:*

1. $\widehat{\theta}_n + \widehat{\theta}'_n \xrightarrow{p} \theta + \theta'$;
2. $\widehat{\theta}_n \times \widehat{\theta}'_n \xrightarrow{p} \theta \times \theta'$;
3. $\widehat{\theta}_n / \widehat{\theta}'_n \xrightarrow{p} \theta/\theta'$ *provided that* $\theta' \neq 0$;
4. *For any continuous function* $g(u)$, $g(\widehat{\theta}_n) \xrightarrow{p} g(\theta)$;
5. *For any continuous function* $g(u, v)$, $g(\widehat{\theta}_n, \widehat{\theta}'_n) \xrightarrow{p} g(\theta, \theta')$;
6. *For a sequence of numbers* $\{a_n, n = 1, \dots\}$, $a_n \to a$ *(in the calculus sense) implies that* $a_n \xrightarrow{\mathbb{P}} a$ *(a_n's are viewed as special random variables).*

This result is called the continuous mapping theorem for the convergence in probability.

We omit the proof.

Example 3.1.2 (S^2 is a consistent estimator of σ^2.). By definition

$$S_n^2 = \frac{1}{n-1}\left(\sum_{i=1}^n Y_i^2 - n\overline{Y}^2\right) = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^n Y_i^2 - \overline{Y}^2\right) \tag{3.1}$$

We want to show that

$$S_n^2 \xrightarrow{p} \sigma^2$$

- By the law of large numbers, we have:

$$\frac{1}{n}\sum_{i=1}^n Y_i^2 \xrightarrow{p} \mathbb{E}(Y_i^2) = \sigma^2 + \mu^2, \tag{3.2}$$

- Also, by LLN,

$$\overline{Y} \xrightarrow{p} \mathbb{E}(Y_1) = \mu \tag{3.3}$$

(Note that $\overline{Y}$ is actually a sequence of random variables that depends on the sample size n. This dependence is suppressed in the notation but we should remember about it to understand what it means that $\overline{Y} \xrightarrow{\mathbb{P}} \mathbb{E}(Y_i)$.)

- By an application of the continuous mapping theorem with $g(u) = u^2$, we have (3):
$$(\overline{Y})^2 \xrightarrow{p} \mu^2$$

- Consider $g(u, v) = u - v$ and apply the continuous mapping theorem again,
$$\frac{1}{n}\sum_{i=1}^{n} Y_i^2 - (\overline{Y}_n)^2 \xrightarrow{p} \sigma^2 + \mu^2 - \mu^2 = \sigma^2$$

Recall that
$$S_n^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \overline{Y}_n{}^2\right),$$
and we have just shown that
$$\frac{1}{n}\sum_{i=1}^{n} Y_i^2 - (\overline{Y}_n)^2 \xrightarrow{p} \sigma^2$$
It remains to notice that
$$\frac{n}{n-1} \to 1 \text{ implies } \frac{n}{n-1} \xrightarrow{p} 1,$$
and another application of the continuous mapping theorem (with function $g(u, v) = uv$, and variables $u = n/(n-1)$ and $v = \left(\frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \overline{Y}_n^2\right)$ shows that
$$S_n^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \overline{Y}_n^2\right) \xrightarrow{p} 1 \times \sigma^2 = \sigma^2.$$

Example 3.1.3 (Another estimator of σ^2). Another estimator of σ can be defined as
$$S_0^2 \equiv \frac{1}{n}\left(\sum_{i=1}^{n} Y_i^2 - n\overline{Y}_n^2\right)$$
The denominator of S^2 is $(n-1)$ while that of his brother S_0^2 is n. Estimator S^2 is an unbiased estimator of σ^2) and estimator S_0^2 is biased. Is S_0^2 a consistent estimator of σ^2? Yes! (By the continuous mapping theorem.)

Example 3.1.4. Is $S := \sqrt{S^2}$ a continuous estimator of σ?

Yes! By the continuous mapping theorem, if plim $S_n^2 = \sigma^2$, then plim $\sqrt{S_n^2} = \sqrt{\sigma^2} = \sigma$.

Is consistency really important?

Yes. If an estimator is not consistent, then it will not produce the correct estimation even if we are given the luxury of getting unlimited amount of data for free. It's a shame if one cannot get the correct answer in this situation. An inconsistent estimator is a waste of time.

Does consistency guarantee good performance?

Not necessarily. We still live in a finite sample world. Something that is ultimately good for very large sample, may not be good enough for a realistic sample size.

3.2 Asymptotic normality

Definition 3.2.1. An estimator $\widehat{\theta}_n$ is called *asymptotically normal* if $(\widehat{\theta}_n - \theta)/\sqrt{\mathbb{V}\mathrm{ar}(\widehat{\theta}_n)}$ converges in distribution to the standard normal distribution.

Typically, $\mathbb{V}\mathrm{ar}(\widehat{\theta}_n) \sim \sigma^2/n$, and the constant σ^2 is called the *asymptotic variance* of the estimator. Intuitively, as n grows, the error of the estimator becomes becomes more and more like a normal random variable with variance σ^2/n. In particular we can use the techniques that we learned in the previous chapter in order to build the asymptotic confidence intervals.

In order to prove the asymptotic normality, we usually use the CLT (Central Limit Theorem).

Example 3.2.2. Let $X_1, \ldots, X_n$ be a sample from a distribution with mean $\mathbb{E}(X_i) = \mu$ and variance $\mathbb{V}\mathrm{ar}(X_i) = \sigma^2$. Then $\overline{X}$ is asymptotically normal with asymptotic variance σ^2.

This is exactly the statement of the central limit theorem, namely: If $X_1, \ldots, X_n$ are i.i.d. with mean μ and finite variance σ^2, then

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

in distribution, where Z is the standard normal r.v.

Many other estimators are also asymptotically normal, but it might be not so easy to find their asymptotic variance.

The following example is meant to illustrate that sometimes there are estimators that have smaller asymptotic variance than sample mean.

Example 3.2.3. Let $X_1, \dots, X_n$ be a sample from the Laplace distribution shifted by θ, that is from the distribution with density

$$p_\theta(x) = \frac{1}{2} e^{-|x-\theta|}.$$

By symmetry, it is clear that $\mathbb{E}(X_i) = \theta$, so we can estimate by using either the sample mean or the sample median. Let us compare the asymptotic variance of the estimators $\overline{X}$ and $\widehat{\theta}_{med}$.

It turns out that it is possible to prove that $\widehat{\theta}_{med}$ is asymptotically normal estimator of θ with asymptotic variance $1/(4p(0)^2) = 1/(4 \times (1/2)^2) = 1$.

On the other hand, for $\overline{X}$, the asymptotic variance equals to the variance of the Poisson distribution, which can be computed as 2. [1] It follows that in this example the sample median has smaller asymptotic variance than the sample mean.

Exercise 3.2.4. Show that if $X_1, \dots, X_n$ is a sample from the standard normal distribution, then the sample mean has smaller asymptotic variance than the sample median.

These examples show that the answer to the question of which estimator is better often depends on the distribution from which we draw the sample.

Now let us finish this section by proving the result that we used in the section about asymptotic confidence intervals.

Example 3.2.5. Let $X_1, \dots, X_n$ be a sample from a distribution with mean $\mathbb{E}(X_i) = \mu$ and variance $\mathbb{V}\mathrm{ar}(X_i) = \sigma^2$. Then, the statistics

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

[1]

$$\sigma^2_{\overline{X}} = \frac{1}{2} \int_{-\infty}^{\infty} x^2 e^{-|x|}\, dx = \int_0^{\infty} x^2 e^{-x}\, dx = \Gamma(3) = 2! = 2.$$

converges to the standard normal distribution.

This is a direct consequence of the following result which we give without proof.

Theorem 3.2.6 (Slutsky's theorem)**.** *If X_n converges* in distribution *to a variable X, and Y_n converges* in probability *to a constant c, then*

- $X_n + Y_n \to X + c$ *in distribution;*
- $X_n Y_n \to cX$ *in distribution;*
- $X_n/Y_n \to X/c$ *in distribution, provided that $c \neq 0$.*

3.3 Risk functions and comparison of point estimators

In Section 1.2, we defined the Mean Squared Error of a point estimator:

$$MSE_{\widehat{\theta}}(\theta) := \mathbb{E}(\widehat{\theta} - \theta)^2.$$

We wrote it here as a function of θ to emphasize that the MSE depends on the true value of the parameter θ.

This is a particular case of the *risk function* of an estimator. More generally, the risk function is

$$R_{\widehat{\theta}}(\theta) := \mathbb{E}\big[u(\widehat{\theta} - \theta)\big],$$

where $u(x)$ is some non-negative function, which might depend on a particular application. The function $u(x)$ is called the *loss* function. So, intuitively the risk function is the expected loss from a mistake made while predicting the parameter θ. In the case of the MSE, the loss function is simply a quadratic function: $u(x) = x^2$.

In the following we will use the MSE, but everything also holds for other risk functions.

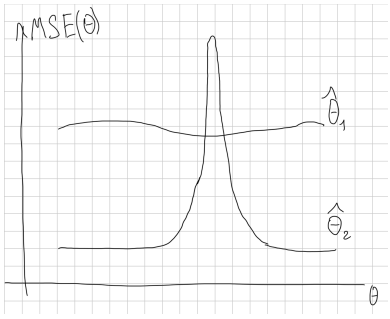


Figure 3.1: MSE (risk functions) for two different estimators, $\widehat{\theta}_1$ and $\widehat{\theta}_2$

The important thing is that MSE and, more generally, any risk function depends on θ, which we do not know. Ideally, an estimator $\widehat{\theta}_1$ is better than another estimator $\widehat{\theta}_2$ if its MSE is smaller for every $\theta \in \Theta$. However, it might happen that MSE of $\widehat{\theta}_1$ is smaller than MSE of $\widehat{\theta}_2$ for one value of parameter θ and larger for another. See the picture.

In general there are two approaches how to deal with this situation. In the first approach we simply compute the average MSE of the estimators over the set of possible parameters and compare these averages. This is called the *Bayesian approach* since it is popular in the branch of statistical theory called the Bayesian statistics.

In the other approach one finds the values of θ which give the largest MSE for every of the estimators. The better estimator will have the smaller of these MSE. This is called the *minimax approach*.

Exercise 3.3.1. According to the minimax criterion, which of the estimators is better for the situation pictured in Figure 3.1? Which one is better according to the Bayesian criterion?

Example 3.3.2. Let $X_1, \ldots, X_n$ be sampled from the exponential distribution with mean θ. Consider estimators of θ that have the form $\widehat{\theta} = \mu_n(X_1 + \ldots + X_n)$. Calculate the MSE of these estimators. Which of them best according to the Bayesian criterion? according to the minimax approach?

Let us calculate the MSE. In the calculation, we will use the fact that if

$X_i \sim Exp(\theta)$ then $\mathbb{E}(X_i) = \theta$, $\mathbb{V}\text{ar}(X_i) = \theta_2$ and therefore $\mathbb{E}(X_i^2) = 2\theta^2$.

$$\begin{aligned} MSE(\theta) &:= \mathbb{E}\Big[\mu_n(X_1 + \ldots + X_n) - \theta\Big]^2 \\ &= \mathbb{E}\Big[\mu_n^2(\sum_{i=1}^n X_i^2 + \sum_{i\neq j} X_i X_j) - 2\mu_n\theta\sum_{i=1}^n X_i + \theta^2\Big] \\ &= \mu_n^2(2n\theta^2 + n(n-1)\theta^2) - 2\mu_n(n\theta^2) + \theta^2) \\ &= \theta^2\Big(n(n+1)\mu_n^2 - 2n\mu_n + 1\Big). \end{aligned}$$

What is important here is that the MSE is θ^2 multiplied by a constant that depends on μ_n. So, it does not matter if we use the Bayesian or the minimax criteria. According to both, the best estimator is the estimator that minimizes the constant

$$n(n+1)\mu_n^2 - 2n\mu_n + 1.$$

It is easy to check that the minimum is reached for

$$\mu_n^* = \frac{2n}{2n(n+1)} = \frac{1}{n+1}.$$

So the best estimator is

$$\widehat{\theta}\frac{1}{n+1}\sum_{i=1}^n X_i = \frac{n}{n+1}\overline{X}.$$

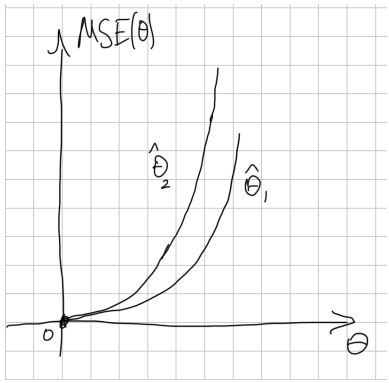


Figure 3.2: MSE (risk functions) for two different estimators, $\widehat{\theta}_1$ and $\widehat{\theta}_2$

Since we know that $\overline{X}$ is an unbiased estimator of the mean, we found that the best estimator of θ is biased!

3.4 Relative efficiency

In this section, we suppose that the estimators that we evaluate are *unbiased*. In this case, the MSE of an estimator equals its variance. Suppose we compare two unbiased estimators. Which of these estimators is better?

Usually, an estimator with smaller variance in preferable. One quantitative measure that statisticians use to compare the unbiased estimators is their relative efficiency.

Definition 3.4.1. Given two unbiased estimators $\widehat{\theta}_1$ and $\widehat{\theta}_0$ of the same parameter θ, the relative efficiency of $\widehat{\theta}_1$ relative to $\widehat{\theta}_0$ is defined to be the ratio of their variances

$$eff(\widehat{\theta}_1, \widehat{\theta}_0) = \frac{Var(\widehat{\theta}_0)}{Var(\widehat{\theta}_1)}.$$

We can think about $\widehat{\theta}_0$ as a reference estimator. The estimator $\widehat{\theta}_1$ with relative efficiency which is greater than 1 is better than $\widehat{\theta}_0$ since its variance is smaller and so $\widehat{\theta}_1$ is a more accurate estimator of θ than $\widehat{\theta}_0$! Note that we assumed from outset that both estimators are unbiased.

- Simple example: $Y_1, \ldots, Y_9 \sim N(\mu_Y, 1)$. Want to estimate μ_Y.
  1. $\widehat{\theta}_1 = \frac{1}{9}\sum_{i=1}^{9} Y_i$,
  2. $\widehat{\theta}_2 = \frac{1}{2}(Y_1 + Y_2)$.
- Both $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are unbiased estimator for θ!
  - $\mathbb{E}\widehat{\theta}_1 = \theta$ and $\mathbb{E}\widehat{\theta}_2 = \theta$;
- $\mathbb{V}\text{ar}(\widehat{\theta}_1) = \frac{1}{9}\mathbb{V}\text{ar}(Y_1) = \frac{1}{9}$;
- $\mathbb{V}\text{ar}(\widehat{\theta}_2) = \frac{1}{2}\mathbb{V}\text{ar}(Y_1) = \frac{1}{2}$;
- $Reff(\widehat{\theta}_1, \widehat{\theta}_2) = \frac{9}{2} > 1$;
- $\widehat{\theta}_1$ is better than $\widehat{\theta}_2$; why?
- the estimator $\widehat{\theta}_2$ does not use all information available, and so it is less efficient.

Example 3.4.2. Let $Y_1, Y_2, \ldots, Y_n$ denote a random sample from the uniform distribution on the interval $[0, \theta]$. Two unbiased estimators for θ are

$$\widehat{\theta}_1 = 2\overline{Y} \quad \text{and } \widehat{\theta}_2 = \frac{n+1}{n} Y_{(n)},$$

where $Y_{(n)} = \max\{Y_1, Y_2, \ldots, Y_n\}$. Find the efficiency of $\widehat{\theta}_1$ relative to $\widehat{\theta}_2$.

Both are unbiased and therefore we only need to compute their variances. $\mathbb{V}\mathrm{ar}(\widehat{\theta}_1) = \mathbb{V}\mathrm{ar}(2\overline{Y}) = 4\mathbb{V}\mathrm{ar}(Y)/n = \theta^2/(3n)$. (This is because $Y = \theta X$, where X is a random variable that has uniform distribution on $[0,1]$ and the variance of X is $1/12$.

In order to compute $\mathbb{V}\mathrm{ar}(\widehat{\theta}_2) = \mathbb{V}\mathrm{ar}(\frac{n+1}{n}\max\{Y_1, Y_2, \ldots, Y_n\})$, we note that $Y_i = \theta X_i$, where X_i is uniformly distributed on $[0,1]$. Then,

$$\begin{aligned}\mathbb{V}\mathrm{ar}(\frac{n+1}{n}\max\{Y_1, Y_2, \ldots, Y_n\}) &= \mathbb{V}\mathrm{ar}(\theta\frac{n+1}{n}\max\{X_1, X_2, \ldots, X_n\}) \\ &= \theta^2\Big(\frac{n+1}{n}\Big)^2\mathbb{V}\mathrm{ar}(\max\{X_1, X_2, \ldots, X_n\}).\end{aligned}$$

So we only need to compute $\mathbb{V}\mathrm{ar}(\max\{X_1, X_2, \ldots, X_n\})$. To do this we need to find the density of $X_{(n)} = \max\{X_1, X_2, \ldots, X_n\}$.

Recall that $\mathbb{P}(X_{(n)} \le x) = \mathbb{P}(X_1 \le x, X_2 \le x, \ldots, X_n \le x)$, so we have for the cdf, $F_{X_{(n)}}(x) = (F_{X_1}(x))^n$, and for the density, $f_{X_{(n)}}(x) = nf_{X_1}(x)(F_{X_1}(x))^{n-1}$. In our particular case, $f_{X_{(n)}}(x) = nx^{n-1}$.

Then we calculate:

$$\begin{aligned}\mathbb{E}X_{(n)} &= n\int_0^1 xx^{n-1}dx = \frac{n}{n+1}, \\ \mathbb{E}X_{(n)}^2 &= n\int_0^1 x^2x^{n-1}dx = \frac{n}{n+2}, \\ \mathbb{V}\mathrm{ar}(X_{(n)}) &= \mathbb{E}X_{(n)}^2 - (\mathbb{E}X_{(n)})^2 = \frac{n}{n+2} - \frac{n^2}{(n+1)^2} = \frac{n(n+1)^2 - n^2(n+2)}{(n+2)(n+1)^2} \\ &= \frac{n}{(n+2)(n+1)^2},\end{aligned}$$

It follows that

$$\mathbb{V}\mathrm{ar}(\widehat{\theta}_2) = \theta^2\Big(\frac{n+1}{n}\Big)^2\frac{n}{(n+2)(n+1)^2} = \theta^2\frac{1}{n(n+2)},$$

and the relative efficiency

$$eff(\widehat{\theta}_2, \widehat{\theta}_1) = \frac{\mathbb{V}\mathrm{ar}(\widehat{\theta}_1)}{\mathbb{V}\mathrm{ar}(\widehat{\theta}_2)} = \frac{\theta^2}{3n}\Big/\frac{\theta^2}{n(n+2)} = \frac{n+2}{3}.$$

Hence, the second estimator is much more efficient than the first one.

3.5 Sufficient statistics

There is a huge multitude of functions of the data that we can consider in the search for a good estimator. So it is worthwhile to check if we can reduce the data to one or just a few of summary statistics. This is the main idea behind the concept of sufficient statistics.

Definition 3.5.1. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution with unknown parameter θ. A statistic $T = T(X_1, X_2, \ldots, X_n)$ is said to be **sufficient for** θ if the conditional distribution of $(X_1, X_2, \ldots, X_n)$, given $T = t$, does not depend on θ. That is,

$$\mathbb{P}_\theta(X_1, \ldots, X_n | T(X_1, \ldots, X_n) = t)$$

depends only on t but does not depend on θ.

Intuitively, if we know that $T = t$, then revealing the complete information about $X_1, X_2, \ldots, X_n$ does not give us any additional information about θ.

Example 3.5.2. Let $X_1, \ldots, X_n$ be an i.i.d. sample from the Bernoulli distribution with parameter $\theta = p$. That is, X_i takes values 1 and 0 with probabilities p and $1 - p$ respectively. Consider a statistic $T = X_1 + \ldots + X_n$. Then

$$\begin{aligned} p(x_1, \ldots, x_n | T = t) &= \frac{p(x_1, \ldots, x_n, T = t)}{p(T = t)} \\ &= \frac{p^{\sum_i x_i}(1-p)^{n-\sum_i x_i}}{\binom{n}{t} p^{\sum_i x_i}(1-p)^{n-\sum_i x_i}} \delta(\sum_i x_i - t) \\ &= \frac{1}{\binom{n}{t}} \delta(\sum_i x_i - t), \end{aligned}$$

where $\delta(\sum_i x_i - t)$ equals 1 if $\sum_i x_i = t$ and 0, otherwise. We can see that this conditional probability does not depend on p.

In principle, a sufficient statistic can be a vector, that is, it can consists of several functions. For example, if you take a vector of order statistics,

$T = (X_{(1)}, X_{(2)}, \dots X_{(n)})$, then it is a sufficient statistic. However, we don't gain very much by considering such statistics since they do not reduce the data.

If we take a function of a vector of sufficient statistics and reduce the dimension, then it can potentially break the sufficiency, however in some cases the resulting function is still sufficient. For example any invertible function of a sufficient statistic is sufficient, – it does not loose any information.

A sufficient statistic is minimal if it can be written as a function of any other of sufficient statistics. (A minimal sufficient statistic exists under mild conditions on the distribution of the data but there are some counterexamples.)

Why do we care about sufficient statistics?

In some cases, we can find a good estimator of a parameter θ by a 2-step procedure:

1. Find a sufficient statistic $T(X_1, X_2, \dots X_n)$ for parameter θ. Informally, it contains all information about the parameter which is available in the data.

2. Find an unbiased estimator of θ, which is a function of T. We can hope that all relevant information in data was used and the estimator cannot be further improved.

How we can find a good sufficient statistics? We should try to factorize the likelihood function.

Recall that the likelihood function is the *same thing* as the joint distribution density or joint distribution pmf of data $X_1, \dots, X_n$. In this course we consider only the situation when $X_1, \dots, X_n$ are independent and identically distributed, so

1. If $X_1, \dots, X_n$ are discrete random variables, and $p(x) := \mathbb{P}(X = x|\theta)$ is the probability mass function of each of them, then

$$L(\theta \mid x_1, \dots, x_n) = p_{X_1,\dots,X_n}(x_1, \dots, x_n \mid \theta) = \prod_{i=1}^{n} p(x_i|\theta).$$

2. If $X_1, \ldots, X_n$ are continuous random variables and $f(x|\theta)$ is the density of each of them, then

$$L(\theta \mid x_1, \ldots, x_n) = f_{X_1,\ldots,X_n}(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

To simplify the notation, we sometimes write $L(\theta)$ instead of $L(\theta \mid x_1, \ldots, x_n)$.

Theorem 3.5.3 (Fisher's Factorization Criterion)**.** *A statistic T is a sufficient statistic for parameter θ if and only if $L(\theta)$ can be factorized into two nonnegative functions:*

$$L(\theta|x_1, \ldots x_n) = g(\theta, t(x_1, \ldots, x_n)) \times h(x_1, \ldots, x_n)$$

Here $g(\theta, t)$ is a function only of t (the observed value of T) and θ and the function $h(x_1, \ldots, x_n)$ does not depend on θ at all.

Let us indicate how one can prove one direction of this theorem. Suppose that the factorization holds. Then, one can write:

$$\begin{aligned} p(x_1, \ldots, x_n|T = t) &= \frac{p(x_1, \ldots, x_n, T = t)}{p(T = t)} \\ &= \frac{g(\theta, t)h(x_1, \ldots, x_n)\delta\big(t(x_1, \ldots, x_n) - t\big)}{\sum_{\vec{x}:t(\vec{x})=t} g(\theta, t)h(x_1, \ldots, x_n)} \\ &= \frac{h(x_1, \ldots, x_n)\delta\big(t(x_1, \ldots, x_n) - t\big)}{\sum_{\vec{x}:t(\vec{x})=t} h(x_1, \ldots, x_n)}, \end{aligned}$$

and the result does not depend on the parameter θ.

Example 3.5.4.

- $Y_1, \ldots, Y_n \sim B(p)$. Find a sufficient statistic for p.

- Likelihood:

$$\begin{aligned} L(p) &= \prod_{i=1}^{n} \{p^{y_i}(1-p)^{(1-y_i)}\} \\ &= p^{\sum_{i=1}^n y_i}(1-p)^{(n-\sum_{i=1}^n y_i)} \\ &= \left\{\frac{p}{1-p}\right\}^{\sum_{i=1}^n y_i} (1-p)^n \times 1 \end{aligned}$$

- We can define a statistic $T = \sum_{i=1}^{n} Y_i$. Then we will have
  - $g_p(t) = \left\{\frac{p}{1-p}\right\}^t (1-p)^n$, and $h(y_1, \ldots, y_n) = 1$
  - The first term only depends on p and T (or t)
  - The second term does not depend on p

Example 3.5.5.

- $Y_i \sim_{iid} Poisson(\lambda)$, i.e. $\sim p(y) = e^{-\lambda}\frac{\lambda^y}{y!}$
- Likelihood (use independence):

$$L(\lambda) = \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{y_i}}{y_i!}$$

- $$L(\lambda) = e^{-n\lambda}\frac{\lambda^{\sum_{i=1}^n y_i}}{\prod_{i=1}^n (y_i!)} = e^{-n\lambda}\lambda^{\sum_{i=1}^n y_i} \times \frac{1}{\prod_{i=1}^n (y_i!)}$$

Example 3.5.6. Let $X_1, X_2, \ldots, X_n$ be a random sample in which X_i is exponentially distributed (remember life of smartphones?) with parameter θ. Find a sufficient statistic for θ.

Example 3.5.7.

- $Y_i \sim_{iid} N(\mu, \sigma^2)$, i.e.

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

- Likelihood:

$$\begin{aligned} L(\cdot) &= \prod_{i=1}^{n}\left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(y_i-\mu)^2}{2\sigma^2}\}\right] \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{\sum_{i=1}^n (y_i-\mu)^2}{2\sigma^2}\right\} \end{aligned}$$

- Note that $\sum_{i=1}^n (y_i-\mu)^2 = \sum_{i=1}^n (y_i - \overline{y} + \overline{y} - \mu)^2 = \sum_{i=1}^n (y_i - \overline{y})^2 + n(\overline{y}-\mu)^2 = (n-1)s^2 + n(\overline{y}-\mu)^2$

- Thus we have

$$L(\cdot) = (2\pi\sigma^2)^{-\frac{n}{2}} \cdot \exp\left\{-\frac{(n-1)s^2}{2\sigma^2}\right\} \cdot \exp\left\{-\frac{n(\overline{y}-\mu)^2}{2\sigma^2}\right\}$$

$$L(\cdot) = (2\pi\sigma^2)^{-\frac{n}{2}} \cdot \exp\left\{-\frac{(n-1)s^2}{2\sigma^2}\right\} \cdot \exp\left\{-\frac{n(\overline{y}-\mu)^2}{2\sigma^2}\right\}$$

- The argument in $L(\cdot)$ is not specified because there are two situations.

1. μ is unknown and σ^2 is known $\Rightarrow L(\mu)$

$$L(\cdot) = \exp\left\{-\frac{n(\overline{y}-\mu)^2}{2\sigma^2}\right\} \cdot (2\pi\sigma^2)^{-\frac{n}{2}} \cdot \exp\left\{-\frac{(n-1)s^2}{2\sigma^2}\right\}$$

   $\overline{Y}$ is a sufficient statistic for μ

2. Both μ and σ^2 are unknown $\Rightarrow L(\mu, \sigma^2)$

$$L(\cdot) = (2\pi\sigma^2)^{-\frac{n}{2}} \cdot \exp\left\{-\frac{(n-1)s^2}{2\sigma^2}\right\} \cdot \exp\left\{-\frac{n(\overline{y}-\mu)^2}{2\sigma^2}\right\} \times 1$$

   The pair $(\overline{Y}, S^2)$ is a sufficient statistic for μ and σ^2

Quiz 3.5.8. Every function of a sufficient statistics is a sufficient statistic.

A. True

B. False

Quiz 3.5.9. Every strictly decreasing function of a sufficient statistics is a sufficient statistic.

A. True

B. False

Even a minimal sufficient statistic is not unique!

Example 3.5.10 (Uniform distribution). Let $X_1, X_2, \ldots, X_n$ be uniformly distributed in $(0, \theta)$. What is a sufficient statistic for θ?

- Density of one random variable:

$$f_{X_i}(x_i) = \begin{cases} 1/\theta, & 0 < x_i < \theta, \\ 0, & \text{otherwise} \end{cases}$$
$$= \frac{1}{\theta} \mathbb{1}_{0<x_i<\theta},$$

  where $\mathbb{1}_{0<x_i<\theta}$ is the indicator function of the event $\{0 < x_i < \theta\}$.

- Likelihood:

$$L(\theta) = \prod_{i=1}^{n} \left[\mathbb{1}_{0 \le x_i \le \theta} \frac{1}{\theta} \right]$$

- Note that $\prod_{i=1}^{n} \mathbb{1}_{\theta \ge x_i} = \mathbb{1}_{\theta \ge x_{(n)}}$. Not very obvious. Think!

- Thus the likelihood $L(\theta) = \mathbb{1}_{\theta \ge x_{(n)}} \frac{1}{\theta^n} = \mathbb{1}_{\theta \ge x_{(n)}} \frac{1}{\theta^n} \times 1$

- Hence $T = X_{(n)}$ is a sufficient statistic for θ.

- **Any 1-to-1 function of a sufficient statistic is also a sufficient statistic for the same parameter.** $\Rightarrow$ the unbiased estimator $\widehat{\theta} = \frac{n+1}{n} X_{(n)}$ is also a sufficient statistic for θ.

- Note that previously we found that this estimator is much more efficient than another unbiased estimator $2\overline{X}$. This is a reflection of a general fact which is called the Rao-Blackwell theorem.

Exercise 3.5.11.

- $f(y) = \exp[-(y - \theta)]$ for $y > \theta$

- Sufficient statistic?

3.6 Rao-Blackwell Theorem and Minimum-Variance Unbiased Estimator

This section relates sufficiency with efficiency and unbiasedness. It will tell us why a sufficient statistic is very useful in statistical inference.

- We have learned two good qualities of an estimator $\widehat{\theta}$ for a parameter θ:
  - Unbiasenss: $E\widehat{\theta} = \theta$;
  - Low variance: $Var(\widehat{\theta})$ is small;
- Relative efficiency of two estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$

$$\text{Reff}(\widehat{\theta}_1, \widehat{\theta}_2) = \frac{Var(\widehat{\theta}_2)}{Var(\widehat{\theta}_1)};$$

  whichever has the smaller variance is more efficient (better)!

Definition 3.6.1. An unbiased estimator $\widehat{\theta}$ is called *MVUE (Minimal Variance Unbiased Estimator)* if for every other estimator $\hat{\theta}'$ and every value of the parameter $\theta \in \Theta$, $\mathbb{V}\text{ar}(\widehat{\theta}) \leq \mathbb{V}\text{ar}(\hat{\theta}')$.

(Sometimes it is called UMVUE, where the first U stands for "uniform" to emphasize that the minimal variance property should hold for every $\theta \in \Theta$.)

Since the unbiased estimators does not always exist, MVUE are even more rare. However, if an MVUE exists, how can we find it?

The main idea of the following theorem is that we can always improve any unbiased estimator by conditioning it on a sufficient statistic. In particular, if an MVUE exists, it must be a function of a sufficient statistic.

Theorem 3.6.2 (Rao-Blackwell Theorem)**.** *Let $\widehat{\theta}$ be an unbiased estimator for θ such that $\mathbb{V}\text{ar}(\widehat{\theta}) < \infty$. If T is a sufficient statistic for θ, define $\widehat{\theta}^* = \mathbb{E}(\widehat{\theta}|T)$. Then, for all θ, $\mathbb{E}\widehat{\theta}^* = \theta$ and $\mathbb{V}\text{ar}(\widehat{\theta}^*) \leq \mathbb{V}\text{ar}(\widehat{\theta})$.*

- Given an unbiased estimator $\widehat{\theta}$ and a sufficient statistic T, we can find a modified estimator $\widehat{\theta}^*$, which is improved in the sense that
  - $\widehat{\theta}^*$ is still unbiased;
  - $\widehat{\theta}^*$ has a smaller (or at least no larger) variance than $\widehat{\theta}$;

- Remarks:
  - $\widehat{\theta}^*$ is a function of T
  - $\widehat{\theta}^*$ is random
  - If $\widehat{\theta}$ is already a function of T, then $\mathbb{E}(\widehat{\theta}\,|T) = \widehat{\theta}$, i.e., taking the conditional expectation does not change anything, in particular it does not improve the efficiency.

Proof. Because T is sufficient for θ, the conditional distribution of any statistic (including $\widehat{\theta}$), given T, does not depend on θ. Thus, $\widehat{\theta}^* = \mathbb{E}(\widehat{\theta}|T)$ is not a function of θ and is therefore a statistic. The fact that $\widehat{\theta}^*$ is almost obvious from the law of repeated expectation.

$$\mathbb{E}\widehat{\theta}^* = \mathbb{E}\Big[\mathbb{E}(\widehat{\theta}|T)\Big] = \mathbb{E}\widehat{\theta} = \theta.$$

For the variance we use another theorem about conditional expectations:

$$\begin{aligned}\mathbb{V}\mathrm{ar}(\widehat{\theta}) &= \mathbb{V}\mathrm{ar}\Big[\mathbb{E}(\widehat{\theta}|T)\Big] + \mathbb{E}\Big[\mathbb{V}\mathrm{ar}(\widehat{\theta}|T)\Big] \\ &= \mathbb{V}\mathrm{ar}(\widehat{\theta}^*) + \mathbb{E}\Big[\mathbb{V}\mathrm{ar}(\widehat{\theta}|T)\Big].\end{aligned}$$

Since the second term is non-negative, we find that $\mathbb{V}\mathrm{ar}(\widehat{\theta}^*) < \mathbb{V}\mathrm{ar}(\widehat{\theta})$. □

This theorem implies that if an unbiased estimator $\widehat{\theta}$ is NOT a function of a sufficient statistics T the we can find another unbiased estimator $\widehat{\theta}^* = \mathbb{E}(\widehat{\theta}|T)$ which is a function of T at least as good as $\widehat{\theta}$. However, it does NOT imply that this estimator is an MVUE. Perhaps we can find another sufficient statistic T_2 and improve $\widehat{\theta}^*$ by taking a conditional expectation with respect to T_2.

A natural conjecture is that if T is a *minimal* sufficient statistic and some function $\widehat{\theta} = \widehat{\theta}(T)$ is an unbiased estimator of θ, then $\widehat{\theta}$ is an MVUE. This is again not quite correct. One has to impose a stronger requirement that T is a *complete* sufficient statistic. In this case, it is also guaranteed to be a minimal sufficient statistic and a function $\widehat{\theta} = \widehat{\theta}(T)$ which is unbiased for θ is indeed MVUE.

This raises questions about what a complete sufficient statistic is and how one can check that a sufficient statistic is complete. We will not be concerned with these questions in this course and simply promise that in all our examples the sufficient statistics obtained by factorization theorem will be complete and sufficient.

Routine to find the MVUE

1. Factorize the likelihood function and find a (minimal) sufficient statistic T;

2. find a function of T which is unbiased for the parameter of interest θ;

Example 3.6.3 (Exponential). Suppose $X_1, X_2, \ldots, X_n$ all from the exponential distribution with the parameter θ. What is an MVUE for θ ?

We showed that $T = (X_1 + \ldots + X_n)$ is a sufficient statistic. In fact it is a minimal and complete statistic, and since $\overline{X} = T/n$ is unbiased for θ hence it is an MVUE.

Example 3.6.4 (Bernoulli). Let X_i's are iid Bernoulli with parameter p. What is an MVUE for p?

Same argument works as in the previous example. We already proved that $T = \sum_{i=1}^n X_i$ is a sufficient statistic for p. Hence $\widehat{p} = \overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ is MVUE (for p).

Example 3.6.5 (Normal). $X_1, \ldots, X_n$ are i.i.d. from normal distribution $N(\mu, \sigma^2)$. What is the MVUE for μ and σ^2.

We have shown that $\overline{X}$ and S^2 are joint sufficient statistics for μ and σ^2. In addition, we know that $\overline{X}$ and S^2 are unbiased estimators of μ and σ. Therefore, $\overline{X}$ and S^2 as the MVUEs for μ and σ^2, respectively.

Example 3.6.6. $X_1, \ldots, X_n$ are i.i.d. from a distribution $f(x|\theta) = (2y/\theta)e^{-y^2/\theta}$ for $y > 0$. MVUE for θ?

[Quizzes]

Chapter 4

Methods of estimation

Suppose that we are looking for a good estimator $\widehat{\theta}(X_1, \ldots, X_n)$ for a parameter θ. The method in the previous section asks us to find a minimal complete sufficient statistic $T = T(X_1, \ldots, X_n)$ by using a factorization criterion and then find a function of T which would be an unbiased estimator of θ. This will lead us to a MVUE. Unfortunately, even if T is known, it is often difficult to find $\widehat{\theta}(T)$ which would be unbiased for θ. In fact, in some cases no unbiased estimator for θ exists.

For this reason we are looking for other methods to construct an estimator, which would be easy to construct and which would have a small MSE in a large sample.

We will consider two such methods, Method of Moments Estimation (MME) and Maximum Likelihood Estimation (MLE).

4.1 Method of Moments Estimation

We consider our usual situation when data $X_1, \ldots, X_n$ is an i.i.d sample from a distribution $F_\theta(x)$ which belongs to a family of distributions parameterized by $\theta \in \Theta$. In general, parameter θ can be a vector $\theta = (\theta_1, \ldots, \theta_s)$ that consists of several components. We want to estimate θ using the data sample.

Recall that the k-th population moment is simply the theoretical expec-

tation of an observation X_i, that is,

$$\mu_k(\theta) := \mathbb{E}(X_i)^k = \begin{cases} \int x^k f(x|\theta)\, dx & \text{if } X_i \text{ are continuous r.v.,} \\ \sum_x x^k p(x|\theta) & \text{if } X_i \text{ are discrete r.v..} \end{cases}$$

Note that the population moments are all functions of the parameter θ (and do not depend on the sample data).

In contrast, the sample moments are functions of the data sample X_i. (They are random quantities and their distribution depends on θ.) We denote the k-th sample moment m_k.

$$m_k = m_k(X_1, \ldots, X_n) := \frac{1}{n} \sum_{i=1}^{n} X_i^k.$$

For example, the first sample moment equals to the sample mean $m_1 = \overline{X}$, the second sample moment can be expressed in terms of the sample variance and the sample mean: $m_2 = (n-1)S^2 + n(\overline{X})^2$.

It should be emphasized that the sample moments are all functions of the data, i.e., they can all be calculated using the data. And they are all random.

The main idea behind Method of Moments Estimator is that by the Law of Large Numbers the sample moments converge to population moments:

$$m_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k \xrightarrow{\mathbb{P}} \mathbb{E}X_i^k = \mu_k(\theta),$$

in probability as $n \to \infty$.

So for large n we have,

$$m_k(X_1, \ldots, X_n) = \mu_k(\theta) + \varepsilon_k,$$

where ε_k is very small with probability close to 1. This means that empirical moments are consistent estimators of the population moments. In addition, we know the form of the functions $\mu_k(\theta)$, although we do not know the value of θ. Hence we can invert the system of these functions and get an estimator of θ from the estimator $\widehat{\mu}_k = m_k(X_1, \ldots, X_n)$ of $\mu_k(\theta)$.

So, the idea is to ignore ε_k and solve the system of equations

$$m_k(X_1, \ldots, X_n) = \mu_k(\widehat{\theta}), \quad , k = 1, 2, \ldots, s$$

for $\widehat{\theta}$. (It might be just one equation if the parameter is one-dimensional vector, that is, a real number.)

The solution is given by the inverse function $\widehat{\theta} = \mu_k^{(-1)}(m_k)$. If the inverse function $\mu_k^{(-1)}$ is a continuous function then we can use the continuous mapping theorem and show that $\widehat{\theta} \to \theta$ in probability as $n \to \infty$, in other words, that $\widehat{\theta}$ is a consistent estimator of θ.

How many moments we need for estimation? Typically, if we need to estimate a vector that consists of s parameters $\theta_1, \ldots \theta_s$, we use the first s moments. However it might happen that one of the first theoretical moments $\mu_k(\theta)$ actually does not depend on the parameter of interest. For example, if for every θ the distribution $F_\theta(x)$ has the density function symmetric relative to the origin $x = 0$, then the first population moment (population mean or expectation) is zero for every θ, and therefore this first moment is not going to help us in the estimation of parameters.

Practical steps:

1. Calculate the first K population moments $\mu_k(\vec{\theta})$ as functions of the vector of unknown parameters $\vec{\theta} = (\theta_1, \ldots, \theta_s)$. (Typically, if there are s unknown parameters, then one needs to calculate the first s moments: $K = s$.)

2. Write the first s sample moments m_k, $k = 1, \ldots, K$, as functions of data vector X_i.

3. Match the moments: Solve the system of K equations $\mu_k(\widehat{\theta}) = m_k(\vec{X})$ for $\widehat{\theta}$.

4. The solution is a (vector) function $\widehat{\theta}$ of the sample moments m_k and hence of the data X_i, because the sample moments are functions of the data. Since we believe that the equations are approximately true, we also believe that the solution $\widehat{\theta}$ is close to the true value of the

parameter θ. These solutions give us the desired Method of Moments Estimator ($\widehat{\theta}_{MME}$).

MME and Consistency Under some mild regularity conditions on the distribution of data, Method of Moment estimators are consistent. This is roughly because

- Under the conditions, we have

$$m_k(X_1, \ldots, X_n) \xrightarrow{\mathbb{P}} \mu_k(\theta_1, \ldots, \theta_s), \text{ for } k = 1, \ldots, s,$$

  when $n \to \infty$.

- The solution to the sistem of equations "$m_k = \mu_k(\widehat{\theta}_1, \ldots, \widehat{\theta}_s)$, $k = 1, \ldots, s$" is a continuous map

$$\widehat{\theta}_k = \widehat{\theta}_k(m_1, \ldots, m_s),$$

  where $k = 1, \ldots, s$. It is the inverse of the moment transformation μ, that sends parameters $\theta_1, \theta, \theta_s$ to moments $\mu_1, \ldots, mu_s$, so we could write (in vector form) $\widehat{\theta}(m_1, \ldots m_s) = \mu^{(-1)}(m_1, \ldots, m_s)$.

- By the continuous mapping theorem, we may conclude (again for conciseness using vector notation for the parameter θ and the estimator $\widehat{\theta}$).

$$\widehat{\theta} = \widehat{\theta}(m_1, \ldots, m_s) \xrightarrow{\mathbb{P}} \mu^{-1}(\mu_1(\theta), \ldots \mu_s(\theta)) = \theta.$$

Example 4.1.1 (Normal). The data $X_1, \ldots, X_n$ are distributed according the normal distribution $N(\mu, \sigma^2)$

- There are 2 parameters to estimate. The vector parameter θ has two components, $\theta = (\mu, \sigma^2)$

- The first population moment is $\mu_1(\theta) := \mathbb{E}(X_i) = \mu$. (There is a clash of notation here: $\mu_1(\theta)$ is the first population moment and μ also denotes the first component of the parameter vector θ.)

- The second population moment is $\mu_2(\theta) := \mathbb{E}X_i^2 = \mathbb{V}\text{ar}(X_i) + (\mathbb{E}X_i)^2 = \sigma^2 + \mu^2$.

- The first sample moment is $m_1 := \frac{1}{n}\sum_{i=1}^n X_i = \overline{X}$.
- The second sample moment is $m_2 := \frac{1}{n}\sum_{i=1}^n X_i^2$.

By matching the population and sample moments we obtain equations:

$$\widehat{\mu} = m_1 = \overline{X}$$
$$\widehat{\sigma}^2 + \widehat{\mu}^2 = m_2 = \frac{1}{n}\sum_{i=1}^n X_i^2$$

After solving these equations we get:

$$\widehat{\mu} = \overline{X};$$
$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^n X_i^2 - (\overline{X})^2$$

These are the MME estimators of the parameters. Note that the MME estimator for the variance is different from the standard estimator, the sample variance, which is

$$S^2 = \frac{1}{n-1}\Big(\sum_{i=1}^n X_i^2 - n(\overline{X})^2\Big)$$

It is easy to see that they are related by the formula:

$$\widehat{\sigma^2}_{MME} = \frac{n-1}{n}S^2.$$

In particular, the MME estimator is biased.

Example 4.1.2 (Poisson with unusual parameter). The data sample $X_1, \dots X_n$ is distributed according to the Poisson distribution with parameter λ. Find the estimator of the parameter $\theta = 1/\lambda$.

- Only 1 parameter to be estimated.
- The first population moment is $\mu(\theta) = \mathbb{E}X_i = \lambda = 1/\theta$
- The first sample moment is $m_1 = \overline{X}$

- Match the two quantities above and solve for $\widehat{\theta}$:

$$1/\widehat{\theta} = \overline{X}$$

- Hence,

$$\widehat{\theta}^{MME} = 1/\overline{X}$$

If you attentively examine the example above then you will see that we could estimate the parameter λ (the population mean) by the sample mean $\overline{X}$. This would give an MME estimator of λ. Then, since parameters θ and λ are in one-to-one correspondence to each other, therefore solving MME equations for $\widehat{\theta}$ can be done by solving MME equations for $\widehat{\lambda}$ and then using the one-to-one relation between the parameters. This would give us the MME estimator $\widehat{\theta} = 1/\widehat{\lambda} = 1/\overline{X}$.

This is a manifestation of the general principle valid for MME estimators.

Plug-in or Invariance property of MME

- If the parameter of interest ψ is a function of another parameter θ whose MME is relative easy to find, i.e., if

$$\psi = h(\theta)$$

and $\widehat{\theta}_{MME}$ is easy to obtain, then

$$\widehat{\psi}_{MME} = h(\widehat{\theta}_{MME}),$$

i.e. we can apply the function h to $\widehat{\theta}_{MME}$ to obtain $\widehat{\psi}_{MME}$

- This is often easier then the "standard" procedure to find MME of ψ, where you need to redo the whole process.

Example 4.1.3. Let $X_1, \ldots, X_n$ be i.i.d observation from the uniform distribution on the interval $[0, \theta]$. What is the Method of Moments estimator of θ?

The first population moment is $\mu_1 = \theta/2$. The first sample moment is $\overline{X}$. Therefore, the MM estimator is $\widehat{\theta} = 2\overline{X}$. It is an unbiased estimator since $\mathbb{E}\widehat{\theta} = 2\mathbb{E}\overline{X} = 2\mathbb{E}X_i = \theta$.

Note, however, that $\overline{X}$ is not a sufficient statistic for θ. Indeed, the minimal sufficient statistic in this example is $X_{(n)} = \max\{X_1, \ldots, X_n\}$.

So, this is an example of an MME estimator which is not a function of a sufficient statistic. So it has no chance to be an MVUE. Even though it is unbiased, its variance could be reduced by conditioning on the sufficient statistic.

Reflection

- The four point estimators back in Chapter 8 (for the mean, proportion, differences in means and proportions) were all MMEs.
- In practice, Method of Moments is a very intuitive way to find an estimator. It requires only the ability to calculate the moments in terms of the parameters and invert this relation.
- Sometimes, MME gives biased estimators but at least it is consistent (under very mild conditions).
- One of its deficiencies that it is not always a function of a sufficient statistic.

Here is a couple of additional examples.

Exercise 4.1.4. X_i's are Gamma(α, β) distribution.

Exercise 4.1.5.

$$Y_i \sim f(y) = \left(\frac{2}{\theta^2}\right)(\theta - y)\mathbb{1}_{0 \le y \le \theta}$$

- Use MME to find an estimator for θ
- Is this estimator is a function of a sufficient statistic?

4.2 Maximum Likelihood Estimation (MLE)

The most popular method to find an estimator is the method of maximum likelihood, MLE.

Definition 4.2.1. The maximum likelihood estimator $\widehat{\theta}$ is the value of the parameter θ, at which the likelihood function $L(\theta|x_1, \dots x_n)$ takes its maximum value.

Informally, if we know that the probability to observe data sample $(x_1, \dots, x_n)$ is 90% if the value of the parameter were θ_1 versus 10% if the value of the parameter were θ_2, then we might prefer θ_1 as an estimator of true unknown θ.

The steps in finding MLE are easy: write down the maximum likelihood function and find its maximum with respect to the parameter θ. It turns out that in many cases, it is technically easier to look for maximum of log-likelihood function $\ell(\theta|\vec{x}) := \log(L(\theta|\vec{x}))$ instead $L(\theta)$. (It is maximized at the same value of θ.)

The main technical question in ML estimation is how to find the global maximum.

The maximization can be done by a computer algorithm or analytically. If the maximization is done by computer, there are many specialized algorithms suitable for statistical applications. One of them is EM algorithm. If the maximization is done analytically, we aim to solve equations $\frac{d}{d\theta}\ell(\theta|\vec{(y)}) = 0$. (They are called the first order conditions "FOC" for the extremal points.) Note, however, that the solution of these equations are all local extrema: local maxima, local minima and saddle points, so one should be careful to choose the global maximum among all possible solutions. In addition, the maximum can occur on the boundary of the set of all possible values of θ. In this case, the FOC will not give you information about the global maximum.

Consistency: The important fact about MLE is that this estimator is consistent under the mild distributional conditions. It shares this good property with the Method of Moments estimator.

Let us consider some examples.

Example 4.2.2 (Exponential). Let $X_1, \dots, X_n$ be i. i. d. observations distributed according to the exponential distribution with parameter θ. What is the ML estimator for θ?

The density of an individual observation is

$$f(x_i) = \frac{1}{\theta} e^{-x_i/\theta}.$$

Since the observations are independent, the likelihood is just the product of the density functions:

$$L(\theta|x_1, \ldots, x_n) = \frac{1}{\theta^n} \prod_{i=1}^{n} e^{-x_i/\theta} = \frac{1}{\theta^n} e^{-(\sum_{i=1}^{n} x_i)/\theta}$$

Hence the log-likelihood is

$$\ell(\theta|x_1, \ldots, x_n) = \log L(\theta|x_1, \ldots, x_n) = -n \log \theta - \frac{1}{\theta} \sum_{i=1}^{n} x_i$$

The first-order condition equation is

$$\begin{aligned} \frac{d}{d\theta} \ell(\theta|x_1, \ldots, x_n) &= 0, \\ -n\frac{1}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^{n} x_i &= 0, \\ \sum_{i=1}^{n} x_i &= n\theta, \end{aligned}$$

and the ML estimator is $\widehat{\theta} = \overline{X}$. Note that it coincides with the MM estimator.

Example 4.2.3. The data $Y_1, \ldots, Y_n$ are from the Bernoulli distribution with parameter p. What is the MLE for p?

- The likelihood function is

$$L(p|\vec{y}) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i} = p^t(1-p)^{n-t},$$

  where $t = \sum_{i=1}^{n} y_i$.

- The log-likelihood function is

$$\ell(p|\vec{y}) := \log(L(p|\vec{y})) = t \log(p) + (n-t) \log(1-p)$$

- $$\frac{d}{dp}\ell(p|\vec{y}) = \frac{t}{p} - \frac{n-t}{1-p}$$

- Set
$$\ell'(p) = \frac{t}{p} - \frac{n-t}{1-p} = 0$$
We obtain
$$\frac{t}{p} = \frac{n-t}{1-p} \Rightarrow t - tp = np - tp \Rightarrow p = \frac{t}{n}$$

- Hence $\widehat{p}_{MLE} = \frac{T}{n}$
- Recall that we proved that this is an unbiased estimator, and since it is a function of a minimal sufficient statistic, it is the MVUE. It is also consistent since its variance goes to 0.
- Same as the MME estimator.

Exercise 4.2.4. A sample $X_1, \ldots, X_n$ is taken from the binomial distribution with parameters (k, θ). Assume k is known. What is the MLE for θ?

Hint. The likelihood function is
$$L(\theta, x_1, \ldots x_n) = \prod_{i=1}^{n} \binom{k}{x_i} \theta^{x_i}(1-\theta)^{k-x_i}.$$

Example 4.2.5 (Normal). Let $X_1, \ldots, X_n$ be i.i.d observations from a normal distribution, $N(\mu, \sigma^2)$. What are the ML estimators of μ and σ^2?

- Likelihood function:
$$\begin{aligned} L(\mu, \sigma^2) &= \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(x_i-\mu)^2}{2\sigma^2}\} \right] \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{ -\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2} \right\} \end{aligned}$$

- Log-likelihood:

$$\ell(\mu,\sigma^2) = \log(L(\mu,\sigma^2)) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}$$
$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}$$

- There are two unknown variables in this function. We want to calculate the partial derivative separately and set each to be zero, and then solve the unknowns.

$$\frac{\partial\{\ell(\mu,\sigma^2)\}}{\partial\mu} = \frac{2\sum_{i=1}^n (x_i-\mu)}{2\sigma^2} = \frac{\sum_{i=1}^n y_i - n\mu}{\sigma^2} = \frac{\overline{x}-\mu}{\sigma^2/n}$$

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

- Set both to zero.

$$\frac{\overline{x}-\mu}{\sigma^2/n} = 0 \Rightarrow \mu = \overline{x}$$

$$-\frac{n}{2}\frac{1}{\sigma^2} + \frac{\sum_{i=1}^n (x_i-\mu)^2}{2}\frac{1}{\sigma^4} = 0 \Rightarrow \sigma^2 = \frac{1}{n}\sum_{i=1}^n (x_i-\mu)^2 = \frac{1}{n}\sum_{i=1}^n (x_i-\overline{x})^2$$

- Hence, $\widehat{\mu}_{MLE} = \overline{X}$, $\widehat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^n (X_i-\overline{X})^2 = S_0^2$. These are exactly the sample mean, and the alternative version of sample variance (but not S^2, the original sample variance). We know that $\overline{X}$ is unbiased and both $\overline{X}$ and S_0^2 are consistent. However, S_0^2 is a biased estimator of σ^2.

  In fact, as in the previous example, the ML estimator coincide with the MM estimator.

So far, all ML estimators coincided with MM estimators. Is there any advantage in ML estimation? Here is an example, where we get an ML estimator which is different from the MM estimator and which is better than MM estimator. It also illustrates that it is sometimes not enough to solve the first order conditions.

Example 4.2.6 (Uniform on $(0, \theta)$). Let $X_1, \ldots, X_n$ be i.i.d observations from the uniform distribution on the interval $[0, \theta]$. What is the ML estimator for θ.

- The density of X_i is $f(x_i) = 1/\theta$ if $x_i \in [0, \theta]$, and 0 otherwise.
- The likelihood function is the product of the densities, so

$$L(\theta | x_1, \ldots, x_n) = \frac{1}{\theta^n},$$

  if $\min x_1, \ldots x_n \geq 0$, and $\max x_1, \ldots, x_n \leq \theta$ and 0 otherwise.
- We can also write this

$$L(\theta) = \theta^{-n} \mathbb{1}_{\{x_{(n)} \leq \theta\}} \mathbb{1}_{\{x_{(1)} \geq 0\}}$$

- Or equivalently $L(\theta) = \theta^{-n}$ for $\theta \geq x_{(n)}$, and 0, otherwise. ← better here for MLE.
- The log-likelihood is $\ell(\theta) = -n \log(\theta)$ for $\theta \geq x_{(n)}$, and $-\infty$, otherwise.
- $\ell'(\theta) = -\frac{n}{\theta}$ for $\theta \geq x_{(n)}$
- But $-\frac{n}{\theta} = 0$ does not have a solution! What can be wrong?
- The maximum of the function $\ell(\theta)$ is reached at the boundary point!

Indeed $\ell'(\theta) = -\frac{n}{\theta} < 0$, hence $\ell(\theta)$ is a decreasing function. The smallest possible value for θ in the domain $\theta \geq x_{(n)}$ is its left boundary point $x_{(n)}$, which is the maximum point for the likelihood.

So, $\widehat{\theta}_{ML} = X_{(n)}$. Note that it is a function of a minimal sufficient statistic $X_{(n)}$. Is it biased? How does its variance compare with the MM estimator?

Let us repeat some information from Example 3.4.2. First, $\mathbb{E}X_{(n)} = \frac{n}{n+1}\theta$, so So the ML estimator is biased with bias $= \theta/(n+1)$. However its variance $\mathbb{V}\mathrm{ar}(X_n) = \frac{n}{(n+2)(n+1)^2}\theta^2 \sim \theta^2/n^2$ is much smaller than the variance of the unbiased MM estimator. ($\mathbb{V}\mathrm{ar}(2\overline{X}) = 4\mathbb{V}\mathrm{ar}(X_i)/n = \theta^2/(3n)$.) In particular, this estimator has smaller MSE for large n.

Here we see a clear difference between ML and MM estimators. The MM estimator $2\overline{X}$ is unbiased but it is not a function of a sufficient statistic. The

ML estimator is biased but it is a function of a minimal sufficient statistic. In particular, the bias of the ML estimator can be corrected and this will lead to an MVUE estimator.

Note, by the way, that if the domain of the density function depends on the parameter θ, it is often a warning sign that the boundary point of θ may play a major role in finding the MLE.

MLE is always a function of a sufficient statistic!

Theorem 4.2.7. *Suppose $T(X_1, \ldots, X_n)$ is a sufficient statistic for θ. Then $\widehat{\theta}_{MLE}$ can be written as a function of T, $\widehat{\theta}_{MLE}(X_1, \ldots, X_n) = \widehat{\theta}(T)$.*

In particular, $\widehat{\theta}(T)$ is a function of a complete minimal sufficient statistic as long as such a beast exists. This is rather appealing because this means that if are able to find a function of $\widehat{\theta}_{MLE}$ which is unbiased, then this function is an MVUE.

Proof. If T is a sufficient statistic, then by the factorization criterion,

$$L(\theta) = g(t, \theta)h(x_1, \ldots, x_n) \text{ and so } \ell(\theta) = \log(g(t, \theta)) + \log(h(y_1, \ldots, y_n))$$

- Since $\log(h(x_1, \ldots, x_n))$ has nothing to do with θ, as far as θ is concerned, $\log(h(x_1, \ldots, x_n))$ is a constant; hence the maximizer of $\ell(\theta)$ over θ is the same as the maximizer of $\log(g(t, \theta))$.
- The maximizer of $\log(g(t, \theta))$, over all possible θ, has to depend only on t.

Thus the ML estimator is a function of T. □

Plug-in property for MLE

Example 4.2.8 (Poisson with usual and unusual parameter). Suppose that $X_1, \ldots, X_n$ are i.i.d. from the Poisson distribution with parameter λ. What is the ML estimator for λ?

The Poisson distribution is discrete so we work with probability mass functions ("pmf"s). The pmf of one observation X_i is

$$p_{X_i}(x_i|\lambda) := \mathbb{P}[X_i = x_i] = e^{-\lambda}\frac{\lambda^{x_i}}{x_i!},$$

where x_i can take values $0, 1, 2, \ldots$. Since the observations are independent, the likelihood function is simply the product of pmfs of individual observations.

$$L(\lambda|x_1, \ldots, x_n) = \lambda^{\sum_{i=1}^n x_i} e^{-n\lambda} / \prod_{i=1}^n (x_i!).$$

Hence, the log-likelihood is

$$\ell(\lambda) = \Big(\sum_{i=1}^n x_i \Big) \log \lambda - n\lambda - \log(\prod_{i=1}^n (x_i!)).$$

So we can write the first order condition as

$$\ell'(\lambda) = \Big(\sum_{i=1}^n x_i \Big) \frac{1}{\lambda} - n = 0$$

The solution gives us the ML estimator $\widehat{\lambda}_{ML} = \overline{X}$. It coincides with the MM estimator.

Now suppose we use a different parameter in the model $\theta = 1/\lambda$ and want to find an ML estimator for θ. This simply means that now we write the distribution function for the observations in terms of θ not λ:

$$p_{X_i}(x_i|\theta) := e^{-1/\theta} \frac{(1/\theta)^{x_i}}{x_i!},$$

So the likelihood function will be

$$L(\theta|x_1, \ldots, x_n) = (1/\theta)^{\sum_{i=1}^n x_i} e^{-n(1/\theta)} / \prod_{i=1}^n (x_i!).$$

Now it is rather obvious that if $L(\lambda|x_1, \ldots, x_n)$ is maximized at $\lambda = \widehat{\lambda}$, then $L(\theta|x_1, \ldots, x_n)$ is maximized at the point that corresponds to this point, namely at $\theta = 1/\widehat{\lambda}$. Therefore,

$$\widehat{\theta}_{ML} = 1/\widehat{\lambda}_{ML} = 1/\overline{X}.$$

The principle that the relation between parameters are transferred to their estimates is called the invariance, or plug-in, principle.

Theorem 4.2.9. *Suppose that $X_1, \ldots, X_n$ are observations from the distribution that depends on parameter θ. If $\widehat{\theta}_{ML} = \widehat{\theta}_{ML}(X_1, \ldots, X_n)$ is the maximum likelihood estimator for θ and $g(\cdot)$ is a one-to-one function, then $g(\widehat{\theta}_{ML})$ is the maximum likelihood for parameter $\psi := g(\theta)$, i.e.,*

$$\widehat{\psi}_{ML} = g(\widehat{\theta}_{ML})$$

Proof. We have the identity

$$L(\theta|y) = L(g^{-1}(\psi)|y),$$

and the expression on the right is the likelihood for ψ. If the MLE of ψ were $\psi^* \neq g(\widehat{\theta}_{MLE})$, then it would follow that

$$L(g^{-1}(\psi^*)) > L(g^{-1}(g(\widehat{\theta}_{MLE}))) = L(\widehat{\theta}_{MLE})$$

But this would contradict the fact that $\widehat{\theta}_{MLE}$ maximizes $L(\theta)$. □

- The invariance property actually holds for any function ψ of a parameter θ (and not only for the one-to-one functions), once we define appropriately, what do we mean by the maximum likelihood estimator of $\psi(\theta)$ in this case.
- A discussion and a proof can be seen in Casella and Berger (2002) [Math 502]

Example 4.2.10. For example, suppose you want to estimate the probability that a random variable $X > 2$, you know that it is a Poisson r.v. and have a data sample X_i, $i = 1, \ldots n$.

Note that

$$\begin{aligned}\Pr\{X > 2\} &= 1 - \Pr\{X = 0\} - \Pr\{X = 1\} - \Pr\{X = 2\} \\ &= 1 - e^{-\lambda} - e^{-\lambda}\frac{\lambda}{1!} - e^{-\lambda}\frac{\lambda^2}{2!}\end{aligned}$$

The invariance principle tells us that the ML estimator for $\Pr\{X > 2\}$ is simply

$$1 - e^{-\widehat{\lambda}} - e^{-\widehat{\lambda}}\frac{\widehat{\lambda}}{1!} - e^{-\widehat{\lambda}}\frac{\widehat{\lambda}^2}{2!},$$

where $\widehat{\lambda} = \widehat{\lambda}_{ML}$ is the maximum likelihood estimator for λ.

Since we know that $\widehat{\lambda}_{ML} = \overline{X}$, therefore the ML estimator for $\Pr\{X > 2\}$ is

$$1 - e^{-\overline{X}}\Big(1 + \frac{\overline{X}}{1!} + \frac{(\overline{X})^2}{2!}\Big).$$

Comparison of MM and MLE

1. MM estimator is usually easier to calculate.
2. In many cases MM estimator coincides with MLE.
3. MME might be not as efficient as MLE.
4. While MME is not as efficient as MLE, sometimes MM can be applied in situations when MLE is not available. (This happens when the likelihood function is not available but one can make some assumptions about the moments of random variables associated with the model.) In particular, a generalized method of moments (GMM) was developed in 1980s by an econometrist Lars Peter Hansen who received Nobel Prize in Economics in 2013 in part for this work.

Additional examples:

Example 4.2.11. Let $Y_1, \ldots, Y_n$ be taken from distribution with the following density:

$$f_Y(y) = \frac{1}{\theta} r y^{r-1} e^{-y^r/\theta} \mathbb{1}_{y>0}, \quad \text{where } \theta > 0 \text{ and } r \text{ is known.}$$

Find a sufficient statistic for θ. Find the MLE of θ. Is it MVUE?

- $L(\theta) = \prod_{i=1}^n \left\{\frac{1}{\theta} r y_i^{r-1} e^{-y_i^r/\theta}\right\} = \frac{1}{\theta^n} r^n (\prod_{i=1}^n y_i)^{r-1} e^{-\frac{1}{\theta}\sum_{i=1}^n y_i^r}$
- Clearly the sufficient statistic is $\sum_{i=1}^n Y_i^r$
- $L(\theta) = C \cdot \frac{1}{\theta^n} e^{-\frac{1}{\theta}\sum_{i=1}^n y_i^r}$ where C has nothing to do with θ.
- $\ell(\theta) = \log(C) - n\log(\theta) - \frac{1}{\theta}\sum_{i=1}^n y_i^r$

- $\ell'(\theta) = -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^n y_i^r$. Note that $\log(C)$ disappears.
- Set $\ell'(\theta) = 0 \Rightarrow \frac{n}{\theta} = \frac{1}{\theta^2}\sum_{i=1}^n y_i^r \Rightarrow \theta^* = \frac{1}{n}\sum_{i=1}^n y_i^r$
- So, $\frac{1}{n}\sum_{i=1}^n Y_i^r$ is the MLE for θ.
- MVUE??? We know that the estimator is a sufficient statistic. Need to check unbiasedness.
- $\mathbb{E}(\frac{1}{n}\sum_{i=1}^n Y_i^r) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}(Y_i^r)$
- Note that

$$\begin{aligned}\mathbb{E}(Y_i^r) &= \int_0^\infty \frac{1}{\theta} r y^{r-1} e^{-y^r/\theta} \cdot y^r dy \\ &= \int_0^\infty e^{-y^r/\theta} \cdot y^r d(y^r/\theta) \\ &\text{(Let } u = y^r/\theta\text{)} \\ &= \theta \int_0^\infty e^{-u} \cdot u du\end{aligned}$$

- One can either calculate the integral explicitly or note that e^{-u} is the density of the exponential distribution with parameter 1 and the integral $\int_0^\infty e^{-u} \cdot u\, du$ calculate its expectation, which, as we already know, equals 1. Hence, we have

$$\mathbb{E}(Y_i^r) = \theta$$

- Therefore, $\mathbb{E}(\frac{1}{n}\sum_{i=1}^n Y_i^r) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}(Y_i^r) = \frac{1}{n} n\theta = \theta$

It follows that the maximum likelihood estimator $\frac{1}{n}\sum_{i=1}^n \mathbb{E}(Y_i^r)$ is the MVUE for θ.

Another example:

Example 4.2.12. Consider the situation when we have two samples. One of them is $X_1, X_2, \ldots, X_m$ from normal distribution $N(\mu_1, \sigma^2)$. The other is $Y_1, Y_2, \ldots, Y_n$ from normal distribution $N(\mu_2, \sigma^2)$. Here we assumed that the variance in both distributions is the same. What is the ML estimators for the parameters μ_1, μ_2 and σ^2?

- Given the observations $x_1, \ldots, x_m, y_1, \ldots, y_n$, the likelihood for μ_1, μ_2, σ^2 is the product of all the densities (including the X's and the Y's)

$$\begin{aligned} L(\mu_1, \mu_2, \sigma^2) &= \prod_{i=1}^{m} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(x_i - \mu_1)^2}{2\sigma^2}\} \right] \times \\ &\quad \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(y_i - \mu_2)^2}{2\sigma^2}\} \right] \\ &= (2\pi\sigma^2)^{-\frac{m}{2}} \exp\left\{ -\frac{\sum_{i=1}^{m}(x_i - \mu_1)^2}{2\sigma^2} \right\} \times \\ &\quad (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{ -\frac{\sum_{i=1}^{n}(y_i - \mu_2)^2}{2\sigma^2} \right\} \end{aligned}$$

- So the log-likelihood is

$$\begin{aligned} \ell(\mu_1, \mu_2, \sigma^2) = &-\frac{m}{2}\log(2\pi) - \frac{m}{2}\log(\sigma^2) - \frac{\sum_{i=1}^{n}(x_i - \mu_1)^2}{2\sigma^2} \\ &- \frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{\sum_{i=1}^{n}(y_i - \mu_2)^2}{2\sigma^2} \end{aligned}$$

Let us use the notation $\theta = \sigma^2$. The partial derivatives of the log-likelihood function with respect to parameters are:

$$\ell_{\mu_1} = \frac{\overline{x} - \mu_1}{\theta/n}$$

$$\ell_{\mu_2} = \frac{\overline{y} - \mu_2}{\theta/n}$$

$$\ell_\theta = \left(-\frac{m}{2}\frac{1}{\theta} + \frac{\sum_{i=1}^{m}(x_i - \mu_1)^2}{2}\frac{1}{\theta^2} \right) + \left(-\frac{n}{2}\frac{1}{\theta} + \frac{\sum_{i=1}^{n}(y_i - \mu_2)^2}{2}\frac{1}{\theta^2} \right)$$

- Set all three to zero and get the solution.

$$\begin{aligned} &\widehat{\mu}_1 = \overline{X},\ \widehat{\mu}_2 = \overline{Y}, \\ &\widehat{\sigma^2} = \frac{\sum_{i=1}^{m}(x_i - \overline{X})^2 + \sum_{i=1}^{n}(y_i - \overline{Y})^2}{m+n} \end{aligned}$$

- Does $\widehat{(\sigma^2)}_{ML}$ look a bit familiar?

Exercise 4.2.13. $Y_1, Y_2, \ldots, Y_n$ is a sample of observations from $N(5, \theta)$ where the **variance** θ is unknown and is the parameter of interest:

$$f(y) = \frac{1}{\sqrt{2\pi\theta}} \exp[-\frac{(y-5)^2}{2\theta}].$$

(a). Find the sufficient statistic for θ.

(b). Find the Method of Moment Estimator (MME) $\widehat{\theta}^{MM}$ for θ.

(c). Find the Maximum Likelihood Estimator (MLE) $\widehat{\theta}^{ML}$ for θ

(d). Show directly (without using the general theorem that MM estimators are consistent) that $\widehat{\theta}^{MM}$ is consistent for θ.

(e). Show directly that $\widehat{\theta}^{ML}$ is consistent for θ.

(f). Prove that $\widehat{\theta}^{ML}$ is the minimal variance unbiased estimator (MVUE) for θ

Exercise 4.2.14. Let Y_1, Y_2, Y_3 be three i.i.d. observations from the distribution with density:

$$f_Y(y) = \frac{ye^{-y/\theta}}{\theta^2} \mathbb{1}_{y>0}$$

In a data sample these random variables are observed to be 120, 130 and 128, respectively.

- Find the ML extimator of θ
- Is the ML estimator unbiased in this model? Explain.
- What is the ML estimator for the variance of Y_1?

Exercise 4.2.15. Let $Y_1, \ldots, Y_n$ be from the distribution with density

$$f(y) = (\theta + 1)y^\theta, \ 0 < y < 1,$$

where $\theta > -1$. Find the MLE.

4.3 Cramer-Rao Lower Bound and large sample properties of MLE

In this section we learn about the Cramer-Rao lower bound on the variance of any *unbiased* estimator of θ. It is not possible do get smaller variance even if you use the MVUE. We also learn that in the limit, for $n \to \infty$, the maximum likelihood estimator achieves this bound. In this sense, the ML estimator is an *asymptotically* Minimal Variance Unbiased Estimator.

The idea behind the Cramer-Rao bound is that if the likelihood function is flat and does not depend on the parameter θ then it will be difficult to estimate the parameter from the data. The measure of the likelihood function flatness that the Cramer-Rao bound uses is the Fisher information. Essentially, it is the average squared sensitivity of the log-likelihood function to the parameter.

For the formal definition, let us define the score function $s(x, \theta)$ of a random variable X as $\log f(x, \theta)$ if X is continuous with probability density $f(x, \theta)$ and as $\log p(x, \theta)$ if it is discrete with probablity mass function $p(x, \theta)$.

We are talking here about a single variable X, and a log-likelihood function is the sum of the values of score function at x_i: $\ln L(\theta) = \sum_{i=1}^{n} s(x_i, \theta)$.

Definition 4.3.1. The Fisher information of a random variable X is defined as

$$I_X(\theta) = \mathbb{E}\Big[\Big(\frac{d}{d\theta} s(X, \theta)\Big)^2\Big]$$

Informally, the larger the Fisher information is, the more sensitive the log-likelihood function is with respect to parameter θ.

Example 4.3.2. Let us calculate the Fisher information for exponential random variable with mean θ. The density is $f(x, \theta) = \frac{1}{\theta} e^{-x/\theta}$, so the score is

$$s(x, \theta) = -\log \theta - x/\theta.$$

By definition,

$$\begin{aligned} I(\theta) &= \mathbb{E}\Big[\Big(\frac{d}{d\theta}(-\log\theta - X/\theta)\Big)^2\Big] \\ &= \mathbb{E}\Big[\Big(-1/\theta + X/\theta^2\Big)^2\Big] \\ &= \mathbb{E}\Big[1/\theta^2 - 2X/\theta^3 + X^2/\theta^4\Big] \end{aligned}$$

Recollect that the exponential distribution with parameter θ has mean θ and variance θ^2. Hence $\mathbb{E}X = \theta$ and $\mathbb{E}X^2 = \theta^2 + \theta^2 = 2\theta^2$. So after substitution we get

$$I(\theta) = 1/\theta^2 - 2\theta/\theta^3 + 2\theta^2/\theta^4 = 1/\theta^2.$$

In many cases, it is easier to calculate the Fisher information by using a different formula. Namely, under some regularity conditions, one has:

$$I_X(\theta) = -\mathbb{E}\Big[\frac{d^2}{d\theta^2}s(X,\theta)\Big]. \tag{4.1}$$

For example, the following result holds.

Lemma 4.3.3. *Suppose that X is a continuous random variable and the range of X (the set where the density is positive) does not depend on θ. Suppose also that the density is continuously differentiable in θ on the range. Then the equality (4.1) holds.*

This lemma can be proved by rather challenging integral manipulations. Identity 4.1 often gives a shorter path to calculate Fisher's information. In the previous example, this identity gives

$$\begin{aligned} I(\theta) &= -\mathbb{E}\Big[\frac{d^2}{d\theta^2}(-\log\theta - X/\theta)\Big] \\ &= -\mathbb{E}\Big[1/\theta^2 - 2X/\theta^3)\Big] = 1/\theta^2. \end{aligned}$$

We formulate the following result as a theorem although we do not specify the regularity conditions precisely. Check for them in graduate level textbooks.

Theorem 4.3.4 (Cramer-Rao bound)**.** *Let $X_1, \ldots, X_n$ be a sample of independent identically distributed observations from the distribution that depends on parameter θ. Under certain regularity conditions on the distribution, for every unbiased estimator $\widehat{\theta}$,*

$$\mathbb{V}\mathrm{ar}(\widehat{\theta}) \geq \frac{1}{nI_X(\theta)}$$

For example, the regularity conditions are satisfied if $X_1, \ldots, X_n$ are i.i.d. continuous random variables with density $f(x, \theta)$ provided that the support of the distribution (that is, the set of x where $f(x, \theta)$ is positive) does not depend on θ and that the density is continuously differentiable in θ.

Note 1: If we are able to find an unbiased estimator such that its variance equals the Cramer-Rao bound (and the regularity conditions hold), then this estimator is MVUE (minimal variance unbiased estimator).

Note 2: Sometimes the Cramer - Rao bound is not sharp. That is, sometimes the variance of the MVUE will be larger than the bound given by the Cramer - Rao inequality.

Note 3: The regularity conditions are often violated if the parameter involves the domain of the density. (Like estimation of the θ for the random variable uniformly distributed on $[0, \theta]$.) In this case, the Cramer-Rao bound is invalid: there can be an estimator with smaller variance than the bound predicts.

Example 4.3.5 (Exponential). Suppose that $X_1, \ldots, X_n$ are i.i.d. observations from the exponential distribution with parameter θ. We have calculated that the Fisher Information of this distribution is $I_X(\theta) = 1/\theta^2$. Hence, the Cramer-Rao inequality says that every unbiased estimator $\widehat{\theta}$ has variance $\geq \theta^2/n$. On the other hand the estimator $\widehat{\theta} = \overline{X}$ is unbiased and has variance $\mathbb{V}\mathrm{ar}(X_i)/n = \theta^2/n$. This fact means that this unbiased estimator attains the Cramer-Rao bound and so it is MVUE.

Example 4.3.6 (Uniform on $(0, \theta)$). Now suppose that $X_1, \ldots, X_n$ are i.i.d. observations from the uniform distribution on the interval $[0, \theta]$. First, let

us calculate the Fisher Information for this distribution. The density is $1/\theta$ and the score function is $-\log(\theta)$, so by definition:

$$I_X(\theta) = \mathbb{E}\Big[\Big(\frac{d}{d\theta}s(X,\theta)\Big)^2\Big] = \mathbb{E}\Big[\frac{1}{\theta^2}\Big] = \frac{1}{\theta^2}$$

(If we use identity (4.1), we get:

$$I_X(\theta) = -\mathbb{E}\Big[\frac{d^2}{d\theta^2}s(X,\theta)\Big] = -\mathbb{E}\Big[\frac{d^2}{d\theta^2}s(X,\theta)\Big] = \mathbb{E}\Big[\frac{1}{\theta^2}\Big] = \frac{1}{\theta^2},$$

which is the same. Note, however, that the conditions of Lemma 4.3.3 that justified (4.1) are not satisfied in this example, so we are lucky that (4.1) gives the correct result.) So the Cramer-Rao inequality predicts that every unbiased estimator should have variance $\geq \theta^2/n$. We have calculated earlier the expectation and variance of the ML estimator in this example which is $X_{(n)}$. This estimator is biased but we can correct its bias and consider the unbiased estimator $\widehat{\theta} = \frac{n+1}{n}X_{(n)}$. Its variance is

$$\mathbb{V}\mathrm{ar}(\widehat{\theta}) = \theta^2\frac{(n+1)^2}{n^2}\frac{n}{(n+2)(n+1)^2} = \theta^2\frac{1}{n(n+2)}$$

So this estimator clearly violates the Cramer-Rao bound. The reason is that the conditions of the Theorem 4.3.4 are not satisfied.

We will explain some ideas behind the proof of the Cramer-Rao bound below. Now let us turn to the asymptotic optimality of MLE. The main result here is that under some regularity conditions,

$$n\mathbb{V}\mathrm{ar}\Big(\widehat{\theta}_{ML}\Big) \to \frac{1}{I_X(\theta)}$$

as $n \to \infty$.

The point is that the MLE in the limit attains the Cramer-Rao bound. In this sense it is an asymptotically MVUE, or in other terminology it is asymptotically efficient.

Ideas of the proof of the Cramer-Rao bound

We are going to prove the bound for the case when X is continuous and its range does not depend on the parameter and when $n = 1$. The proof in the general case is difficult and you can find it in graduate level textbooks.

Lemma 4.3.7. *Assume the range of X does not depend on θ and the density is positive and continuously differentiable in θ. Then,*

$$\mathbb{E}\Big[\frac{d}{d\theta}s(X,\theta)\Big] = 0.$$

Proof. Note that by chain rule:

$$\frac{d}{d\theta}s(X,\theta) = \frac{\frac{d}{d\theta}f(x,\theta)}{f(x,\theta)}.$$

$$\begin{aligned}\mathbb{E}\Big[\frac{d}{d\theta}s(X,\theta)\Big] &= \int_a^b \frac{d}{d\theta}s(x,\theta)f(x,\theta)dx = \int_a^b \frac{d}{d\theta}f(x,\theta)dx \\ &= \frac{d}{d\theta}\int_a^b f(x,\theta)dx = \frac{d}{d\theta}1 = 0.\end{aligned}$$

□

Corollary of Lemma 4.3.7: $I(\theta) = \mathbb{V}\mathrm{ar}\big(\frac{d}{d\theta}s(X,\theta)\big)$.

Proof of the Cramer-Rao bound for $n = 1$. Let $\widehat{\theta}(X)$ be an unbiased estimator of θ based on just one datapoint X. Let us write $s'(\theta)$ instead of $\frac{d}{d\theta}s(X,\theta)$. By using Lemma 4.3.7 and the Cauchy - Schwarz inequality for covariance:

$$|\mathbb{E}\big(s'(\theta)\widehat{\theta}\big)| = |\mathrm{Cov}\big(s'(\theta),\widehat{\theta}\big)| \leq \sqrt{\mathbb{V}\mathrm{ar}(s'(\theta))\mathbb{V}\mathrm{ar}(\widehat{\theta})}.$$

Or,

$$\mathbb{V}\mathrm{ar}(\widehat{\theta}) \geq \frac{|\mathbb{E}\big(s'(\theta)\widehat{\theta}\big)|}{I(\theta)}$$

All this would hold if $\widehat{\theta}$ was biased. The next step is crucial.

$$\begin{aligned}\mathbb{E}\big(s'(\theta)\widehat{\theta}\big) &= \int_a^b \frac{d}{d\theta}f(x,\theta)\widehat{\theta}(x)dx = \frac{d}{d\theta}\int_a^b f(x,\theta)\widehat{\theta}(x)dx \\ &= \frac{d}{d\theta}\mathbb{E}\widehat{\theta} = \frac{d}{d\theta}\theta = 1.\end{aligned}$$

□

Example 4.3.8.

- $Y_i \sim \text{Bernoulli}(p)$ with PMF $p^{y_i}(1-p)^{1-y_i}$
- We know that $\hat{p} = \sum_{i=1}^{n} Y_i/n$ is the MLE for p
- Try to derive the Cramer-Rao Lower Bound for $\hat{p}$

$$\frac{1}{n\mathbb{E}\left[-\frac{\partial^2 \log f(Y|\theta)}{\partial\theta^2}\right]} = \frac{1}{n\mathbb{E}\left[-\frac{\partial^2 \log[p^Y(1-p)^{1-Y}]}{\partial p^2}\right]}$$

$$= \frac{1}{n\mathbb{E}\left[-\frac{\partial^2[Y\log(p)+(1-Y)\log(1-p)]}{\partial p^2}\right]} = \frac{1}{n\mathbb{E}\left[\frac{Y}{p^2}+\frac{1-Y}{(1-p)^2}\right]}$$

$$= \frac{1}{n\left[\frac{p}{p^2}+\frac{1-p}{(1-p)^2}\right]} = \frac{1}{n\left[\frac{1}{p}+\frac{1}{1-p}\right]} = \frac{p(1-p)}{n}$$

- We already know that $\mathbb{V}\text{ar}(\hat{p}) = \frac{p(1-p)}{n}$. So indeed, $\hat{p}$ is the MVUE.

Chapter 5

Hypothesis testing

5.1 Basic definitions

- Chapter 8: make statistical inference about the population (parameter) by
  - point estimators (with unbiasedness & low variance), and
  - confidence intervals.
- Chapter 9: different mathematical properties of point estimators, and how to find good estimators.
- Chapter 10:
  - Hypothesis tests: answer scientific questions using statistics
  - Different philosophy and goal.
  - Some connection with confidence interval.

Examples of "tests" in real life

- What people may want to know:
  1. Does smoking cause lung cancer?
  2. Is global warming real?

3. Are men more likely to run a stop sign than women?
4. Does chemotherapy really cure cancer?
5. Is a new medicine effective in increasing longevity?

- Beyond just scientific interest.
  - Business decisions, military actions, political strategies.

The basic philosophy of statistical testing is "Proof by Contradiction".

1. Identify the question of interest:
   DOES SMOKING LEAD TO LUNG CANCER?
2. Try to prove causality by assuming the opposite theory ("does not lead"), which is called null hypothesis, and showing that it leads to a contradiction.
3. Namely, if the data looks very improbable under the null hypothesis, then you can conclude that the data contradicts the null hypothesis, so it should be rejected and your theory should be accepted instead.
4. However, if the data does not looks very improbable under null hypothesis, then you cannot reject it and so you don't have enough evidence in support of your, alternative, point of view.

Terminology

- Hypothesis
  - A statement about a population, usually of the form that a parameter takes a particular numerical value (e.g. $\theta = 2$) or falls in a certain range of values (e.g. $\theta > 2$).
- Null Hypothesis H_0
  - The statement of **no effect**.
  - This is the statement that we will assume as true when we will try to show that it leads to improbable conclusions.

  - It is usually denoted by H_0 and it is usually very specific. For example it can state: "the treatment has no effect".

- Alternative Hypothesis H_a
  - The statement of **some effect**.
  - The statement that we actually want to confirm by showing that H_0 should be **rejected**.
  - Usually it is denoted by H_a; it can be specific like: the percentage of recoveries after a medicine was used increased by 10%. This is a point hypothesis. Or it can be less specific: the percent of recoveries after the treatment increased by at least 10%. This is called the composite alternative hypothesis.

- The hypotheses must be stated before collecting, viewing or analyzing the data.

Besides the null and alternative hypothesis, the statistical test is defined by a *test statistic* and a *rejection region*.

- Recall: a statistic is a function of random observations Y_i (the data) A statistic cannot be defined in terms of the unknown θ.

- A test statistic (TS) is a statistic, that is, a function of the data. Its intention is different from an estimator which is also a function a data. The test statistic should help us to answer the question "*how close is the data sample to what we would expect if the null hypothesis H_0 were true?*"

- If the null hypothesis is expressed in terms of a parameter of the model, for example if it has the form like $\theta = \theta_0$, where θ_0 is a specific value, then often you can use a test statistic in the form

$$TS = \frac{\widehat{\theta} - \theta_0}{\sigma_{\widehat{\theta}}},$$

  where $\widehat{\theta}$ is an estimator of θ.

- In this case, if $|TS|$ is large (so TS is far from 0) then it might indicate that the data is not compatible with the null hypothesis.

Definition 5.1.1. The reject region RR is a set of possible values of a test statistic $TS(X_1, \ldots, X_n)$ so that if value of TS for the observed data is in the rejection rejection RR, then we reject the null hypothesis H_0.

- Example: if we reject H_0 when $TS > 1$, then $RR = \{x : x > 1\}$, that is $RR = (1, \infty)$.
- Often, it is more convenient to write the RR as some inequality, such as $RR : TS < a$. This is equivalent to $RR = \{x : x < a\} = (-\infty, a)$.
- If the TS is not in the RR, then we fail to reject the null hypothesis H_0.
- Other commonly used terms: do not reject, do not have enough evidence to reject, etc.
- Next: How to find the reject region?

Example 5.1.2. Let $X_1, \ldots, X_n$ be distributed with Bernoulli distribution with parameter p.

- Null hypothesis: H_0: $p = 0.5$
- Alternative hypothesis H_a: $p > 0.5$.
- Test the hypothesis H_0 against alternative H_a.
- We can define our test statistic to be

$$TS = \frac{\hat{p} - 0.5}{SE(\hat{p})},$$

  where $\hat{p}$ is the sample proportion and $SE(\hat{p}) = \sqrt{0.5(1 - 0.5)/n}$
- A suitable reject region seems to be $TS > t$, since large TS indicates that p estimated from data is much larger then 0.5.
- But which t should we choose?

	Does not reject H_0	Reject H_0 in favor of H_a
H_0 is true	Correct decision	**Type I Error**
H_a is true	**Type II Error**	Correct decision

Types of errors that a test can make

(a) Type I Error or False Positive or False Discovery

- occurs if we reject H_0 and accept H_a when H_0 is in fact true.
- Probability of making a Type I error, also called (significance) **LEVEL of the test**:

  Level of the test, $\alpha = P(\text{Type I Error}|t) = P(\text{Reject } H_0|H_0, t)$.

(b) Type II Error or False Negagive

- Occurs if we fail to reject H_0 when H_0 is false. We failed to make a discovery
- Probability of making a Type II error. :

  $$\beta = P(\text{Type II Error}|t) = P(\text{Does not reject } H_0|\theta \in H_a, t).$$

- The quantity $1 - \beta$ is also called the **POWER of the test**

In the working example:

- As $t > 0$ increases, harder to reject H_0. Then $\alpha \downarrow$, $\beta \uparrow$
- As $t > 0$ decreases, easier to reject H_0. Then $\alpha \uparrow$, $\beta \downarrow$
- α and β are always inversely related;
  - It is impossible to minimize both at the same time.
- In scientific practice and in drug development, researchers typically consider a Type I error ("False Discovery") more serious error than a Type II error ("Failure to make a discovery"). In medical application, such as testing for a decease, however, it is often more important to minimize Type II error. Here we proceed with the assumption that Type I error is more important.

  - In this case, a value for α is chosen before initiating a hypothesis test.
  - Common values for α are 0.01, 0.05 and 0.10;
  - Choose the rejection region so that

$$\mathbb{P}(\text{Type I Error}) = \mathbb{P}(t \in RR|H_0) = \alpha$$

- If $\alpha = 0.05$, this choice of the rejection region means that in 5% of the data samples from a population where H_0 is actually true, the test will reject the H_0.

Type I against Type II errors

- What is the consequence of a Type I error?
  - Conclude that a drug is effective when in fact that it is not.
  - Conclude that a foreign policy is working when in fact that it is not.
  - Ultimately: huge amount of money spent for nothing
- What is the consequence of a Type II error?
  - Conclude that a drug is ineffective when in fact it is a good drug.
  - Conclude that a potentially working foreign policy is not useful.
  - Ultimately: Lost opportunity

Eventually, the choice of balance between type I and type II error depend on a cost-benefit analysis, which is outside of the area of statistics.

Summary Design of the test:

- Set up H_0
- Set up H_a
- Define a reasonable statistic TS.
- Figure out the distribution of TS under H_0

- Choose a small significance level α (like 5%) and find a reject region (RR) so that $\mathbb{P}(\text{make a type I error}) = \mathbb{P}(TS \in RR | H_0 \text{ is true}) = \alpha$.

Application of the test: If the observed value of the TS is in the RR, then reject H_0 in favor of H_a; otherwise, decide that the test fails to reject H_0, and conclude that there is no sufficient evidence at the significance level α that H_a is true.

5.2 Calculating the Level and Power of a Test

5.2.1 Basic examples

Given a test statistic and a reject region (RR) of a test, how do we find the probabilities of errors α and β?

- Type I Error:
  - Occurs if we reject H_0 when H_0 is true.
  - Probability of making a Type I error:

$$\begin{aligned}\alpha &= P(\text{Type I Error}) = P(\text{rejecting } H_0 \mid H_0 \text{ is true}) \\ &= P(TS \in RR \mid H_0 \text{ is true}).\end{aligned}$$

- Type II Error:
  - Occurs if we fail to reject H_0 when H_0 is false.
  - Probability of making a Type II error:

$$\begin{aligned}\beta &= P(\text{Type II Error}) = P(\text{fail to reject } H_0 \mid \theta \in H_a) \\ &= P(TS \notin RR \mid \theta \in H_a).\end{aligned}$$

Example 5.2.1. An experimenter has prepared a drug dosage level that she claims will induce sleep for 80% of people suffering from insomnia. After examining the dosage, we feel that her claims regarding the effectiveness of the dosage are inflated. In an attempt to disprove her claim, we administer her prescribed dosage to 20 insomniacs and we observe Y , the number for whom the drug dose induces sleep. We wish to test the hypothesis

$H_0 : p = .8$ versus the alternative, $H_a : p < .8$. Assume that the rejection region $\{y \leq 12\}$ is used.

(a) What is the probability of type I error (level of the test) α?

(b) What is the probability of type II error β if $H_a : p = 0.6$?

(c) What is the probability of type II error β if $H_a : p = 0.4$?

(d) If we want the size of the test $\alpha \approx 0.01$ how should we choose the threshold r in the rejection region $RR = \{y \leq r\}$?

(a) In this example Y is our test statistic (the complete data consists of observations for each insomniac). This statistic is distributed according to the binomial distribution with parameters $n = 20$ and p. If we assume H_0 then $p = 0.8$. Then we need to calculate

$$\alpha = \mathbb{P}(Y \in RR | H_0 : p = 0.8) = \mathbb{P}(Y \leq 12 | H_0 : p = 0.8).$$

We can do it using tables or issuing R command pbinom(12, 20, 0.8). which gives us the result $\alpha = 0.03214266$. This is the significance level (or size) of this test.

(b) Under the alternative hypothesis $H_a : p = 0.6$, we have

$$\begin{aligned} \beta &= \mathbb{P}(Y \notin RR | H_0 : p = 0.8) = \mathbb{P}(Y > 12 | H_a : p = 0.6) \\ &= 1 - \mathbb{P}(Y \leq 12 | H_a : p = 0.6). \end{aligned}$$

We can calculate it as $\beta = 1 - pbinom(12, 20, 0.6) = 0.4158929$. This is a rather large probability of error. We can also say that the power of the test $1 - \beta \approx 58\%$ is small under this alternative hypothesis.

(c) If $H_a : p = 0.4$, then $\beta = 1 - pbinom(12, 20, 0.4) = 0.02102893$, and the power is $1 - \beta \approx 98\%$. Here both the probabilities of type I and type II errors are small because in this case it is easy to detect from the data whether a null hypothesis or an alternative is true. The probability that we encounter the data which would be likely under both alternatives is small.

(d) Now if we want to make $\alpha = 0.01$, then we need to find r such that $\mathbb{P}(Y \leq r | H_0 : p = 0.8) = 0.01$. Unfortunately, there is no r such that this equality is satisfied exactly. However, we can solve it approximately by using the R command $qbinom$. If we issue thhe command $qbinom(0.01, size = 20, p = 0.8)$, it gives us $r = 12$, since by definition it produces the smallest r such that $\mathbb{P}(Y \leq r)$ is ≥ 0.01. However, we know that $\mathbb{P}(Y \leq 12) = pbinom(12, 20, 0.8) = 0.03214266$ and this is not satisfactory if we want to ensure that the probability of type I error is smaller than 0.01. For this reason we should choose $r = 11$. In fact, for this choice we have $\alpha = \mathbb{P}(Y \leq 11) = pbinom(11, 20, 0.8) = 0.009981786$ which is very close to 0.01.

In the previous example we were given the test statistic and the rejection region. How can we choose them in a typical exam? Here is one method which is useful if we are interested in testing a statement about a parameter θ, and are given a desired probability of type I error (i.e., the significance level of the test α).

- The null hypothesis $H_0 : \theta = \theta_0$
- The alternative hypothesis could be one of the following
  - $H_a : \theta > \theta_0$ (one-sided test)
  - $H_a : \theta < \theta_0$ (one-sided test)
  - $H_a : \theta \neq \theta_0$ (two-sided test)
- Using the sample data to find an estimator of θ, denote it by $\widehat{\theta}$;
- Define the test statistic
$$TS = \frac{\widehat{\theta} - \theta_0}{\sigma_{\widehat{\theta}}}.$$
- For the reject region (RR):
  - $\{TS > t\}$ (one-sided test)
  - $\{TS < t\}$ (one-sided test)

– $\{|TS| > t\}$ (two-sided test)

- Cutoff t is chosen such that

$$P(TS \text{ is in RR} \mid H_0 \text{ is true}) = P(TS \text{ is in RR } |\theta = \theta_0) = \alpha$$

Example 5.2.2. A machine in a factory must be repaired if it produces more than 10% defectives among the large lot of items that it produces in a day. A random sample of 100 items from the day's production contains 15 defectives, and the supervisor says that the machine must be repaired. Does the sample evidence support his decision? Use a test with level .01.

- The null hypothesis $H_0 : p = p_0$ (where $p_0 = 0.1$ here.)
- The alternative hypothesis $H_a : p > p_0$;
- An estimator of p is $\widehat{p}$ the sample proportion;
- Define the test statistic

$$TS = \frac{\widehat{p} - p_0}{\sigma_{\widehat{p}}} = \frac{\widehat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}.$$

  Note that we use $p_0 = 0.1$ to calculate $\sigma_{\widehat{p}}$, not the estimator of $\widehat{p} = 0.15$. This is because we are aiming to use fully the assumptions of the null hypothesis to calculate the test statistic.

- The reject region (RR): $\{TS > t\}$; with the cutoff t chosen such that $P(TS > t \big| p = p_0) = \alpha$.

Under the assumption H_0 is true, TS is approximately $N(0, 1)$, because the sample size is large ($n = 100$). So the equation becomes a statement about a normal random variable.

$$P(TS > t \big| p = p_0) = \alpha \Rightarrow P(Z > t) = \alpha,$$

where Z is a normal random variable. We know that it holds when $t = z_\alpha$. So the reject region RR becomes

$$RR: \quad \{TS > z_\alpha\};$$

The problem asks for level $\alpha = 0.01$, so $z_{0.01} = 2.33$, and the reject region is

$$RR: \quad \{TS > 2.33\};$$

By using the data provided in this problem, we calculate the observed value of the test statistic TS as

$$ts = \frac{\widehat{p} - p_0}{\sigma_{\widehat{p}}} = \frac{0.15 - 0.10}{\sqrt{0.1(1-0.1)/100}} = \frac{5}{3} = 1.667.$$

Since ts is NOT in the reject region, we fail to reject H_0 at the level $\alpha = 0.01$. and come to the conclusion that there is NO sufficient evidence to support the statement that the machine must be repaired, at the significance level $\alpha = 0.01$.

Note that if we used a different α, say $\alpha = 0.05$, then $z_{0.05} = 1.645$, and the reject region would be

$$RR: \quad \{TS > 1.645\};$$

Then, ts would be in the reject region, and we would reject H_0 at the level $\alpha = 0.05$. The conclusion would become: at the significance level $\alpha = 0.05$, there is sufficient evidence to support the statement that the machine must be repaired.

So, the decision of the hypothesis test depends on the value of α – the level of tolerance for the type I error. The report about the decision should always specify the value of α.

Not that the decision of a hypothesis test has a random nature! It depends on the realized data. In particular, if H_0 is true, we incorrectly reject it with probability α.

Now let us look at the calculation of the probability of type II error in the situation when we have a large sample and an estimator of the parameter which is distributed approximately normally. Let us consider the same example as before.

Example 5.2.3. Let $X_1, \ldots, X_n$ be distributed according to the Bernoulli distribution with parameter p.

(a) Null hypothesis: H_0: $p = 0.1$

(b) Alternative hypothesis H_a: $p > 0.1$.

(c) Test the hypothesis H_0 against alternative H_a at level α.

(d) What can be said about the probability of type II error and the power of this test?

As in the previous example, we define the test statistic as

$$TS = \frac{\hat{p} - 0.1}{SE(\hat{p})},$$

where $\hat{p}$ is the sample proportion and $SE(\hat{p}) = \sqrt{0.1(1-0.1)/n}$ The rejection region is $TS > z_\alpha$, where α is the pre-specified level (=size) of the test, which is the probability of type I error.

Now, what about the probability of type II error for this test? In order to calculate this probability, we need to have a specific alternative hypothesis H_a about the parameter p. Suppose, for example that $H_a : p = 0.15$. This is a natural choice since we observed 15 defective machine out of 100. If we assume that the alternative hypothesis is true, then we know that for large n, the quantity

$$Z = \frac{\widehat{p} - 0.15}{\sqrt{0.15(1-0.15)/n}}$$

is distributed as a standard normal random variable. Therefore,

$$\widehat{p} = 0.15 + \big(\sqrt{0.15(1-0.15)/n}\big) Z,$$

and we re-write the test statistic as

$$TS = \frac{\widehat{p} - 0.1}{\sqrt{0.1(1-0.1)/n}} = \frac{0.15 - 0.1 + Z\sqrt{0.15(1-0.15)/n}}{\sqrt{0.1(1-0.1)/n}}$$

So the probability of type II error for the test with level α is

$$\begin{aligned}
\beta = \mathbb{P}(TS \leq z_\alpha | H_a) &= \mathbb{P}\Big(\frac{0.05 + Z\sqrt{0.15(1-0.15)/n}}{\sqrt{0.1(1-0.1)/n}} \leq z_\alpha\Big) \\
&= \mathbb{P}\Big(Z\sqrt{0.15(1-0.15)/n} \leq -0.05 + z_\alpha\sqrt{0.1(1-0.1)/n}\Big) \\
&= \mathbb{P}\Big(Z \leq \frac{-0.05 + z_\alpha\sqrt{0.1(1-0.1)/n}}{\sqrt{0.15(1-0.15)/n}}\Big)
\end{aligned}$$

and this quantity is easy to evaluate by using software or by referring to tables. If we use $\alpha = 0.01$ and $n = 100$, then $z_\alpha = 2.33$ we calculate that

$$\frac{-0.05 + z_\alpha \sqrt{0.1(1-0.1)/n}}{\sqrt{0.15(1-0.15)/n}} = 0.5573$$

and

$$\mathbb{P}(Z \leq 0.5573) = 0.71.$$

So the probability that we make an error of the type II is rather large, 71%

Note that the *power* of the test is by definition $1 - \beta$, so we have a method to calculate both the probability of type II error and the power of the test. In this example the power of the test is $1 - 0.71 = 29\%$.

It is important to note that β (and the power) depends on the value of the parameter under the alternative hypothesis. For example, if we changed our alternative hypothesis to $H_a : p = 0.2$, then the probability of type II error would be equal to

$$\beta = \mathbb{P}\Big(Z \leq \frac{-0.1 + z_\alpha \sqrt{0.1(1-0.1)/n}}{\sqrt{0.2(1-0.2)/n}}\Big) = \mathbb{P}(Z \leq -0.7525) = 22.6\%$$

This is certainly much better outcome. Under this alternative hypothesis the test has much more statistical power. The power equals $1 - 0.226 = 77.4\%$.

Example 5.2.4 (Covid-19 in New York and California). The total number of cases of Covid-19 in New York State is $\approx 203,123$ with number of total deaths 10,834 as of April 14. The corresponding numbers for California are $25,536$ and 782. Test the hypothesis that the mortality rate in New York is higher than the mortality rate in California.

- The null hypothesis $H_0 : \theta = \theta_0$ (where $\theta = p_1 - p_2$ and $\theta_0 = 0$.)
- The alternative hypothesis $H_a : \theta > \theta_0$;
- An estimator of $\theta = p_1 - p_2$ is $\widehat{\theta} = \widehat{p}_1 - \widehat{p}_2$, the difference in sample proportion;

- Find the test statistic

$$TS = \frac{\widehat{\theta} - \theta_0}{\sigma_{\widehat{\theta}}} = \frac{\widehat{p}_1 - \widehat{p}_2 - 0}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \approx ?.$$

- Form the reject region (RR): $TS > z_\alpha$.

An additional difficulty here is that the null hypothesis does not specify the exact values of p_1 and p_2. It only says that $p_1 = p_2$. For this reason, we need to estimate p_1 and p_2. We use the **"pooled sample proportion"**, suggested by the fact that H_0 claims that $p_1 = p_2$.

$$\tilde{p} := \frac{Y_1 + Y_2}{n_1 + n_2}$$

This is the best guess about p_1 and p_2 we can obtain when $p_1 = p_2$ (that is, under the assumption that H_0 is true.)

Using the data provided in this problem, we can calculate:

$$\begin{aligned} \widehat{p}_1 &= \frac{10834}{203123} = 0.05333, \\ \widehat{p}_2 &= \frac{782}{25536} = 0.03062, \\ \widehat{p}_1 - \widehat{p}_2 &= 0.02271 \end{aligned}$$

$$\begin{aligned} \tilde{p} &= \frac{10834 + 782}{203123 + 25536} = 0.05080, \\ \sigma_{\widehat{p}_1 - \widehat{p}_2} &= \sqrt{\tilde{p}(1-\tilde{p})\Big(\frac{1}{n_1} + \frac{1}{n_2}\Big)} = 0.001457, \end{aligned}$$

and the observed value of the test statistic TS is

$$ts = \frac{\widehat{\theta} - \theta_0}{\sigma_{\widehat{\theta}}} = \frac{0.05080 - 0}{0.001457} = 15.5789$$

So, in this case it is obvious that the null hypothesis can be rejected at $\alpha = 0.01$. The data give strong support to the hypothesis that the mortality rate in New York is higher than that in California.

The previous two examples were about testing hypotheses about population proportions.

Now let us look at the hypotheses about population means. We still maintain the assumption that the sample size is large and therefore we can rely on the normality of the parameter estimator distribution.

Example 5.2.5. A random sample of 37 second graders who participated in sports had manual dexterity scores with mean 32.19 and standard deviation 4.34. An independent sample of 37 second graders who did not participate in sports had manual dexterity scores with mean 31.68 and standard deviation 4.56.

a. Test to see whether sufficient evidence exists to indicate that second graders who participate in sports have a higher mean dexterity score. Use $\alpha = .05$.

b. For the rejection region used in part (a), calculate β when $\mu_1 - \mu_2 = 3$.

The null hypothesis is $H_0 : \mu_1 = \mu_2$ and the alternative is $H_a : \mu_1 > \mu_2$.

A suitable estimator for $\theta = \mu_1 - \mu_2$ is $\widehat{\theta} = \overline{X} - \overline{Y}$, the difference of the sample means. Its variance is

$$\sigma_{\widehat{\theta}} = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}.$$

Since we do not know the exact values of σ_1^2 and σ_2^2, we will use estimates for these variances. Then, the test statistic is

$$\begin{aligned} TS &= \frac{\widehat{\theta} - \theta_0}{\widehat{\sigma}_{\widehat{\theta}}} = \frac{\overline{X} - \overline{Y} - 0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \\ &= \frac{32.19 - 31.68}{\sqrt{4.34^2/37 + 4.56^2/37}} = 0.4928 \end{aligned}$$

Since the samples are relatively large ($n_1 = n_2 > 30$), the test statistic is distributed as a standard normal random variable. Since $\alpha = 0.05$ and $TS \leq z_{0.05} = 1.645$, the test statistic is not in the rejection region and we are not able to reject the null hypothesis. The data does not give enough evidence to indicate that second graders who participate in sports have a higher mean dexterity score.

Now let us consider the second question. What is β if $\mu_1 - \mu_2 = 3$?

In this case, we know that

$$Z = \frac{\overline{X} - \overline{Y} - 3}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

is approximately standard normal random variable. What we need to calculate is the probability that we do not reject the null hypothesis, that is $\mathbb{P}(TS \leq z_\alpha)$. So, we need to express TS in terms of Z:

$$TS = \frac{\overline{X} - \overline{Y}}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} = \frac{\overline{X} - \overline{Y} - 3 + 3}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} = Z + \frac{3}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

Then, the desired probability is

$$\begin{aligned}\beta = \mathbb{P}(TS \leq z_\alpha) &= \mathbb{P}\Big[Z + \frac{3}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \leq z_\alpha\Big] \\ &= \mathbb{P}\Big[Z \leq z_\alpha - \frac{3}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}\Big]\end{aligned}$$

After plugging in numbers, we get

$$\begin{aligned}\beta &= \mathbb{P}\Big[Z \leq 1.645 - \frac{3}{\sqrt{(4.34^2 + 4.56^2)/37}}\Big] \\ &= \mathbb{P}\Big[Z \leq -1.2538] = pnorm(-1.2538) = 0.105...\end{aligned}$$

So, $\beta = 10.5\%$ and the power of this test is $1 - \beta = 89.5\%$.

More generally, if we use the test statistic

$$TS = \frac{\widehat{\theta} - \theta_0}{\widehat{\sigma}_{\widehat{\theta}}},$$

where the sample size is large, the estimator $\widehat{\theta}$ has an approximately normal distribution, and the standard deviation of the estimator $\widehat{\theta}$ is estimated from the data (not calculated on the basis of the hypothesis), then we can write simple formulas for β.

If the alternative hypothesis is $\theta_a > \theta_0$, and we use the rejection region $TS > z_\alpha$, then

$$\beta = \mathbb{P}\Big[Z \leq z_\alpha - \frac{\theta_a - \theta_0}{\widehat{\sigma}_{\widehat{\theta}}}\Big]$$

If the alternative hypothesis is $\theta_a < \theta_0$ and the rejection region is $TS < -z_\alpha$, then

$$\beta = \mathbb{P}\Big[Z \geq -z_\alpha - \frac{\theta_a - \theta_0}{\widehat{\sigma}_{\widehat{\theta}}}\Big],$$

which can also be written as

$$\beta = \mathbb{P}\Big[Z \leq z_\alpha - \frac{\theta_0 - \theta_a}{\widehat{\sigma}_{\widehat{\theta}}}\Big]$$

by the symmetry of the distribution of the normal random variable.

Finally, if the alternative hypothesis is $\theta_a \neq \theta_0$ and we decided to use the symmetric rejection region $|TS| \geq z_{\alpha/2}$, then

$$\beta = \mathbb{P}\Big[Z \leq z_{\alpha/2} - \frac{\theta_a - \theta_0}{\widehat{\sigma}_{\widehat{\theta}}}\Big] - \mathbb{P}\Big[Z \leq -z_{\alpha/2} - \frac{\theta_a - \theta_0}{\widehat{\sigma}_{\widehat{\theta}}}\Big],$$

What happens if $\theta_a = \theta_0$, of for example, if $\theta_a = \theta_0 + \varepsilon$, where ε is very small? This is the situation when the alternative hypothesis is barely distinguishable from the null hypothesis. It is easy to see that in this case $\beta = 1 - \alpha$, and the power $= \alpha$. This is the worst case scenario for the test and we conclude that the power of a test cannot drop down below its size.

Now what happens if $|\theta_a - \theta_0|$ becomes larger. In the case of one-sided hypotheses, it is easy to see from formulas that β declines. In fact, it is possible to check that this also holds for the two-sided hypothesis. The more the alternative differs from the null hypothesis, the less is the probability of type II error β (for a fixed level α).

5.2.2 Additional examples

Example 5.2.6 (Shear strength of soils). Shear strength of soils is a quantity important in civil engineering. Shear strength measurements derived from unconfined compression tests for two types of soils gave the results shown in the following table (measurements in tons per square foot). Do the soils appear to differ with respect to average shear strength, at the 1% significance level?

Soil Type I	Soil Type II
$n_1 = 30$	$n_2 = 35$
$\overline{y}_1 = 1.65$	$\overline{y}_2 = 1.43$
$s_1 = 0.26$	$s_2 = 0.22$

The null hypothesis $H_0 : \theta = \theta_0$ (where $\theta = \mu_1 - \mu_2$ and $\theta_0 = 0$.) This is simply a different formulation of the hypothesis $H_0 : \mu_1 = \mu_2$.) The alternative hypothesis $H_a : \theta \neq \theta_0$.

An estimator of $\theta = \mu_1 - \mu_2$ is $\widehat{\theta} = \overline{Y}_1 - \overline{Y}_2$, the difference in sample mean. So we take the test statistic

$$TS = \frac{\widehat{\theta} - \theta_0}{\sigma_{\widehat{\theta}}} = \frac{\widehat{\theta} - \theta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \approx \frac{\widehat{\theta} - \theta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

The rejection region (RR): $\{|TS| > t\}$ and choose the cutoff t so that

$$P(|TS| > t \big| \theta = \theta_0) = \alpha$$

When H_0 is true, the TS is approximately $N(0, 1)$, and we can take $t = z_{\alpha/2}$.

$$RR: \quad \{|TS| > z_{\alpha/2}\};$$

For $\alpha = 0.01$, $z_{0.01/2} = 2.575$, which gives $RR: \quad \{|TS| > 2.575\}$.

By using the data,

$$ts = \frac{\widehat{\theta} - \theta_0}{\sigma_{\widehat{\theta}}} = \frac{1.65 - 1.43 - 0}{\sqrt{\frac{0.26^2}{30} + \frac{0.22^2}{35}}} = 3.65.$$

Since ts in the reject region, our decision: reject H_0 at the level $\alpha = 0.01$. We conclude that at the significance level $\alpha = 0.01$, there is sufficient evidence that the soils appear to differ with respect to average shear strength.

Example 5.2.7. A political researcher believes that the fraction p_1 of Republicans strongly in favor of the death penalty is greater than the fraction p_2 of Democrats strongly in favor of the death penalty. He acquired independent random samples of 200 Republicans and 200 Democrats and found 46 Republicans and 34 Democrats strongly favoring the death penalty. Does this evidence provide statistical support for the researcher's belief? Use $\alpha = .05$.

Using the data provided in this problem, we can calculate:

$$\tilde{p} = \frac{Y_1 + Y_2}{n_1 + n_2} = \frac{46 + 34}{400} = 0.2,$$
$$\sigma_{\widehat{p}_1 - \widehat{p}_2} = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n_1} + \frac{\tilde{p}(1-\tilde{p})}{n_2}} = \sqrt{\frac{2 \times 0.2 \times 0.8}{200}} = 0.04,$$

and the observed value of the test statistic TS is

$$ts = \frac{\widehat{\theta} - \theta_0}{\sigma_{\widehat{\theta}}} = \frac{(46/200 - 34/200) - 0}{0.04} = 1.5$$

What should be the conclusion?

5.3 Determining the sample size

Now, consider Example 5.2.5 again. Suppose that we are not satisfied that the probability of type II error β is around 10%. What should the sample size that would give $\beta = 5\%$?

Recall that the formula for β is

$$\beta = \mathbb{P}\Big[Z \leq z_\alpha - \frac{3}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}\Big].$$

We can use the symmetry of the normal distribution and write it as

$$\beta = \mathbb{P}\Big[Z > -z_\alpha + \frac{3}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}\Big],$$

which can be re-written as

$$z_\beta = -z_\alpha + \frac{3}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

If we assume that $n_1 = n_2$, then this is the same as

$$z_\beta + z_\alpha = \frac{3\sqrt{n}}{\sqrt{s_1^2 + s_2^2}},$$

and the formula for the appropriate sample size is

$$n = (s_1^2 + s_2^2)\Big[\frac{z_\beta + z_\alpha}{3}\Big]^2,$$

Note that 3 here represents $\theta_a - \theta_0$, so the general formula is

$$n = (s_1^2 + s_2^2)\Big[\frac{z_\beta + z_\alpha}{\theta_a - \theta_0}\Big]^2,$$

This formula works for both right-tailed and left-tailed one-sided tests. However, there is no simple formula for two-sided hypotheses.

By plugging in the numbers from the example, we get

$$n = (4.34^2 + 4.56^2)\Big[\frac{1.645 + 1.645}{3}\Big]^2 \approx 47.66$$

Since we cannot use fraction as a sample size, we conclude that sample size $n = 48$ would be sufficient.

5.4 Relation with confidence intervals

Suppose we are still working with large size samples and we know that the estimator $\widehat{\theta}$ for a parameter θ has a normal distribution. In fact let us assume that we calculated the confidence interval, with the confidence level α. For concreteness, let us focus on one-sided lower bound confidence interval

$$CI = (\widehat{\theta} - z_\alpha \widehat{\sigma}_{\widehat{\theta}}, \infty).$$

Does it help us with hypothesis testing? Well, the confidence interval says that the true value of the parameter is likely to be larger than $\widehat{\theta} - z_\alpha \widehat{\sigma}_{\widehat{\theta}}$. So if we test the null hypothesis $H_0 \colon \theta = \theta_0$, and it happens that θ_0 is outside of the confidence interval, that is, if $\theta_0 < \widehat{\theta} - z_\alpha \widehat{\sigma}_{\widehat{\theta}}$, then we should reject the null hypothesis.

The only provision here is that the CI should be in agreement with the alternative hypothesis, that is, the alternative hypothesis should be $H_a : \theta > \theta_0$.

If our alternative hypothesis is that $\theta < \theta_0$, the it is more appropriate to consider the upper bound confidence interval:

$$CI = (-\infty, \widehat{\theta} + z_\alpha \widehat{\sigma}_{\widehat{\theta}}).$$

This confidence interval tells us that the true value of the parameter is likely to be large than $\widehat{\theta} + z_\alpha \widehat{\sigma}_{\widehat{\theta}}$, so if θ_0 is greater then this quantity we should reject the null hypothesis.

Similarly, if we use a two-sided alternative hypothesis $H_a : \theta \neq \theta_0$, then it is appropriate to use the two-sided confidence interval

$$CI = (\widehat{\theta} - z_{\alpha/2}\widehat{\sigma}_{\widehat{\theta}}, \widehat{\theta} + z_{\alpha/2}\widehat{\sigma}_{\widehat{\theta}}),$$

and reject the null hypothesis $H_0 : \theta = \theta_0$ if θ_0 is outside of the confidence interval.

On the formal level, consider, say, the first case, when the alternative is $H_a : \theta > \theta_0$.

Then the rejection region is

$$TS = \frac{\widehat{\theta} - \theta_0}{\widehat{\sigma}_{\widehat{\theta}}} > z_\alpha.$$

But this condition can be re-written as

$$\theta_0 < \widehat{\theta} - z_\alpha \widehat{\sigma}_{\widehat{\theta}},$$

and this is exactly the condition that θ_0 is outside of the low-bound confidence interval, as we claimed above. The other two cases can be done similarly.

5.5 p-values

The p-value of a test is useful if one wants to report how strongly the evidence in the data speaks against the null hypothesis.

Recall that we saw several times the situation when for $\alpha = 0.01$ we could not reject the null hypothesis, the evidence was not strong enough, but for $\alpha = 0.05$, we could reject the null. (This is because when $\alpha = 0.05$ we could allow to make type I error more frequently.)

For any data sample if we consider very large α then the test statistic is likely to land in the rejection region, which is very wide in this case and we are likely to reject the test. However, as we gradually decrease α, we

become more conservative, the rejection region shrinks, and at some point we switch from rejecting H_0 for this data sample to saying that there is not enough evidence in the data to support the rejection. This point is called the p-value of the test.

Note especially that unlike the level and the power of the test, the p-value depends both on the test (that is, on the way to calculate the test statistic and the rejection region) and on the data. If the data sample looks more unlikely for the null hypothesis than another sample, that is, if it has a larger test statistic, then the switch from rejection to non-rejection happens later, for smaller α, and p - value for such data sample is *smaller*!

Definition 5.5.1. The *p-value* is the smallest significance level α at which the observed data indicates that H_0 should be rejected.

While definition above is easy to use, it is a bit difficult to grasp or to explain to a client who does not know what is the significance level of a test. In this case, the following equivalent definition might be useful.

Definition 5.5.2. The p-value is the probability, – calculated assuming that the null hypothesis is true, – of obtaining a value of the test statistic, which is at least as contradictory to H_0 as the value calculated from the available sample.

It is very easy to calculate the p-value. We just set the threshold in the rejection region equal to the observed value of the test statistic and calculate the probability of this rejection region under the null hypothesis.

Say, let the test have the rejection region $RR : \{TS > t\}$ and let ts be the observed value of the test statistic. Then the *p-value* is $\Pr\{TS > ts|H_0\}$.

In practice, for large sample tests it often boils down to calculating the cumulative function of the standard normal distribution at the test statistic value ts.

Benefits of the p-value:

- It is a universal measure of the strength of the evidence.
- It describes how extreme the data would be if the H_0 were true.

- It answers the question: "Assuming that the null is true, what is the chance of observing a sample like this, or even worse?"

Example 5.5.3. Urban storm water can be contaminated by many sources, including discarded batteries. When ruptured, these batteries release metals of environmental significance. The paper "Urban Battery Litter" (J. Environ. Engr., 2009: 46–57) presented summary data for characteristics of a variety of batteries found in urban areas around Cleveland. A sample of 51 Panasonic AAA batteries gave a sample mean zinc mass of 2.06 g. and a sample standard deviation of .141 g. Does this data provide compelling evidence for concluding that the population mean zinc mass exceeds 2.0 g.?

With m denoting the true average zinc mass for such batteries, the relevant hypotheses are $H_0 : m = 2.0$ versus $H_a : m > 2.0$. The sample size is large enough so that a z-test can be used without making any specific assumption about the shape of the population distribution. The test statistic value is

$$z = \frac{\overline{x} - 2.0}{s/\sqrt{n}} = \frac{2.06 - 2.0}{.141/\sqrt{51}} = 3.04$$

So, we calculate the p-value:

$$p - \text{value} = \mathbb{P}(Z > 3.04) = 1 - pnorm(3.04) = 0.118\%$$

This means that the null hypothesis would be rejected by tests with $\alpha = 5\%$, $\alpha = 1\%$, and even with $\alpha = 0.2\%$, although it could not be rejected at the level of $\alpha = 0.1\%$. We would conclude that the sample appears to highly contradictory to the null hypothesis, and so there is a compelling evidence that the the population mean zinc mass exceeds 2.0 g.

p-values for large sample tests (aka z-tests)

1. The parameter θ is one of the following: μ, p, $\mu_1 - \mu_2$ and $p_1 - p_2$;
2. The sample size n is large enough.
3. The null hypothesis is $H_0 : \theta = \theta_0$

4. The test statistic is

$$TS = \frac{\widehat{\theta} - \theta_0}{\sigma_{\widehat{\theta}}} \sim N(0,1) \text{ under } H_0$$

and the observed test statistic using the given data is ts;

If the alternative hypothesis is

- $H_a : \theta > \theta_0$ then p-value $= P(TS > ts | \theta = \theta_0) = 1 - \Phi(ts)$
- $H_a : \theta < \theta_0$ then p-value $= P(TS < ts | \theta = \theta_0) = \Phi(ts)$
- $H_a : \theta \neq \theta_0$ then p-value $= P(|TS| > |ts| | \theta = \theta_0)$
  $= P(TS > |ts| | \theta = \theta_0) + P(TS < -|ts| | \theta = \theta_0) = 2(1 - \Phi(|ts|))$

5.6 Small-sample hypothesis tests for population means

If the sample size is small ($n < 30$) then we cannot hope that the Central Limit Theorem will ensure that the test statistic

$$TS = \frac{\widehat{\theta} - \theta_0}{\widehat{\sigma}_{\widehat{\theta}}}$$

has the standard normal distribution. In this case the only way out is to make sure that the data is at least approximately normal, perhaps by applying an appropriate transformation to the data.

From now on, in this section we will assume that the data is normal. Even in this case, the distribution of the test statistic differs significantly from the normal distribution. This means that when we calculate the probabilities of type I and II errors, or when we calculate the p-values, we cannot calculate probabilities like

$$\mathbb{P}(TS > x)$$

as if the TS were a standard normal random variable. This would result in wrong probabilities.

Luckily, the distribution for this test statistics is still known and can be calculated by a computer algorithm. It can also can be found in tables.

$$TS = \frac{\widehat{\theta} - \theta_0}{\sigma_{\widehat{\theta}}} \sim t\text{-distribution if } H_0 \text{ is true } (\theta = \theta_0),$$

The degrees of freedom for t-distributions depends on whether $\theta = \mu$ or $\mu_1 - \mu_2$.

If we are interested in testing of $H_0 : \mu = \mu_0$, then we use the test statistic

$$TS = \frac{\overline{Y} - \mu_0}{S/\sqrt{n}} \sim t_{n-1} \text{ if } H_0 \text{ is true } (\mu = \mu_0)$$

If we have two samples, $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$,with population means μ_1 and μ_2, respectively, then we are often interested in testing $H_0 : \mu_1 - \mu_2 = \theta_0$.

$$v = \frac{\left(\frac{s_1^2}{m} + \frac{s_2^2}{n}\right)^2}{\frac{(s_1^2/m)^2}{m-1} + \frac{(s_2^2/n)^2}{n-1}} = \frac{\left[(se_1)^2 + (se_2)^2\right]^2}{\frac{(se_1)^4}{m-1} + \frac{(se_2)^4}{n-1}}$$

where

$$se_1 = \frac{s_1}{\sqrt{m}} \qquad se_2 = \frac{s_2}{\sqrt{n}}$$

(round v down to the nearest integer).

Figure 5.1: Degrees of freedom for the test statistic when the variances are not the same

Here, two different situations are possible. A bit simple situation is when we can assume that the variances in two samples are the same $\sigma_1^2 = \sigma_2^2 = \sigma^2$. (We could check this assumption by an appropriate test!) Then we can use the test statistic

$$TS = \frac{\overline{X} - \overline{Y} - \theta_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2} \text{ if } H_0 \text{ is true } (\theta = \theta_0);$$

where we use the pooled-sample standard deviation as an estimator for σ^2.

$$S_p = \sqrt{S_p^2} = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}} = \dots$$

In this case the degrees of freedom of the t-distribution are $df = n_1 + n_2 - 2$.

A more difficult situation arises when we cannot assume that the variances σ_1^2 and σ_2^2 are equal. Then we have to use this test statistic:

$$TS = \frac{\overline{X} - \overline{Y} - \theta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}},$$

where S_1^2 and S_2^2 are sample variances in the two samples. It turns out that the distribution of this TS is *approximately* a t-distribution, but the formula for the degrees of freedom is quite complicated. See Figure 5.1.

Note, however, that some researchers suggested that this procedure should be used if there are doubts about whether the variances are same.

After the distribution of the test-statistic is determined the rest is simple, one would only need to replace z_α with the t_α that is calculated for the t-distribution with correct number of degrees of freedom.

- $H_a : \theta > \theta_0 \Leftrightarrow RR : \{TS > t_\alpha\}$
- $H_a : \theta < \theta_0 \Leftrightarrow RR : \{TS < -t_\alpha\}$
- $H_a : \theta \neq \theta_0 \Leftrightarrow RR : \{|TS| > t_{\alpha/2}\}$

The quantities t_α can be found from the tables or by using the R command *qt*. In particular t_α for ν degrees of freedom can be calculated as $\texttt{qt}(1-\alpha, \nu)$.

The calculation of the probability of type II error β and the power $1-\beta$ is in fact very similar to the calculations in the case of the normal distribution. Again, one only needs to use the t-distribution with the correct number of degrees of freedom instead of the standard normal distribution.

This is also true for p-values.

- $H_a : \theta > \theta_0 \Leftrightarrow p\text{-value} = P(TS > ts|\theta = \theta_0)$
- $H_a : \theta < \theta_0 \Leftrightarrow p\text{-value} = P(TS < ts|\theta = \theta_0)$
- $H_a : \theta \neq \theta_0 \Leftrightarrow p\text{-value} = P(|TS| > |ts||\theta = \theta_0) = 2P(TS > |ts|)$

where the test statistic has the t-distribution with an appropriate number of degrees of freedom. The tables only give a range for p-value. For precise probability, one must use R command `pt(a,df)` whose output is $\mathbb{P}(T < a)$ where $T \sim t(df)$.

Example 5.6.1. An Article in *American Demographics* investigated consumer habits at the mall. We tend to spend the most money when shopping on weekends, particularly on Sundays between 4:00 and 6:00PM, while Wednesday-morning shoppers spend the least.

Independent random samples of weekend and weekday shoppers were selected and the amount spent per trip to the mall was recorded as shown in the following table:

Weekends	Weekdays
$n_1 = 20$	$n_2 = 20$
$\overline{y}_1 = \$78$	$\overline{y}_2 = \$67$
$s_1 = \$22$	$s_2 = \$20$

- Is there sufficient evidence to claim that there is a difference in the average amount spent per trip on weekends and weekdays? Use $\alpha = 0.05$.
- What is the attained significance level (p-value)?
- What if $n_1 = 10$ and $n_2 = 10$?

Let us do this example for $n_1 = n_2 = 20$ and assume that $\sigma_1^2 = \sigma_2^2$.

Our null hypothesis is that $\mu_1 = \mu_2$, so we want to calculate

$$TS = \frac{\overline{X} - \overline{Y}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where

$$S_p = \sqrt{S_p^2} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{19 \times 22^2 + 19 \times 20^2}{20 + 20 - 2}}$$
$$= 21.0238$$

So,

$$TS = \frac{78 - 67}{21.0238\sqrt{\frac{1}{20} + \frac{1}{20}}} = 1.654556$$

For the normal distribution $z_{0.05} = 1.645$ so the test would reject the null hypothesis if the sample were large.

If we want to test H_0 at the level $\alpha = 0.05$ and use the t-distribution we want t_α for $\nu = 20 + 20 - 2 = 38$.

We calculate it as $qt(.95, 38) = 1.685954$, so we conclude that the evidence is not sufficient to reject H_0 at the level $\alpha = 0.05$.

The p-value can be calculated as $1 - pt(1.654556, 38) = 5.31\%$.

5.7 Hypothesis testing for population variances

Occasionally, we are interested in testing variances. The most frequent example is when we test equality of variance in two samples, in order to see if the corresponding populations are really different in a certain aspect. Sometimes we might me interested to see that the variance does not exceed a certain threshold. This problem arises in quality control.

Let us consider first testing the hypothesis $H^0 : \sigma^2 = \sigma_0^2$ If we sample is large, then we can use

$$TS = \frac{S^2 - \sigma_0^2}{\sigma_{S^2}},$$

with a suitable estimator for σ_{S^2}.

This approach does not generalizes easily to small samples since it is difficult to calculate the exact distribution of the ration. So we look at an alternative method, which works both for large and small samples.

So let us assume that $X_1, \ldots, X_n$ are from a Normal distribution $N(\mu, \sigma^2)$ with unknown mean μ and unknown variance σ^2.

We use the result that we know from the section about variance estimation.

$$TS = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1) \text{ when } H_0 \text{ is true}$$

Fore the case when the alternative hypothesis is $H_a : \sigma^2 > \sigma_0^2$, the rejection region is similar to the RR in z- and t-tests:

$$RR : TS > \chi^2_\alpha(n-1),$$

where $\chi^2_\alpha(n-1)$ solves the equation $\mathbb{P}(TS > x) = \alpha$ if we know that TS is distributed as a χ^2 random variable with $n-1$ degrees of freedom.

The R command for calculating this quantity is `qchisq`$(1-\alpha, n-1)$.

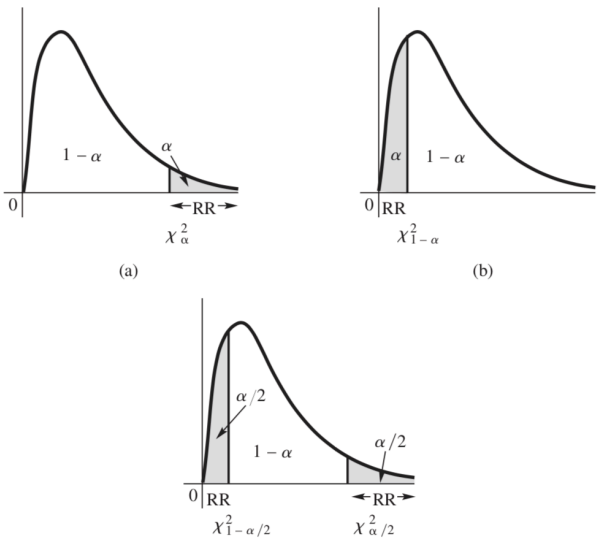


FIGURE **10.10** Rejection regions RR for testing $H_0 : \sigma^2 = \sigma_0^2$ versus (a) $H_a : \sigma^2 > \sigma_0^2$; (b) $H_a : \sigma^2 < \sigma_0^2$; and (c) $H_a : \sigma^2 \neq \sigma_0^2$

For the alternative hypothesis $H_a : \sigma^2 < \sigma_0^2$, there is some difference from the case of z- or t-tests because the χ^2 distribution is not symmetric relative to zero. Instead of using $-\chi^2_\alpha(n-1)$ as a threshold, we use $\chi^2_{1-\alpha}(n-1)$. So the rejection region in this case is

$$RR : TS < \chi^2_{1-\alpha}(n-1),$$

Finally, if the alternative hypothesis is $H_a : \sigma^2 \neq \sigma_0^2$, then it is conventional to use the following rejection region:

$$RR : TS\chi^2_{1-\alpha/2}(n-1) \text{ or } TS > \chi^2_{\alpha/2}(n-1)$$

Correspondingly, the p-values for these alternative hypotheses are as follows. If $H_a : \sigma^2 > \sigma_0^2$, then

$$p\text{-value} = P(TS > ts) = 1 - \texttt{pchisq}(ts, n-1).$$

If $H_a : \sigma^2 < \sigma_0^2$, then

$$p\text{-value} = P(TS < ts) = \texttt{pchisq}(ts, n-1)$$

If $H_a : \sigma^2 \neq \sigma_0^2$, then we need to think a bit harder. When we decrease α then at some point either $\chi^2_{\alpha/2}$ or $\chi^2_{1-\alpha/2}$ hits the value of the tests statistic

ts that was realized in the sample. At this moment the test stops rejecting the null hypothesis. So if $\chi^2_{\alpha/2}$ hits ts first, then we conclude that this is the critical α^* which is equal the p-value. Hence in this case the p-value equals

$$\alpha^* = 2\mathbb{P}(TS > \chi^2_{\alpha^*/2}) = 2\mathbb{P}(TS > ts)$$

Note that at that moment $ts > 1 - \alpha^*/2$ so $\mathbb{P}(TS < ts) > 1 - \alpha^*/2$ and $2P(TS < ts) > 2 - \alpha^* > 1$.

If $\chi^2_{1-\alpha/2}$ hits ts first, then by a similar argument we find that p-value equals

$$\alpha^* = 2\mathbb{P}(TS < \chi^2_{1-\alpha/2}) = 2\mathbb{P}(TS < ts)$$

and $P(TS > ts) > 1$.

So we have two candidates for the p-value: $2P(TS < ts)$ and $2P(TS > ts)$, and we know that if one of them is indeed the p-value (and so less than 1 as a probability), then the other is greater than 1. So we can simply choose the minimal of these two numbers. In summary,

$$p\text{-value} = 2 \times \min[P(TS > ts), P(TS < ts)]$$

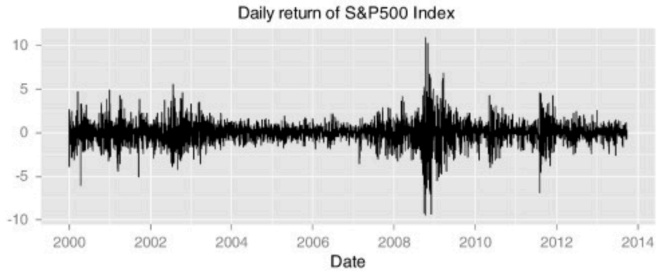
In case you pick up the wrong one between $P(TS > ts)$ or $P(TS < ts)$, your answer will exceed 1, which is an immediate warning sign because probability cannot be greater than 1.

Example 5.7.1. An experimenter was convinced that the variability in his measuring equipment results in a standard deviation of 2. Sixteen measurements yielded $s^2 = 6.1$. Do the data disagree with his claim? Determine the p-value for the test. What would you conclude if you chose $\alpha = 0.05$?

- $H_0 : \sigma^2 = 4$ and $H_a : \sigma^2 \neq 4$
- Test statistic: $\frac{(n-1)S^2}{4}$
- Observed value for test statistic is $ts = \frac{(16-1)6.1}{4} = 22.875$
- R gives $pchisq(22.875, 15) = 0.9132$ and $1 - pchisq(22.875, 15) = 0.0868$.

- Hence p-value $= 2 \times 0.0868 = 17.36\% > 5\%$.

We conclude that the data do not give enough evidence to disagree with his claim.



Now consider the test for equality of variances in two population. In some situation, a researcher is interested to know whether the data variation in two samples indicated the different variances in corresponding populations. For example:

- comparing precision of two measuring instruments;
- the variation in quality of a product at two locations or at two different time periods;
- variation is scores for two test procedures.
- variation in outcomes for two medical procedures.
- variation in market returns for two time periods or in two different countries.

Suppose that we have two samples with n_1 and n_2 observations, respectively, from the normal distributions with variances σ_1^2 and σ_2^2 respectively. We want to test the hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ against the alternative $H_a : \sigma_1^2 > \sigma_2^2$.

If S_1^2 and S_2^2 are sample variances, then define the test statistic

$$TS = \frac{S_1^2}{S_2^2}.$$

Under the null hypothesis this ratio is distributed as so-called F-distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom. (F is for Fisher, who designed the test.)

So we can use the Rejection Region

$$RR = \{TS > F_\alpha\}$$

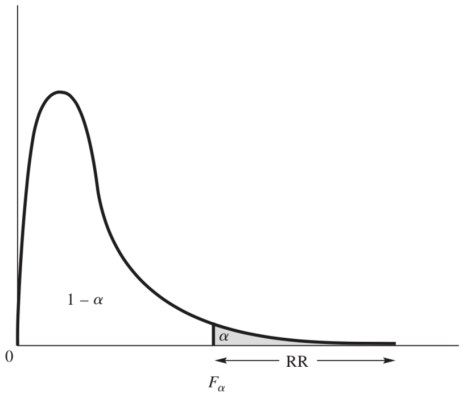


FIGURE **10.12** Rejection region RR for testing $H_0 : \sigma_1^2 = \sigma_2^2$ versus $H_a : \sigma_1^2 > \sigma_2^2$

If the hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ but we test against the alternative $H_a : \sigma_1^2 < \sigma_2^2$. (instead of $H_a : \sigma_1^2 > \sigma_2^2$) then we can simply use

$$TS = \frac{S_2^2}{S_1^2}.$$

instead of

$$TS = \frac{S_1^2}{S_2^2}.$$

The new tests statistics is distributed as F-random variable with degrees of freedom $n_2 - 1$ and $n_1 - 1$.

What if we want to test H_0 against the alternative hypothesis $H_a : \sigma_1^2 \neq \sigma_2^2$?

It turns out that in this case we can use the test statistics

$$TS = \frac{S_1^2}{S_2^2},$$

which is distributed as the F-random variable with $n_1 - 1$ and $n_2 - 1$ degrees of freedom and use the rejection region

$$RR = \left\{ \frac{1}{TS} > F^{n_2-1}_{n_1-1;\alpha/2} \text{ or } TS > F^{n_1-1}_{n_2-1;\alpha/2} \right\}$$

Notice also the degrees of freedom in the numerator and denominator of the thresholds!

It is worthwhile to repeat: the test is very sensitive to the assumption that the data are normally distributed.

Example 5.7.2. A study was conducted by the Florida Game and Fish Commission to assess the amounts of chemical residues found in the brain tissue of brown pelicans. In a test for DDT, random samples of n1 = 10 juveniles and n2 = 13 nestlings produced the results shown in the accompanying table (measurements in parts per million, ppm).

Juveniles	Nestlings
$n_1 = 10$	$n_2 = 13$
$\overline{y}_1 = .041$	$\overline{y}_2 = .026$
$s_1 = .017$	$s_2 = .006$

Are you willing to assume that the underlying population variances are equal? Test $\sigma_1^2 = \sigma_2^2$ against $\sigma_1^2 > \sigma_2^2$ at $\alpha = 0.01$. What is the p-value?

The test statistic is

$$TS = \frac{0.017^2}{0.006^2} = 8.027778.$$

p-value $= 1 - \texttt{pf}(8.027778, 9, 12) = 0.07\%$.

So we would reject the null at 1% level.

5.8 Neyman - Pearson Lemma and Uniformly Most Powerful Tests

So far we talked about specific tests and have not be concerned with evaluating and comparing tests.

Recall that for the estimation problem we had the concept of the Minimal Variance Unbiased Estimator. If we can find such an estimator, then we could agree, that this is the best estimator among available.

For a test the natural measure of its goodness is its power. If we can have two tests with the same α, we will prefer the test that has the large power.

Note however that the power of the test is a function of the parameter under the alternative hypothesis. So if one test has larger power than another under one value of the parameter θ_a, it can actually has smaller power under another value of θ_a.

Recall that the power of the test $= 1 - \beta = \Pr($ reject $H_0|$ if H_a is true).

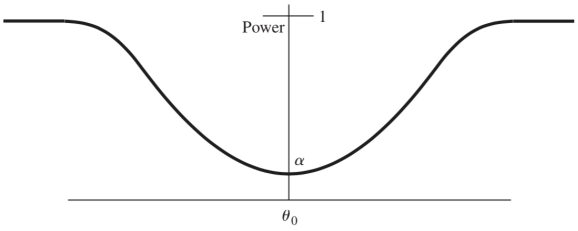
It can be calculated only if H_a is a simple hypothesis.

Definition 5.8.1. A hypothesis is said to be simple if this hypothesis uniquely specifies the distribution of the population from which the sample is taken. Any hypothesis which is not simple is called a composite hypothesis.

For example, the hypothesis $\theta = 2$ is simple, the hypothesis $\theta > 2$ is composite.

Power is a curve, it depends on the value of the parameter in the alternative hypothesis. Can we build a test with the "best" power curve?

FIGURE **10.13**
A typical power curve for the test of $H_0 : \theta = \theta_0$ against the alternative $H_a : \theta \neq \theta_0$



First, let us not to be too ambitious and try to find the test with the maximum power when the significance level is α and when the alternative hypothesis is simple, $H_a : \theta = \theta_a$.

Theorem 5.8.2 (The Neyman-Pearson Lemma)**.** *For testing between* $H_0 : \theta = \theta_0$ *vs* $H_a : \theta = \theta_a$*, the test with the reject region*

$$RR = \left\{ \frac{L(\theta_0|X)}{L(\theta_a|X)} < t \right\}$$

where t *is chosen so that* $\mathbb{P}(\frac{L(\theta_0)}{L(\theta_a)} < t|\theta = \theta_0) = \alpha$*, is the most powerful* α*-level test for* H_0 *versus* H_a*.*

In other words, if the alternative hypothesis is simple: $\theta = \theta_\alpha$, the best test statistic is

$$TS = \frac{L(\theta_0|X)}{L(\theta_a|X)}$$

This TS measures how likely is the data under the null hypothesis compared with its likelihood under the alternative hypothesis. You reject H_0 if the ratio is too small, with the threshold chose so that the level of this test is α.

The theorem says that this test has the largest power to reject H_0 (among the tests with the same α) provided the alternative is fixed at θ_a.

In order to construct the Neyman-Pearson test, we need to know the distribution of the test statistics, which is not always easy. Here is an example where the distribution of $\frac{L(\theta_0)}{L(\theta_a)}$ is not too hard to find.

Example 5.8.3. Suppose we have just one observation in the sample, $Y \sim f(y|\theta) = \theta y^{\theta-1} \mathbb{1}_{\{0<y<1\}}$. Find the most powerful test for $H_0 : \theta = 2$ against $H_a : \theta = 1$ at significance level $\alpha = 0.05$.

The likelihood here is simply $L(\theta|y) = \theta y^{\theta-1}$. (Only one observation - so no products.)

The ratio is

$$\frac{L(\theta_0|y)}{L(\theta_a|y)} = \frac{\theta_0}{\theta_a} y^{\theta_0-\theta_a}$$

If we want the size of the test be α, we should have

$$\begin{aligned}\Pr\left[\frac{L(\theta_0|Y)}{L(\theta_a|Y)} < t \middle| H_0\right] &= \Pr\left[\frac{\theta_0}{\theta_a} Y^{\theta_0-\theta_a} < t \middle| H_0\right] = \\ &= \Pr\left[Y < \left(\frac{\theta_a}{\theta_0} t\right)^{1/(\theta_0-\theta_a)} \middle| H_0\right] = \alpha.\end{aligned}$$

Under the null hypothesis, Y has density $\theta_0 y^{\theta_0-1}$, so we can calculate the cumulative distribution function as $F_Y(y) = y^{\theta_0}$, and

$$\begin{aligned}\Pr\left[\frac{L(\theta_0|Y)}{L(\theta_a|Y)} < t \middle| H_0\right] &= \Pr\left[Y < \left(\frac{\theta_a}{\theta_0} t\right)^{1/(\theta_0-\theta_a)} \middle| H_0\right] \\ &= \left(\frac{\theta_a}{\theta_0} t\right)^{\theta_0/(\theta_0-\theta_a)} = \alpha.\end{aligned}$$

Hence, the threshold in the test should be set to

$$t = \frac{\theta_0}{\theta_a}\alpha^{(\theta_0-\theta_a)/\theta_0}.$$

In our example, the test statistic is $\frac{\theta_0}{\theta_a}Y^{\theta_0-\theta_a} = 2Y$, and

$$t = \frac{2}{1}0.05^{(2-1)/2} = 2 \times 0.2236.$$

Equivalently, the most powerful test with $\alpha = 0.05$ in this case has $RR = \{Y < 0.2236\}$

In N-P lemma, the test is guaranteed to be most powerful level-α test against a specific alternative hypothesis. What if we try a different alternative hypothesis?

Consider the previous example: We found that the most powerful test has the rejection region:

$$\left\{\frac{\theta_0}{\theta_a}Y^{\theta_0-\theta_a} < \frac{\theta_0}{\theta_a}\alpha^{(\theta_0-\theta_a)/\theta_0}\right\} = \left\{Y^{\theta_0-\theta_a} < \alpha^{(\theta_0-\theta_a)/\theta_0}\right\}$$

If $\theta_a < \theta_0$, then we can take the power $1/(\theta_0 - \theta_a)$ on both sides and get that the rejection region is

$$\left\{Y < \alpha^{1/\theta_0}\right\}.$$

It is the same for all $\theta_a < \theta_0$. But if $\theta_a > \theta_0$ then we would get a completely different test:

$$\left\{Y > (1-\alpha)^{1/\theta_0}\right\}$$

Say if in the previous example we choose $H_a : \theta = 4$ then we would get test $RR = \{Y > \ldots\}$.

When a test (that is a test statistic TS and a rejection region RR) maximizes the power for every value of $\theta \in \Omega_a$, it is said to be a uniformly most powerful (or UMP) test for $H_0 : \theta = \theta_0$ versus composite hypothesis $H_a : \theta \in \Omega_a$.

For example, in our previous example, the Neyman-Pearson test is the UMP for $H_a : \theta > \theta_0$. It is also UMP for $H_a : \theta < \theta_0$. (Note that in this case

it is actually a different test. We call it the Neyman-Pearson test because it was obtained by applying the Neyman - Pearson lemma to a specific $\theta_a < \theta$. It is just happened in this example that all these test coincide for all $\theta_a < \theta$.) However, in case of the two-sided hypothesis, $H_a : \theta_a \neq \theta_0$, Neyman-Pearson is not helpful because it gives two different tests depending on whether $\theta_a < \theta_0$ or $\theta_a > \theta_0$. In fact, in this case there is no uniformly most powerful test.

In many cases, even for one sided hypothesis, UMP tests do not exists. However, they are especially rare if the alternative is two sided $H_a : \theta \neq \theta_0$, or if we test a vector parameter and the alternative hypothesis is not simple. (That is, the alternative hypothesis is not just a specific value $\vec{\theta}_a$, for which the Neyman-Pearson lemma would give us a most powerful test.)

Example 5.8.4 (Neyman-Pearson lemma applied to normal data). $Y_1, \ldots, Y_n \sim N(\mu, \sigma)$. Consider $H_0 : \mu = \mu_0$ versus $H_a : \mu = \mu_1$. We assume that σ^2 is KNOWN (fixed). Otherwise the N-P lemma is not applicable: the hypothesis is not simple. What is the most powerful test with level α?

The likelihood is

$$L(\mu) = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right\}$$

The ratio of the likelihoods defined in the Neyman Pearson lemma is:

$$\begin{aligned} \frac{L(\mu_0)}{L(\mu_1)} &= \frac{(2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu_0)^2\right\}}{(2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu_1)^2\right\}} \\ &= \exp\left\{ -\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i-\mu_0)^2 - \sum_{i=1}^{n}(y_i-\mu_1)^2\right]\right\} \\ &= \exp\left\{ -\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n} 2y_i(\mu_1-\mu_0) + \mu_0^2 - \mu_1^2\right]\right\} \end{aligned}$$

The Neyman Pearson Lemma says that the most powerful test is the one

with some appropriate threshold t which rejects H_0 when

$$\exp\left\{-\frac{1}{2\sigma^2}\Big[\sum_{i=1}^{n} 2y_i(\mu_1-\mu_0)+\mu_0^2-\mu_1^2\Big]\right\} < t$$

$$\iff -\frac{1}{2\sigma^2}\Big[\sum_{i=1}^{n} 2y_i(\mu_1-\mu_0)+\mu_0^2-\mu_1^2\Big] < t'$$

$$\iff \sum_{i=1}^{n} 2y_i(\mu_1-\mu_0)+\mu_0^2-\mu_1^2 > t''$$

$$\iff 2n\overline{y}(\mu_1-\mu_0) > t'''$$

That is, the test tells us to reject H_0 when

$$\overline{y} > t'''/(2n(\mu_1-\mu_0)), \text{ if } \mu_1-\mu_0>0, \text{ or,}$$
$$\overline{y} < t'''/(2n(\mu_1-\mu_0)), \text{ if } \mu_1-\mu_0<0.$$

$$\overline{y} > A, \text{ if } \mu_1-\mu_0>0, \text{ or,}$$
$$\overline{y} < B, \text{ if } \mu_1-\mu_0<0,$$

where the thresholds A and B are chosen so that the level of the test is α. Indeed, such a test would be what our intuition would have driven us.

The NP lemma provides theoretical justification for this test. Is this test uniformly most powerful among all the tests against the composite alternative hypothesis $H_a : \mu > \mu_0$? the composite $H_a : \mu < \mu_0$? Can we construct uniformly most powerful tests for $H_a : \mu \neq \mu_0$?

5.9 Likelihood ratio test

Theoretical Question #2 How do we design a test?

Suppose we have a model with many parameters $\vec{\theta} = (\theta_1, \ldots, \theta_k)$, and want to test that one or more of the parameters is 0. Or may be we want to test some relationship between parameters, such as $\theta_1 = \theta_2$, or more generally that $c_1\theta_1 + c_2\theta_2 + \ldots + c_k\theta_k = 0$.

How do we test such a hypothesis?

One approach for the design of a statistical test is to find an estimator for a parameter that encapsulate our hypothesis, and then use our knowledge about the distribution of this estimator. For example, if we test the hypothesis $\theta_1 = \theta_2$, then we can find the ML estimator for the difference $\theta_1 - \theta_2$ and use the fact that in large samples this estimator is normal and that it is possible to calculate its variance. This is a very useful approach. Its deficiency is that we need to obtain the estimator and its variance before we are able to construct the test. In addition, the variance often depends on the true value of parameter θ so if our null hypothesis is not simple but has some nuisance parameters, then we are in trouble.

The second approach is based on the Neyman-Pearson lemma and uses the ratio $\frac{L(\vec{\theta}_0|\vec{Y})}{L(\vec{\theta}_a|\vec{Y})}$ as the test statistic. This will give the most powerful test. The deficiencies is that we need to find the distribution of this test statistic. Additionally, this approach is quite restrictive. It requires both the null and alternative hypotheses to be simple and not composite.

In this section we consider the third alternative, which is often very convenient since it works for composite hypotheses, and in case of large samples requires essentially no calculation except the maximization of some likelihood functions.

Let Ω_0 be the set of parameters that satisfy our null hypothesis. For example, it can be that $\Omega_0 =$ the set of all parameters $\vec{\theta}$ such that $\theta_1 = \theta_2$, and all other parameters can be arbitrary. Then, let Ω_a be the set of all possible alternative values for parameters. For example, our alternative can be that $\theta_1 > \theta_2$ and all other θ_i are arbitrary. The alternative hypothesis is typically a composite hypothesis in practical applications.

Define the total feasible parameter set $\Omega = \Omega_0 \cup \Omega_a$. Define the likelihood ratio statistic by

$$\lambda = \frac{\max_{\vec{\theta} \in \Omega_0} L(\vec{\theta}|\vec{Y})}{\max_{\vec{\theta} \in \Omega} L(\vec{\theta}|\vec{Y})}$$

where $L(\vec{\theta}|\vec{Y})$ is the likelihood of the vector parameter $\vec{\theta}$ given that the observed data is $\vec{Y} = (Y_1, \ldots, Y_n)$.

Use the rejection region $RR = \{\lambda < k\}$, where the threshold k is deter-

mined by the requirement that the level of the test is α.

This appears to be not an especially useful since we need to do two constrained maximizations and we do not know the distribution of λ, so we cannot calculate the threshold k. In fact, it appears that λ is a quite complicated function of the data: it is the ratio of two constrained maximums of the likelihoods, which are itself complicated functions!

The power of this method is that we can do the maximization numerically and this is a relatively easy given enough computing power. The most important fact, however, is that k can be calculated efficiently when the data sample is large.

Conceptually, the likelihood ratio test makes a lot of sense.

- If H_0 is true, then with high probability, the constrained maximum likelihood (with maximum over Ω_0) would be close to unconstrained maximum likelihood (maximum over Ω), the denominator would give the same result as the numerator, and λ would be around 1.
- If H_0 is false while H_a is true, then the unconstrained maximum likelihood would be much larger than the constrained maximum likelihood, the denominator likely to be much greater than the numerator, which leads to a small value of λ
- Therefore, we may use λ as a test statistic and reject H_0 if $\lambda < k$.

But what threshold k to choose if we want level α test?

The practical applications of the likelihood ratio test are based on the following amazing theorem.

Theorem 5.9.1 (Wilks' Theorem)**.** *Let $X_1, X_2, \ldots, X_n$ are i.i.d random observations and $\lambda(X_1, \ldots X_n)$ is the likelihood ratio for the hypothesis Ω_0 against Ω_a. Let r_0 denote the number of free parameters that are specified by $H_0 \colon \vec{\theta} \in \Omega_0$ and let r denote the number of free parameters specified by the statement $\vec{\theta} \in \Omega$. Then, for large n, for all $\theta_0 \in \Omega_0$, the statistic $-2 \log \lambda$ has approximately a χ^2 distribution with $r - r_0$ degree of freedoms.*

The number of free parameters is the total number of parameters minus the number of *equality* constraints, so the difference $r - r_0$ is simply the

excess in the number of equality constraints that define Ω_0 over the number of equality constraints that define Ω. (Note that the inequality constraints do not count - they are not important for large sample analysis.)

Therefore, the rejection region for the likelihood ratio test in large samples has a very simple form:

$$RR : \{-2 \log \lambda > \chi^2_\alpha (r - r_0)\},$$

The proof of Wilks' Theorem is not easy and will not be given here.

Example 5.9.2. Suppose that an engineer wishes to compare the number of complaints per week filed by union stewards for two different shifts at a manufacturing plant. One hundred independent observations on the number of complaints gave means $\overline{x} = 20$ for shift 1 and $\overline{y} = 22$ for shift 2. Assume that the number of complaints per week on the i-th shift has a Poisson distribution with mean λ_i, for $i = 1, 2$. Use the likelihood ratio method to test $H_0 : \lambda_1 = \lambda_2$ against $H_a : \lambda_1 \neq \lambda_2$ with $\alpha \approx 0.01$.

By taking the product the individual density functions we find the likelihood function:

$$L(\lambda_1, \lambda_2) = \frac{1}{C} e^{-n\lambda_1} (\lambda_1)^{\sum_{i=1}^n x_i} \times e^{-n\lambda_2} (\lambda_2)^{\sum_{i=1}^n y_i},$$

where $C = x_1! \dots x_n! y_1! \dots y_n!$ and $n = 100$.

Here we will be able to do maximizations analytically, although it could also be done numerically.

Log likelihood function:

$$\ell(\lambda_1, \lambda_2) = -\log C + \Big(\sum_{i=1}^n x_i \Big) \log \lambda_1 - n\lambda_1 + \Big(\sum_{i=1}^n y_i \Big) \log \lambda_2 - n\lambda_2.$$

If it is assumed that $\lambda_1 = \lambda_2 = \lambda$, then the maximization of the log-likelihood function leads (after a calculation) to the constrained MLE estimator

$$\widehat{\lambda}^{ML} = \frac{1}{2}(\overline{x} + \overline{y}) = 21.$$

If we do not assume that $\lambda_1 = \lambda_2$, then the unconstrained maximum likelihood estimator of the vector (λ_1, λ_2) is (after a calculation)

$$\widehat{\lambda}_1^{ML} = \overline{x} = 20 \text{ and } \widehat{\lambda}_2^{ML} = \overline{y} = 22.$$

Then, log likelihood ratio is the *difference* of the log-likelihoods evaluated at the constrained and unconstrained MLE estimators.

$$-2\log\text{likelihood ratio} = -2\Big(\ell(\widehat{\lambda}^{ML}, \widehat{\lambda}^{ML}) - \ell(\widehat{\lambda}_1^{ML}, \widehat{\lambda}_2^{ML})\Big)$$

Calculation gives:

$$\begin{aligned}
-2\log\text{likelihood ratio} &= -2\Big(\ell(\widehat{\lambda}^{ML}, \widehat{\lambda}^{ML}) - \ell(\widehat{\lambda}_1^{ML}, \widehat{\lambda}_2^{ML})\Big) \\
&= -2\Big[-\log k + \Big(n\overline{x} + n\overline{y}\Big)\log\widehat{\lambda}^{ML} - 2n\widehat{\lambda}^{ML} \\
&\quad + \log k - n\overline{x}\log\widehat{\lambda}_1^{ML} + n\widehat{\lambda}_1^{ML} - n\overline{y}\log\widehat{\lambda}_2^{ML} + n\widehat{\lambda}_2^{ML}\Big]
\end{aligned}$$

Some terms cancel out and we find

$$\begin{aligned}
-2\log\text{likelihood ratio} &= -2\Big[(100\times 20 + 100\times 22)\log 21 \\
&\quad - 100\times 20\log 20 - 100\times 22\log 22\Big] \\
&= 9.5274
\end{aligned}$$

By Wilks theorem we should use the rejection region $RR = \{-2\log\lambda > \chi^2_{\alpha=0.01, df=1} = 6.635\}$. Hence we reject $H_0 : \lambda_1 = \lambda_0$ at significance level $\alpha = 0.01$.

In fact, in this example, we can also use the first method based on the estimator of the parameter $\lambda_1 - \lambda_2$. Indeed, since parameter λ is the mean of the Poisson distribution, the problem can be thought as a problem about the equality of means in two samples. The difficulty is that the sample standard deviations are not given. However, we know the distribution of data observations (Poisson).

Note that $\overline{x}$ is an estimator of λ_1 which is approximately normal with distribution $\mathcal{N}(\lambda_1, \lambda_1/n)$. Similarly $\overline{y}$ is approximately normal independent random variable with distribution $\mathcal{N}(\lambda_2, \lambda_2/n)$.

Hence, under the null hypothesis, we have that the test statistic

$$TS = \frac{\overline{y} - \overline{x}}{\sigma_{\overline{y}-\overline{x}}} \sim \mathcal{N}(0,1).$$

Under the null hypothesis, a good estimator of $\sigma_{\overline{y}-\overline{x}}$ is

$$\sqrt{\frac{\widehat{\lambda}}{n} + \frac{\widehat{\lambda}}{n}}$$

where $\widehat{\lambda} = \frac{1}{2}(\overline{x} + \overline{y}) = 21$. So, we calculate:

$$TS = \frac{22 - 20}{\sqrt{2 \times \frac{21}{100}}} = \frac{2 \times 10}{\sqrt{42}} = 3.086.$$

This is greater than $z_{0.01} = 2.33$, so $H_0 : \lambda_1 = \lambda_2$ should be rejected at level $\alpha = 0.01$.

5.9.1 An Additional Example

This section gives an example, in which the likelihood ratio test is designed explicitly without using the approximation provided by Wilks' theorem.

Example 5.9.3. A service station has six gas pumps. When no vehicles are at the station, let p_i denote the probability that the next vehicle will select pump i (where $i = 1, 2, \ldots, 6$). We have a sample of size n in which x_i vehicles have chosen pump i. We wish to test $H_0 : p_1 = \ldots = p_6 = \frac{1}{6}$ versus the alternative $H_a : p_1 = p_3 = p_5$; $p_2 = p_4 = p_6 = \theta \neq \frac{1}{6}$.

What is the likelihood function of the parameters $p_1, p_2, \ldots, p_6$ given the sample data if no restriction are imposed on the parameters?

This is a multinomial distribution so

$$L(\vec{p}|\vec{x}) = \binom{n}{x_1, \ldots, x_6} \prod_{i=1}^{6} p_i^{x_i}$$

What are the likelihood functions under the hypothesis H_0 and H_a, respectively?

Under H_0, the likelihood is

$$\binom{n}{x_1, \ldots, x_6} \left(\frac{1}{6}\right)^n.$$

Under H_a, we calculate that $p_1 = p_2 = p_3 = 1/3 - \theta$, and the likelihood

$$\binom{n}{x_1, \ldots, x_6} \left(\frac{1}{3} - \theta\right)^{x_1+x_3+x_5} \theta^{x_2+x_4+x_6}.$$

Suppose that $X = X_2 + X_4 + X_6$ is the number of customers in the sample that select an even numbered pump. What is the maximum likelihood estimator of the parameter θ under the alternative hypothesis H_a?

Maximization of likelihood under H_a is equivalent to maximization of log-likelihood, that is, of

$$\ell(\theta) = c + (n - X) \log\left(\frac{1}{3} - \theta\right) + X \log \theta,$$

Then,

$$\begin{aligned} \ell'(\theta) = -(n - X)\frac{1}{\frac{1}{3} - \theta} + X\frac{1}{\theta} = 0, \\ -\theta n + X\theta + X/3 - X\theta = 0, \\ \widehat{\theta}^{MLE} = \frac{X}{3n}. \end{aligned}$$

Express the likelihood ratio statistic λ in terms of X.

Under H_a, the likelihood

$$\binom{n}{x_1, \ldots, x_6} \left(\frac{1}{3} - \theta\right)^{n-X} \theta^{X}.$$

Substituting the MLE estimate of θ in the definition of λ, we get:

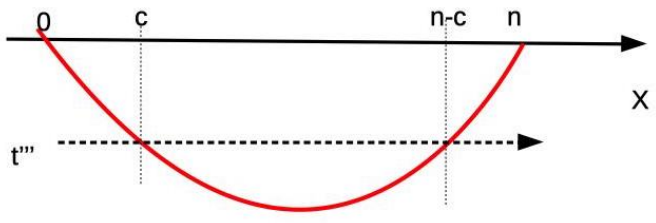
$$\lambda = \frac{(1/6)^n}{\left(\frac{1}{3} - \frac{X}{3n}\right)^{n-X} \left(\frac{X}{3n}\right)^{X}}$$

The rejection region for likelihood ratio test is $\{\lambda \leq t\}$, where t is a threshold. This is the same as $\{-\log \lambda \geq t'\}$, and we can re-write this region in our case as

$$(n - X) \log\left(\frac{n - X}{3n}\right) + X \log\left(\frac{X}{3n}\right) \geq t'',$$

or

$$(n - X) \log\Big(n - X\Big) + X \log\Big(X\Big) \geq t''',$$



The second derivative of the function on the left is

$$\frac{1}{n-X}+\frac{1}{X}>0,$$

which means that this function is convex and so:

1. there can be only two solutions to the equality
2. By symmetry, if one of these solutions is c, then the other is $n-c$.
3. The inequality is satisfied only if $X \geq c$ or if $X \leq n-c$.

Let $n=10$ and $c=9$. Determine significance level α of the test and its power when $\theta = p_2 = p_4 = p_6 = 1/10$.

Under the H_a, the probability that one of the pumps 2, 4, or 6 is visited equals 3θ. Hence X (the number of visits of these pumps) is distributed as binomial with parameter $3\theta = 0.3$. For H_0 it is the binomial with probability 0.5 Hence,

$$\begin{aligned}\alpha &= \Pr(X \geq 9 | n=10, p=0.5) + \Pr(X \leq 1 | n=10, p=0.5) \\ &= 1 - \Pr(X \leq 8 | n=10, p=0.5) + \Pr(X \leq 1 | n=10, p=0.5) \\ &= 1 - 0.989 + 0.011 = 0.022\end{aligned}$$

and

$$\begin{aligned}power = 1-\beta &= \Pr(X \geq 9 | n=10, p=0.3) + \Pr(X \leq 1 | n=10, p=0.3) \\ &= 1 - \Pr(X \leq 8 | n=10, p=0.3) + \Pr(X \leq 1 | n=10, p=0.3) \\ &= 1 - 1.000 + 0.149 = 0.149\end{aligned}$$

5.10 Quizzes

Quiz 5.10.1. A Type I error is when:

A. We reject the null hypothesis when it is actually true

B. We obtain the wrong test statistic

C. We fail to reject the null hypothesis when it's actually false

D. We reject the alternate hypothesis when it's actually true

Quiz 5.10.2. A level of significance (or size of the test) of 5% means:

A. There's a 5% chance there is an error in test decision.

B. There's a 5% chance we'll be wrong if we fail to reject the null hypothesis

C. There's a 5% chance we'll be wrong if we reject the null hypothesis.

D. The alternative hypothesis is not significant.

Quiz 5.10.3. We are interested in this problem: "Is the proportion of babies born male different from 50%?" In the sample of 200 births, we found that 96 babies born were male. We tested the claim using a test with the level of significance 1% and found that the conclusion is "Fail to reject H_0." What could we use as interpretation?

A. The proportion of babies born male is not 0.50.

B. There is not enough evidence to say that the proportion of babies born male is different from 0.50.

C. There is not enough evidence to say that the proportion of babies born male is 0.50.

Quiz 5.10.4. Suppose, we are interested in testing $H_0\colon \mu = \mu_0$ against $H_a\colon \mu > \mu_0$. We will reject H_0 at level $\alpha = 0.05$ if μ_0 is

A. larger than 95% upper confidence bound for μ.

B. larger than 95% lower confidence bound for μ.

C. smaller than 95% upper confidence bound for μ.

D. smaller than 95% lower confidence bound for μ.

Quiz 5.10.5. An educator is interested in determining the number of hours of TV watched by 4-year-old children. She wants to show that the average number of hours watched per day is more than 4 hours. To test her claim she took a random sample of 100 youngsters. Which of the following values for the sample mean would have the largest p-value associated with it.

A. 2

B. 3.9

C. 4

D. 5

Quiz 5.10.6. Suppose that I have collected a random sample to test $H_0 : \mu = \mu_0$ v.s. $H_a : \mu > \mu_0$ and I end up rejecting H_0 at level $\alpha = 0.05$ based on my sample. If I decided to change α from 0.05 to 0.01, then based on the same sample that I have in my hand, I would

A. definitely fail to reject the H_0 at level $\alpha = 0.01$;

B. definitely reject the H_0 at level $\alpha = 0.01$;

C. either reject the H_0 at level $\alpha = 0.01$ or fail to reject the H_0 at level $\alpha = 0.01$ depending on the sample;

D. have to toss a fair coin to decide what to do.

Quiz 5.10.7. Suppose that I am interested in testing $H_0 \colon \mu = \mu_0$ against $H_a \colon \mu \neq \mu_0$. I calculate the type II error probability β using the alternative value of parameter μ_a. Then, β will be smaller if I

A. Decrease the type I error probability α;

B. Decrease the sample size n;

C. Decrease the distance between μ_a and μ_0;

D. None of the above is correct.

Chapter 6

Linear statistical models and the method of least squares

6.1 Linear regression model

In the previous chapter, we have seen an example of two sample problems, in which we observed two sample and tried to understand whether one sample is different from another.

One case of this problem is when we compare the control sample with the treatment sample and try to understand whether treatment have a statistically significant effect.

You can imagine the situation when we have more than two samples corresponding to several treatments (may be different doses of drugs or may be different drugs) and we try to understand what are the effects of these treatments. The models with greater than two samples can be studied by statistical method ANOVA.

In this chapter we are introducing another model in which one variable X affect another one Y and the relation is assumed to be linear up to a random error.

In other words, we have n observations of variables X and Y: (x_1, y_1), (x_2, y_2), …(x_n, y_n) and we assume that they satisfy the following model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{6.1}$$

Here β_0 and β_1 are unknown parameters that we want to estimate. The quantities x_i, $i = 1, \ldots, n$ are known parameters, which are called *explanatory* variables, or *independent* variables. The variables ε_i are *error* terms. They are responsible for randomness in the model. They are always assumed to have zero mean: $\mathbb{E}(\varepsilon_i) = 0$. They are also often but not always assumed to have unknown variance σ_2 that does not depend on i: $\mathbb{E}(\varepsilon_i) = \sigma_2$. Even more restrictively, they are often assumed to be normally distributed: $\varepsilon_i \sim \mathcal{N}(0, \sigma_2)$.

The values y_i are random since they are functions of ε_i. (We could write them Y_i following our usual convention about random variables.) They are usually called the response variable or dependent variable. So, y_1, ..., y_n are n independent observations of the response variable Y.

If we want to relate this model to the models in previous chapters, assume that the error terms are normally distributed and note that then we can think about $y_1, \ldots, y_n$ as an a sample from the distribution $\mathcal{N}(\beta_0+\beta_1 x_i, \sigma^2)$. The observations in this sample are independent but they are not identically distributed! Indeed the mean of the i observation changes with i: $\mathbb{E}(y_i) = \beta_0 + \beta_1 x_i$.

The model (6.1) is called the *simple regression model*. It is often written in a short form that omits the subscript i:

$$Y = \beta_0 + \beta_1 x + \varepsilon.$$

A *general linear regression* model includes more than one explanatory variable:

$$y_i = \beta_0 + \beta_1 x_i^{(1)} + \ldots + \beta_k x_i^{(k)} + \varepsilon_i \tag{6.2}$$

,

or in short notation:

$$Y = \beta_0 + \beta_1 x^{(1)} + \ldots + \beta_k x^{(k)} + \varepsilon$$

This model is very flexible and can be used to model non-linear dependencies as well. For example, if we believe that the response Y depends on

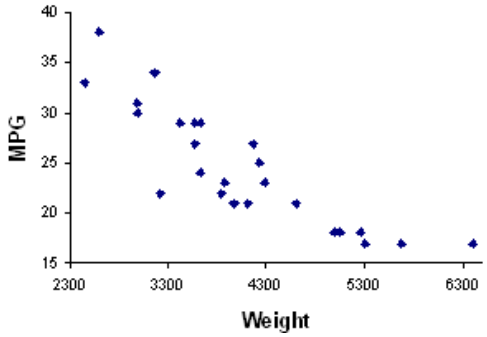


Figure 6.1: A scatter plot for the weight (in pounds) and the miles per gallon (MPG) of a car in a random sample of 25 vehicles.

explanatory variable X as a polynomial of degree three, we can add new explanatory variables that corresponds to squares and cubes of X:

$$\begin{aligned} x_i^{(1)} &= x_i, \\ x_i^{(2)} &= (x_i)^2, \\ x_i^{(3)} &= (x_i)^3. \end{aligned}$$

Then, we only need to estimate the regression model:

$$Y = \beta_0 + \beta_1 x^{(1)} + \beta_2 x^{(2)} + \beta_3 x^{(3)} + \varepsilon,$$

in order to get the desired non-linear relationship between Y and X.

In some cases we can also consider a transformation of random variable Y so as to get a more suitable distribution for random error terms ε.

Examples

The linear regression is a working horse of statistics so there are enormous number of examples. For instance,

- Energy efficiency of a vehicle and its weight.
- Wage of an individual and its education, age, experience.

Note that one of the difficult problems in statistic is to distinguish which variables are explanatory variables and which are response variables.

Goals

As usual, are goals are to estimate parameters β_0, ..., β_k and test some hypothesis about their values.

There is another goal, which we have not seen before. We might be interested in *predicting* response Y for some other values of x. In addition we might be interested in having some kind of a confidence interval for our prediction.

6.2 Simple linear regression

6.2.1 Least squares estimator

Here we look at the simple linear regression $y = \beta_0 + \beta_1 x + \varepsilon$, although the methods are also applicable to the general linear regression.

In order to estimate parameters β_0 and β_1 we could use the maximum likelihood method by writing the likelihood of the random quantities y_i and maximizing it with respect to β_0 and β_1. It turns out that for normally distributed $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ this method gives the same estimates as a simpler method described below.

This simpler method aims to minimize the deviation of the fitted values $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ from the observed values of y_i, by a choice of the estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$. Specifically, the *method of least squares* aims to minimize the Sum of Squared Errors (SSE):

$$SSE := \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2 \to \min \text{ by a choice of } \widehat{\beta}_0, \widehat{\beta}_1 \tag{6.3}$$

As usual, this minimization can be done by using the First Order Conditions.

Definition 6.2.1. The values of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ which solve the problem (6.3) are called the (ordinary) Least Squares estimators of the simple regression model (6.1).

(The estimators are called ordinary LS estimators, because sometimes in the definition of SSE the terms have different weights. In this case the solution is called the weighted least squares estimators.)

It is a bit simpler to do it for a modified model, in which the explanatory variables are centered by subtracting their mean:

$$y_i = \alpha_0 + \beta_1(x_i - \overline{x}) + \varepsilon_i$$

Clearly, this model equivalent to the original simple regression with $\beta_0 = \alpha_0 - \beta_1\overline{x}$. It is also clear that the least squares estimators in these regression problems are related by the similar equation: $\widehat{\beta}_0 = \widehat{\alpha}_0 - \widehat{\beta}_1\overline{x}$.

Theorem 6.2.2. *The least squares estimators are given by the following formulas:*

$$\widehat{\alpha}_0 = \overline{y},$$
$$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}},$$

where

$$S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$
$$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2.$$

This implies that for our original problem, we have also the following least squares estimator for the parameter β_0:

$$\widehat{\beta}_0 = \widehat{\alpha}_0 - \widehat{\beta}_1\overline{x} = \overline{y} - \widehat{\beta}_1\overline{x}$$

Proof.

$$\frac{\partial SSE}{\partial \widehat{\alpha}_0} = \frac{\partial\left\{\sum_{i=1}^{n}[y_i - (\widehat{\alpha}_0 + \widehat{\beta}_1(x_i - \overline{x}))]^2\right\}}{\partial \widehat{\alpha}_0}$$
$$= 2\sum_{i=1}^{n}[y_i - (\widehat{\alpha}_0 + \widehat{\beta}_1(x_i - \overline{x}))] \cdot (-1)$$

- Set this to be 0. We have

$$0 = \sum_{i=1}^{n}[y_i - (\widehat{\alpha}_0 + \widehat{\beta}_1(x_i - \overline{x}))] = \sum_{i=1}^{n} y_i - n\widehat{\alpha}_0 + \widehat{\beta}_1\sum_{i=1}^{n}(x_i - \overline{x})$$

- Since $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$ (we have centered them, remember?), we have $0 = \sum_{i=1}^{n} y_i - n\widehat{\alpha}_0 \Rightarrow$

$$\widehat{\alpha}_0 = \overline{y}$$

$$\begin{aligned}\frac{\partial SSE}{\partial \widehat{\beta}_1} &= \frac{\partial \left\{\sum_{i=1}^{n}[y_i - (\widehat{\alpha}_0 + \widehat{\beta}_1(x_i - \overline{x}))]^2\right\}}{\partial \widehat{\beta}_1} \\ &= 2\sum_{i=1}^{n}[y_i - (\widehat{\alpha}_0 + \widehat{\beta}_1(x_i - \overline{x}))] \cdot (-(x_i - \overline{x}))\end{aligned}$$

- Now plug in $\widehat{\alpha}_0 = \overline{y}$ which we just obtained, and set the whole thing to be 0.

$$\begin{aligned}0 &= 2\sum_{i=1}^{n}[(y_i - \overline{y}) - \widehat{\beta}_1(x_i - \overline{x})] \cdot (-(x_i - \overline{x})) \\ &= -2\sum_{i=1}^{n}\left[(y_i - \overline{y})(x_i - \overline{x}) - \widehat{\beta}_1(x_i - \overline{x})^2\right] \\ &= -2\left\{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x}) - \widehat{\beta}_1\sum_{i=1}^{n}(x_i - \overline{x})^2\right\}\end{aligned}$$

- Thus

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

□

6.2.2 Properties of LS estimator

We aim to calculate the expectation and the variance of LS estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$. This information is important for calculation of the bias of the estimators and for construction of confidence interval.

We start with $\widehat{\beta}_1$, which is typically more useful in practice since β_1 measures the effect of X on Y.

Theorem 6.2.3. *Assume that the error terms in the simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ have the properties $\mathbb{E}\varepsilon_i = 0$ and $\mathbb{V}\mathrm{ar}(\varepsilon_i) = \sigma^2$. Then,*

1. $\mathbb{E}\widehat{\beta}_1 = \beta_1$.
2. $\mathbb{V}\mathrm{ar}(\widehat{\beta}_1) = \sigma^2/S_{xx}$.

If, in addition, ε_i are normal, then $\widehat{\beta}_1$ is also normal.

Before proving this theorem, let us derive some consequences. First, we see that $\widehat{\beta}_1$ is an unbiased estimator of β_1. Second, if $S_{xx} = \sum_{i=1}^n (x_i - \overline{x})^2 \to \infty$ as $n \to \infty$, then $\widehat{\beta}_1$ is a consistent estimator of β_1. The condition $S_{xx} \to \infty$ as $n \to \infty$ means that as the sample grows we continue to observe sufficient variation in explanatory variables x_i.

Proof. It is convenient to write the model in the form: $y_i = \alpha_0 + \beta_1(x - x_i) + \varepsilon_i$.

Recall that x_i are not random. Also note that

$$S_{xy} = \sum (x_i - \overline{x})(y_i - \overline{y}) = \sum (x_i - \overline{x}) y_i.$$

Then, we have

$$\begin{aligned}
\mathbb{E}\widehat{\beta}_1 = \mathbb{E}\frac{S_{xy}}{S_{xx}} &= \frac{1}{S_{xx}} \sum (x_i - \overline{x}) \mathbb{E} y_i \\
&= \frac{1}{S_{xx}} \sum (x_i - \overline{x})(\alpha_0 + \beta_1 (x_i - \overline{x})) = \frac{1}{S_{xx}} \beta_1 \sum (x_i - \overline{x})(x_i - \overline{x}) \\
&= \beta_1,
\end{aligned}$$

which proves the first statement. Similarly, we calculate variance. It is useful to note that $\mathbb{V}\mathrm{ar}(y_i) = \mathbb{V}\mathrm{ar}(\varepsilon_i) = \sigma^2$.

$$\begin{aligned}
\mathbb{V}\mathrm{ar}(\widehat{\beta}_1) = \mathbb{V}\mathrm{ar}\Big(\frac{S_{xy}}{S_{xx}}\Big) &= \frac{1}{S_{xx}^2} \sum (x_i - \overline{x})^2 \mathbb{V}\mathrm{ar}(y_i) \\
&= \frac{1}{S_{xx}^2} S_{xx} \sigma^2 = \frac{\sigma^2}{S_{xx}}.
\end{aligned}$$

Finally, if ε_i are normal, therefore y_i are also normal. Note that $\widehat{\beta}_1$ is a weighted sum of y_i and the coefficients in this sum are non-random. We

know that this implies that the sum itself is normal. This shows that $\widehat{\beta}_1$ is normal if ε_i are normal. □

For other estimators we have similar results.

Theorem 6.2.4. *Assume that the error terms in the simple linear regression model $y_i = \alpha_0 + \beta_1(x_i - \overline{x}) + \varepsilon_i$ have the properties $\mathbb{E}\varepsilon_i = 0$ and $\mathbb{V}\mathrm{ar}(\varepsilon_i) = \sigma^2$. Then,*

1. $\mathbb{E}\widehat{\alpha}_0 = \alpha_0$.
2. $\mathbb{V}\mathrm{ar}(\widehat{\alpha}_0) = \sigma^2/n$.
3. $\mathrm{Cov}(\widehat{\alpha}_0, \widehat{\beta}_1) = 0$.

If, in addition, ε_i are normal, then $\widehat{\alpha}_0$ is also normal.

Therefore, $\widehat{\alpha}_0$ is an unbiased and consistent estimator of α_0.

Proof. For the expectation, we have:

$$\mathbb{E}\widehat{\alpha}_0 = \mathbb{E}\overline{y} = \frac{1}{n}\sum \mathbb{E}y_i = \frac{1}{n}\sum(\alpha_0 + \beta_1(x_i - \overline{x})) = \alpha_0.$$

For the variance,

$$\mathbb{V}\mathrm{ar}(\widehat{\alpha}_0) = \mathbb{V}\mathrm{ar}(\overline{y}) = \frac{1}{n^2}\sum \mathbb{V}\mathrm{ar}(y_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}.$$

Finally, for covariance,

$$\begin{aligned}\mathrm{Cov}(\alpha_0, \beta_1) &= \mathrm{Cov}\Big(\frac{1}{n}\sum y_i, \frac{1}{S_{xx}}\sum(x - x_i)y_i\Big) \\ &= \frac{1}{nS_{xx}}\sum(x - x_i)\mathbb{V}\mathrm{ar}(y_i) = \frac{1}{nS_{xx}}\sigma^2\sum(x - x_i) = 0.\end{aligned}$$

Finally, if ε_i are normal, then y_i are also normal, and since $\widehat{\alpha}_0$ is the average of y_i, $\widehat{\alpha}_0$ is also normal. □

Since $\widehat{\beta}_0 = \widehat{\alpha}_0 - \widehat{\beta}_1\overline{x}$, we have a similar result for β_0.

Theorem 6.2.5. *Assume that the error terms in the simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ have the properties $\mathbb{E}\varepsilon_i = 0$ and $\mathbb{V}\mathrm{ar}(\varepsilon_i) = \sigma^2$. Then,*

1. $\mathbb{E}\widehat{\beta}_0 = \beta_0$.

2.

$$\mathbb{V}\mathrm{ar}(\widehat{\beta}_0) = \sigma^2\Big(\frac{1}{n} + \frac{(\overline{x})^2}{S_{xx}}\Big).$$

3.

$$\mathrm{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) = \sigma^2\Big(\frac{\overline{x}}{S_{xx}}\Big).$$

If, in addition, ε_i are normal, then $\widehat{\beta}_0$ is also normal.

This result can be obtained from the formula $\widehat{\beta}_0 = \widehat{\alpha}_0 - \widehat{\beta}_1\overline{x}$ and previous theorems through an easy calculation and is left as an exercise.

We are almost ready to write down the confidence intervals for the estimates $\widehat{\beta}_1$, $\widehat{\alpha}_0$ and $\widehat{\beta}_0$. However, the variances that we have just calculated include σ_2, which is not known and should be estimated. It turns out that our previous definition of S^2 as an estimator of σ_2 is unapproriate because y_i are no longer identically distributed. A suitable estimator is as follows:

$$\widehat{\sigma}^2 := \frac{1}{n-2}SSE \equiv \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \equiv \sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2.$$

(This estimator is sometimes still denoted S^2.)

Theorem 6.2.6. *$\widehat{\sigma}^2 := \frac{1}{n-2}SSE$ is an unbiased estimator of σ^2. If the error terms are normal, then $\widehat{\sigma}^2$ is independent from $\widehat{\beta}_1$, $\widehat{\alpha}_0$ and $\widehat{\beta}_0$ and $(n-2)\widehat{\sigma}^2/\sigma^2$ has the χ^2 distribution with $n-2$ degrees of freedom.*

The fact that we have $n-2$ in the denominator instead of $n-1$ as in the previous definition of S^2 can be "explained" as that we now estimated two parameters instead of one and so lost two degrees of freedom.

I omit the proof of this theorem.

6.2.3 Confidence intervals and hypothesis tests for coefficients

Once we know the variances of the parameters, it is easy to construct the confidence intervals. The procedure is essentially the same as what we did when the estimated the mean of a sample.

For example, a large sample two-sided confidence interval for the parameter β_1 can be written as follows:

$$\Big(\widehat{\beta}_1 - z_{\alpha/2}\frac{\widehat{\sigma}}{\sqrt{S_{xx}}}, \widehat{\beta}_1 + z_{\alpha/2}\frac{\widehat{\sigma}}{\sqrt{S_{xx}}}\Big),$$

where α is the confidence level.

If the sample is small but we assume that the error terms are normal, we can use our previous theorems to come to conclusion that

$$\frac{\widehat{\beta}_1 - \beta_1}{\widehat{\sigma}/\sqrt{S_{xx}}}$$

is a pivotal quantity that has t distribution with $n-2$ degrees of freedom. In this case an appropriate confidence interval is

$$\Big(\widehat{\beta}_1 - t^{(n-2)}_{\alpha/2}\frac{\widehat{\sigma}}{\sqrt{S_{xx}}}, \widehat{\beta}_1 + t^{(n-2)}_{\alpha/2}\frac{\widehat{\sigma}}{\sqrt{S_{xx}}}\Big).$$

Similarly, if the null hypothesis is $\beta_1 = \beta_1^{(0)}$ we can form the test statistic as

$$T = \frac{\widehat{\beta}_1 - \beta_1^{(0)}}{\widehat{\sigma}/\sqrt{S_{xx}}}$$

and use this test statistic to test the null hypothesis against various alternative. If the sample is large ($n > 30$) then T is distributed as a standard normal random variable. If the sample is small, then we rely on the assumption that ε_i have normal distribution and then T has the t distribution with $df = n - 2$.

Similar procedures can be easily established for other parameters, that is for α_0 or β_0. We only need to use the appropriate variance of the estimator instead of the $\widehat{\sigma}^2/S_{xx}$.

6.2.4 Statistical inference for the regression mean

In applications we sometimes want to make some inferences about linear combinations of parameters. In this section we study a particular example of this problem. Suppose we want to build the confidence interval for the regression mean of Y, when x is equal to a specific value x^*:

$$\mathbb{E}(Y|x^*) = \beta_0 + \beta_1 x^*.$$

The natural estimator for this quantity is the predicted value:

$$\widehat{y}^* = \widehat{\beta}_0 + \widehat{\beta}_1 x^*.$$

This estimator is unbiased, because $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are unbiased estimators of β_0 and β_1. In order to build the confidence interval, we also need to calculate its variance. It is more convenient to use the other form of the regression for this task:

$$\widehat{y}^* = \widehat{\alpha}_0 + \widehat{\beta}_1(x^* - \overline{x}).$$

Then,

$$\begin{aligned}\mathbb{V}\mathrm{ar}(\widehat{y}^*) &= \mathbb{V}\mathrm{ar}(\widehat{\alpha}_0) + (x^* - \overline{x})^2 \mathbb{V}\mathrm{ar}(\widehat{\beta}_1) + 2(x^* - \overline{x})\mathrm{Cov}(\widehat{\alpha}_0, \widehat{\beta}_1) \\ &= \sigma^2\Big(\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}\Big),\end{aligned}$$

where we used our previous results about variances and covariance of estimators α_0 and β_1.

By using this information we can build the confidence intervals for y^*. For example, if the sample size is large then the two-sided confidence interval with significance level α is

$$\widehat{y}^* \pm z_{\alpha/2}\widehat{\sigma}\sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}},$$

where $\widehat{\sigma} = \sqrt{SSE/(n-2)}$ is the estimate for $\sigma = \mathbb{V}\mathrm{ar}(\varepsilon_i)$.

If the sample is small and the errors ε_i are normal, then we can use the t distribution with $n-2$ degrees of freedom and the confidence interval

becomes:

$$\widehat{y}^* \pm t_{\alpha/2}^{(n-2)} \widehat{\sigma} \sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}.$$

And for testing hypothesis $H_0 : y^* = y_0$ we use the statistic:

$$T = \frac{y^* - y_0}{\widehat{\sigma}\sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}}.$$

6.2.5 Prediction interval

When predicting Y we are often interested not in variation of our predictions $\widehat{y}$ around the true regression mean but rather in variations of the *actual* quantities y around the true regression mean. The random quantity y has larger variation than $\widehat{y}$ because in addition to uncertainty due to the error in parameter estimation it also includes the variation due to error terms ε_i.

We define the *prediction interval* with confidence level $1-\alpha$ as a (random) interval (L, U) such that

$$\mathbb{P}(L \leq y_i \leq U) = 1 - \alpha.$$

Here L and U are some statistics, so they must be computable from data.

In order to construct the prediction interval we use the pivotal quantity technique and consider

$$T = \frac{y^* - \widehat{y}^*}{SE(y^* - \widehat{y}^*)},$$

where SE stands for "standard error". Here y^* is a new observation which we try to predict and $\widehat{y}^*$ is the prediction.

Note that that

$$\begin{aligned} y^* - \widehat{y}^* &= \beta_0 + \beta_1 x^* + \varepsilon^* - (\widehat{\beta}_0 + \widehat{\beta}_1 x^*) \\ &= (\beta_0 - \widehat{\beta}_0) + (\beta_1 - \widehat{\beta}_1) x^* + \varepsilon^* \end{aligned}$$

Since $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are unbiased estimators of β_0 and β_1, we see that this quantity has expectation 0.

Moreover, if ε_i are normal then we see that $y^* - \widehat{y}^*$ is also normal.

What is the standard error of $y^* - \widehat{y}^*$? Note that we have

$$\mathbb{V}\mathrm{ar}(y^* - \widehat{y}^*) = \mathbb{V}\mathrm{ar}(\beta_0 + \beta_1 x^* + \varepsilon^* - \widehat{y}^*) = \mathbb{V}\mathrm{ar}(\varepsilon^*) + \mathbb{V}\mathrm{ar}(\widehat{y}^*),$$

because the "new" error term ε^* is un-correllated with prediction $\widehat{y}^*$. Indeed, the coefficients $\widehat{\beta}_0$ and $\widehat{\beta}_1$ were estimated using the old error terms ε_i and x^* is not random.

We calculated the variance of $\widehat{y}^*$ in the previous section, and so we have

$$\mathbb{V}\mathrm{ar}(y^* - \widehat{y}^*) = \sigma^2 + \sigma^2\Big(\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}\Big) = \sigma^2\Big(1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}\Big)$$

It follows that

$$Z = \frac{y^* - \widehat{y}^*}{\sigma\sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}}$$

has the standard normal distribution.

It can be shown that if we use the estimator $\widehat{\sigma} = \sqrt{SSE/(n-2)}$ instead of unknown σ, then the quantity

$$T = \frac{y^* - \widehat{y}^*}{\widehat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}}$$

has the t distribution with $n-2$ degrees of freedom.

So it follows that the prediction interval for y^* can be written as

$$\widehat{y}^* \pm t_{\alpha/2}^{(n-2)}\widehat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}} = \widehat{\beta}_0 + \widehat{\beta}_1 x^* \pm t_{\alpha/2}^{(n-2)}\widehat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}$$

The interpretation is that with probability $1 - \alpha$ the deviation of our prediction $\widehat{y}^*$ from the actual realization of y^* will be smaller than

$$t_{\alpha/2}^{(n-2)}\widehat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}.$$

6.2.6 Correlation and R-squared

Sometimes, x_i can be interpreted as observed values of some random quantity X. That is, we have n observations (x_i, y_i) sampled from the joint distribution of the random quantities X and Y. In this case, the coefficient β_1 in the regression $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ can be interpreted as a measure of dependence between Y and X.

On the other hand, we know that another measure of dependence between Y and X is the correlation coefficient:

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathbb{V}\mathrm{ar}(X)\mathbb{V}\mathrm{ar}(Y)}},$$

and we can estimate it as

$$R = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

Since $\widehat{\beta}_1 = S_{xy}/S_{xx}$, we see that we have the following relation between the estimates of correlation coefficient ρ and linear regression parameter β:

$$R = \beta_1 \sqrt{\frac{S_{yy}}{S_{xx}}}.$$

So there is a clear relationship between these two measures of association.

The statistic r^2 (called *R-squared*) has another useful interpretation, which will be later generalized for multiple linear regression model. Namely, it measures *goodness of fit* in the simple linear regression.

Indeed, it is possible to derive the following useful formula:

$$\begin{aligned} SSE &:= \sum_i (y_i - \widehat{y}_i)^2 = \sum_i (y_i - \overline{y} - \widehat{\beta}_1(x_i - \overline{x}))^2 \\ &= \sum_i (y_i - \overline{y})^2 - 2\widehat{\beta}_1 \sum_i (y_i - \overline{y})(x_i - \overline{x}) + \widehat{\beta}_1^2 \sum_i (x_i - \overline{x})^2 \\ &= S_{yy} - \widehat{\beta}_1^2 S_{xx} = S_{yy} - \frac{S_{xy}^2}{S_{xx}} \end{aligned}$$

Now, $S_{yy} = \sum_i (y_i - \overline{y})^2$ can be thought as the variation in the response variable if no explanatory variable is used, and SSE is the variation in the

response after the explanatory variable is brought in. So the difference is the reduction in the variation due to the explanatory variable X.

In particular,

$$R^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}}$$

is this reduction measured in percentage terms.

To summarize, R^2 is the proportion of response variable variation that is explained by the explanatory variable.

6.3 Multiple linear regression

6.3.1 Estimation

A more general version of linear regression reads:

$$y = \beta_0 + \beta_1 x^{(1)} + \ldots + \beta_p x^{(p)} + \varepsilon,$$

where we have p explanatory variables. In fact this stands for n separate equations, one for each observation:

$$y_i = \beta_0 + \beta_1 x_i^{(1)} + \ldots + \beta_p x_i^{(p)} + \varepsilon_i,$$

where $i = 1, \ldots, n$.

In full glory, we have a big system of equations:

$$\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} + \varepsilon_1 \\
y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} + \varepsilon_2 \\
&\vdots \\
y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \\
&\vdots \\
y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} + \varepsilon_n
\end{aligned}$$

In order to write down this system shorter, we use the matrix notation.

We have $(p+1)$ coefficients (1 for the intercept term and p for the independent variables), and n observations. Each observation is represented by y_i and $\boldsymbol{x}_i := [1, x_{i1}, x_{i2}, \ldots, x_{ip}]$. (We distinguish between column and row vectors, and the vector $\boldsymbol{x}_i$ here is row vector.)

Now stack all observations together to form a vector of responses and a matrix of explanatory variables.

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \underbrace{\begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix}}_{p+1 \text{ columns}} = \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_n \end{bmatrix}$$

Also define a (column) vectors of coefficients and error terms.

$$\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_p]^T, \quad \boldsymbol{\varepsilon} = [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n]^T.$$

(Here superscript T is the notation for the transposition operation. It indicates that the vector, which we wrote as a row vector is in fact the column vector. For example,

$$[\beta_0, \beta_1, \ldots, \beta_p]^T = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}.$$

The transposition can also be defined for matrices and has a useful property: $(AB)^T = B^T A^T$.

Then we can rewrite model as

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

The sum of squared errors can also be written very simply in the matrix notation:

$$
\begin{aligned}
SSE(\widehat{\boldsymbol{\beta}}) =& \sum_{i=1}^{n} \{y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \cdots + \widehat{\beta}_p x_{ip})\}^2 \\
=& \sum_{i=1}^{n} \{y_i - \boldsymbol{x}_i \widehat{\boldsymbol{\beta}}\}^2 \\
=& [\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}]^T [\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}]
\end{aligned}
$$

To minimize $SSE(\widehat{\boldsymbol{\beta}})$, we need to write the first order conditions, which can also be written in matrix form. Namely, for each $j = 1, \ldots, p$ we have:

$$
\frac{\partial SSE(\widehat{\boldsymbol{\beta}})}{\partial \beta_j} = -2 \sum_{i=1}^{n} x_{ij} \{y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \cdots + \widehat{\beta}_p x_{ip})\} = 0
$$

If we stack these $p + 1$ equations together, we obtain the matrix form of these system of equations:

$$
\frac{\partial SSE(\widehat{\boldsymbol{\beta}})}{\partial \widehat{\boldsymbol{\beta}}} = -2\mathbf{X}^T [\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}] = -2\mathbf{X}^T \boldsymbol{y} + 2\mathbf{X}^T \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{0}
$$

Or, re-arranging the terms and simplifying:

$$
\mathbf{X}^T \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^T \boldsymbol{y}.
$$

This system of $(p + 1)$ equations in $p + 1$ unknowns $\widehat{\beta}_i$ is called the *normal equations*. In matrix form, its solution can be written as

$$
\widehat{\boldsymbol{\beta}}_{LS} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{y}.
$$

6.3.2 Properties of least squares estimators

We are interested to learn about the expectations and variances of the estimated parameters β_i, $i = 1, \ldots, p$. It is convenient to express the results using the language of random vectors in matrices. In particular, if $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_p]$ is a vector of random quantities, then its expectation is the vector of expectations: $\mathbb{E}\boldsymbol{\xi} = [\mathbb{E}\xi_1, \ldots, \mathbb{E}\xi_p]$.

We can also define the (variance-)covariance matrix of vector $\boldsymbol{\xi} \in \mathbb{R}^p$ with p components as the $p \times p$ matrix whose ijth element is the covariance between the ith and jth elements of $\boldsymbol{\xi}$, i.e.

$$\mathbb{V}(\boldsymbol{\xi}) = \begin{bmatrix} \mathrm{Cov}(\xi_1, \xi_1) & \mathrm{Cov}(\xi_1, \xi_2) & \dots & \mathrm{Cov}(\xi_1, \xi_p) \\ \mathrm{Cov}(\xi_2, \xi_1) & \mathrm{Cov}(\xi_2, \xi_2) & \dots & \mathrm{Cov}(\xi_2, \xi_p) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(\xi_p, \xi_1) & \mathrm{Cov}(\xi_p, \xi_2) & \dots & \mathrm{Cov}(\xi_p, \xi_p) \end{bmatrix}$$

(For example, $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$ where I_n is the $n \times n$ identity matrix, which is a diagonal matrix with 1's on the diagonal.)

Note that the expectation can also me defined for matrices, as the matrix of expectations of entries. In this case, if ξ_i are all have zero mean, then $\mathbb{V}(\boldsymbol{\xi}) = \mathbb{E}(\boldsymbol{\xi}\boldsymbol{\xi}^T)$. More generally,

$$\mathbb{V}(\boldsymbol{\xi}) = \mathbb{E}((\boldsymbol{\xi} - \mathbb{E}\boldsymbol{\xi})(\boldsymbol{\xi} - \mathbb{E}\boldsymbol{\xi})^T). \tag{6.4}$$

We also have the following simple rules. If If A is an $m \times p$ (non-random) matrix, then we can calculate new vector $A\xi$. For this vector, the expectation and variance can be calculated as follows:

$$\mathbb{E}(A\xi) = A(\mathbb{E}\xi). \tag{6.5}$$

(This is a consequence of linearity of the expectation.) And from formula (6.4) it is easy to get the equality:

$$\mathbb{V}(A\boldsymbol{\xi}) = A\mathbb{V}(\boldsymbol{\xi})A^T. \tag{6.6}$$

Now let us apply this properties to the least squares estimator $\widehat{\boldsymbol{\beta}}$.

Recall that the $\widehat{\boldsymbol{\beta}}$ is a p-vector $[\widehat{\beta}_0, \widehat{\beta}_1, \dots, \widehat{\beta}_p]^T$ and that

$$\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}.$$

Theorem 6.3.1. *The least squares estimator of $\boldsymbol{\beta}$ is unbiased:*

$$\mathbb{E}\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}.$$

Its variance matrix is the following $(p+1) \times (p+1)$ matrix:

$$\mathbb{V}\widehat{\boldsymbol{\beta}} = \sigma(X^T X)^{-1}.$$

Proof. We know that $\mathbb{E}\boldsymbol{y} = X\boldsymbol{\beta}$ and $\mathbb{V}\boldsymbol{y} = \mathbb{V}\boldsymbol{\varepsilon} = \sigma^2 I_n$. Then, we can apply rules (6.5) and (6.6) and get:

$$\mathbb{E}\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbb{E}\boldsymbol{y} = (X^T X)^{-1} X^T X\boldsymbol{\beta} = \boldsymbol{\beta},$$

and

$$\begin{aligned}\mathbb{V}\widehat{\boldsymbol{\beta}} &= (X^T X)^{-1} X^T \mathbb{V}\boldsymbol{y} [(X^T X)^{-1} X^T]^T \\ &= (X^T X)^{-1} X^T \sigma^2 I_n X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1},\end{aligned}$$

after cancellations. □

In addition, if ε_i are normal, $\varepsilon_i \sim N(0, \sigma)$, then it can be shown that $\widehat{\boldsymbol{\beta}}$ is the multivariate normal with mean $\boldsymbol{\beta}$ and variance $\sigma^2 (X^T X)^{-1}$.

Now it is clear how to build confidence intervals and test the hypothesis for the parameters β_i. We simply notice that

$$\mathbb{V}\mathrm{ar}(\beta_i) = \sigma^2 c_{ii},$$

where c_{ii} is the i-th element on the main diagonal of the matrix $(X^T X)^{-1}$:

$$c_{ii} = \left[(X^T X)^{-1}\right]_{ii}. \tag{6.7}$$

So if σ^2 is known, then the confidence interval for β_i is

$$\widehat{\beta}_i \pm z_{\alpha/2} \sigma \sqrt{c_{ii}}.$$

In practice, σ^2 is not known and have to be estimated from data. We can do it using SSE, which is defined similarly to the case of the simple linear regression:

$$SSE = \sum_i (y_i - \widehat{y}_i)^2,$$

where $\widehat{y}_i$ are fitted values for the response variable:

$$\begin{aligned}\widehat{y}_i &= \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \ldots + \widehat{\beta}_p x_{ip} \\ &= X\widehat{\boldsymbol{\beta}} = X(X^T X)^{-1} X^T \boldsymbol{y} = X\beta + X(X^T X)^{-1} X^T \boldsymbol{\varepsilon}.\end{aligned}$$

Theorem 6.3.2. *$\widehat{\sigma}^2 := SSE/(n-p-1)$ is an unbiased estimator of σ^2.*

Proof. (This proof is optional.)

Recall that for any vector $\boldsymbol{a} = [a_1, \ldots, a_n]$, we have the notation $\|\boldsymbol{a}\|^2 := \sum_{i=1}^2 a_i^2$. We can also write this sum as $\boldsymbol{a}^T\boldsymbol{a}$.

Since $\boldsymbol{y} = X\beta + \boldsymbol{\varepsilon}$, we can do the following calculation for SSE:

$$\begin{aligned} SSE &= \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \|X\beta + \boldsymbol{\varepsilon} - (X\beta + X(X^TX)^{-1}X^T\boldsymbol{\varepsilon})\|^2 \\ &= \|(I_n - X(X^TX)^{-1}X^T)\boldsymbol{\varepsilon}\|^2 \\ &= \Big((I_n - X(X^TX)^{-1}X^T)\boldsymbol{\varepsilon}\Big)^T(I - X(X^TX)^{-1}X^T)\boldsymbol{\varepsilon} \\ &= \boldsymbol{\varepsilon}^T(I_n - X(X^TX)^{-1}X^T)^2\boldsymbol{\varepsilon}, \end{aligned}$$

A calculation gives $(I_n - X(X^TX)^{-1}X^T)^2 = I_n - X(X^TX)^{-1}X^T$.

Now, to move further, we need a fact from linear algebra. Recall that the trace of a matrix A is defined as sum of its diagonal entries: $\mathrm{tr}(A) = \sum_i a_{ii}$. An important property of trace is that $\mathrm{tr}(AB) = \mathrm{tr}(BA)$. (It is easy to prove this directly from the definition of trace.) In our case $\boldsymbol{\varepsilon}^T(I_n - X(X^TX)^{-1}X^T)^2\boldsymbol{\varepsilon}$ is a scalar (a 1×1 matrix) so it is equal to its trace. Hence, we can use this property of trace and write:

$$\boldsymbol{\varepsilon}^T(I_n - X(X^TX)^{-1}X^T)\boldsymbol{\varepsilon} = \mathrm{tr}\left[(I_n - X(X^TX)^{-1}X^T)\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^t\right].$$

Next, we use the fact that $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^t) = \mathbb{V}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$, and the property that taking expectations and taking the trace can be performed in any order and get:

$$\begin{aligned} \mathbb{E}(SSE) &= \mathrm{tr}\left[(I_n - X(X^TX)^{-1}X^T)\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^t)\right] \\ &= \sigma^2\,\mathrm{tr}\left[(I_n - X(X^TX)^{-1}X^T)I_n\right] = \sigma^2(\mathrm{tr}(I_n) - \mathrm{tr}\left[X(X^TX)^{-1}X^T\right]) \\ &= \sigma^2(n - \mathrm{tr}\left[X(X^TX)^{-1}X^T\right]) \end{aligned}$$

Now, in order to calculate $\mathrm{tr}\left[X(X^TX)^{-1}X^T\right]$, we use the property of the trace once more and write:

$$\mathrm{tr}\left[X(X^TX)^{-1}X^T\right] = \mathrm{tr}\left[(X^TX)^{-1}X^TX\right] = \mathrm{tr}(I_{p+1}) = p+1.$$

So,

$$\mathbb{E}(SSE) = \sigma^2(n-p-1)$$

and $SSE/(n-p-1)$ is an unbiased estimator of σ^2. □

Moreover, if ε_i are independent normal random variables, $\varepsilon_i \sim N(0, \sigma^2$, then it turns out that SSE/σ^2 has the ξ^2 distribution with $n-p-1$ degrees of freedom, and that this random variable is independent from $\widehat{\boldsymbol{\beta}}$. It follows that

$$\frac{\widehat{\beta}_i - \beta_i}{\widehat{\sigma}\sqrt{c_{ii}}}$$

has the t distribution with $n-p-1$ degrees of freedom. (Here c_{ii} is as defined in (6.7).) Therefore, in this case the confidence interval is

$$\widehat{\beta}_i \pm t_{\alpha/2}^{(n-p-1)} \widehat{\sigma}\sqrt{c_{ii}}.$$

6.3.3 Confidence interval for linear functions of parameters

Sometimes, we want to test a hypothesis about a linear function of parameters. For example, if we have a regression model

$$y = \beta_0 + \beta_1 x^{(1)} + \beta_2 x^{(2)} + \beta_3 x^{(3)} + \varepsilon,$$

then we might want to test the hypothesis that $\beta_1 = \beta_2$. This is equivalent to the hypothesis that $\beta_1 - \beta_2 = 1$. We can approach it by asking what is the confidence interval for $\beta_1 - \beta_2$. This confidence interval should be centered on the estimate $\widehat{\beta}_1 - \widehat{\beta}_2$ and the question is what is the standard error of this estimate.

More generaly, if we have p parameters we might be interested in finding the confidence interval for the linear combination $a_0\beta_0 + a_1\beta_1 + \ldots + a_p\beta_p$, which can be written in matrix notation as $\boldsymbol{a}^t\beta$ where $\boldsymbol{a}$ is a column vector $[a_0, \ldots, a_p]^t$. (In our previous example $\boldsymbol{a} = [0, 1, -1, 0]^t$.) The confidence interval should be centered on $\boldsymbol{a}^t\widehat{\boldsymbol{\beta}}$ and the main question is about the standard error of this estimator. It turns out that it is easy to calculate by using matrices.

Indeed, by using one of the theorems from probability theory course we have:

$$\mathbb{V}\mathrm{ar}(\sum_{i=0}^{p} a_i\widehat{\beta}_i = \sum_{i,j} a_i a_j \mathrm{Cov}(\widehat{\beta}_i, \widehat{\beta}_j).$$

In the matrix notation, this can be written as a very concise formula:

$$\mathbb{V}\mathrm{ar}(\boldsymbol{a}^t\widehat{\boldsymbol{\beta}}) = \boldsymbol{a}^t\mathbb{V}(\widehat{\boldsymbol{\beta}})\boldsymbol{a} = \sigma^2\boldsymbol{a}^t(X^tX)^{-1}\boldsymbol{a},$$

where we used the formula for the variance-covariance matrix of the estimator $\boldsymbol{\beta}$.

It follows that the confidence interval for $\boldsymbol{a}^t\beta$ can be written as

$$\boldsymbol{a}_t\widehat{\beta} \pm z_{\alpha/2}\sigma\sqrt{\boldsymbol{a}^t(X^tX)^{-1}\boldsymbol{a}}.$$

Since σ is unknown, we substitute it with its estimator. In the case of normal errors, it gives the following confidence interval:

$$\boldsymbol{a}_t\widehat{\beta} \pm t_{\alpha/2}^{(n-p-1)}\widehat{\sigma}\sqrt{\boldsymbol{a}^t(X^tX)^{-1}\boldsymbol{a}}.$$

This calculations are useful, for example if we want to calculate the confidence interval for the regression mean $\widehat{y}^* = \widehat{\beta}_0 + \widehat{\beta}_1 x_*^{(1)} + \ldots + \widehat{\beta}_p x_*^{(p)}$, for a new observation $(x_*^{(1)}, \ldots, x_*^{(p)})$.

In this case we use the formulas we derived above by setting the vector $\boldsymbol{a} = [1, x_*^{(1)}, \ldots x_*^{(p)}]^t$.

6.3.4 Prediction

Suppose we obtained a new observations with predictors (i.e., explanatory variables $x_*^{(1)}, x_*^{(2)}, \ldots, x_*^{(p)}$ and we want to predict the response variable y_*. The natural predictor is

$$\widehat{y}_* = \widehat{\beta}_0 + \widehat{\beta}_1 x_*^{(1)} + \ldots \widehat{\beta}_p x_*^{(p)} = \boldsymbol{x}_*^t\widehat{\boldsymbol{\beta}},$$

where $\boldsymbol{x}_*$ is the column vector $[1, x_*^{(1)}, \ldots, x_*^{(p)}]^t$.

This expected value of this predictor equals the regression mean $\mathbb{E}y_*$,

$$\mathbb{E}\widehat{y}_* = \boldsymbol{x}_*^t \mathbb{E}\widehat{\boldsymbol{\beta}} = \boldsymbol{x}_*^t \boldsymbol{\beta}.$$

Let us define the prediction error as the difference between the prediction and the actual realization of the response variable,

$$e_* = y_* - \widehat{y}_*,$$

Then the expected value of the error is zero and it is easy to compute its variance by using results from the previous section:

$$\begin{aligned}\mathbb{V}\mathrm{ar}(e_*) &= \mathbb{V}\mathrm{ar}(y_* - \widehat{y}_*) = \mathbb{V}\mathrm{ar}(\boldsymbol{x}_*^t\boldsymbol{\beta} + \varepsilon_* - \boldsymbol{x}_*^t\widehat{\boldsymbol{\beta}}) \\ &= \mathbb{V}\mathrm{ar}(\varepsilon_*) + \mathbb{V}\mathrm{ar}(\boldsymbol{x}_*^t\widehat{\boldsymbol{\beta}}) \\ &= \sigma^2 + \sigma^2 \boldsymbol{x}_*^t (X^t X)^{-1} \boldsymbol{x}_*.\end{aligned}$$

This allows us to write the prediction interval:

$$\boldsymbol{x}_*^t\widehat{\boldsymbol{\beta}} \pm t_{\alpha/2}^{(n-p-1)} \widehat{\sigma} \sqrt{1 + \boldsymbol{x}_*^t (X^t X)^{-1} \boldsymbol{x}_*}$$

6.4 Goodness of fit and a test for a reduced model

Recall that we defined R^2 statistic earlier as

$$R^2 = \frac{SST - SSE}{SST},$$

where

$$SST = \sum_i (y_i - \overline{y})^2, \text{ and } SSE = \sum_i (y_i - \widehat{y}_i)^2.$$

Essentially, in this calculation we compare a given regression model with a model in which only a constant allowed. The prediction of this reference model is always $\overline{y}$.

More generally, we can compare two models, *reduced* (often also called *restricted*) and *complete* (often called *full* or *unrestricted*).

$$\begin{aligned}y &= \beta_0 + \beta_1 x^{(1)} + \ldots + \beta_p x^{(p)} + \varepsilon, \\ y &= \beta_0 + \beta_1 x^{(1)} + \ldots + \beta_p x^{(p)} + \beta_{p+1} x^{(p+1)} + \ldots + x^{(p+q)} + \varepsilon.\end{aligned}$$

In the complete model, we have q additional predictors.

Then it is reasonable to define a statistic that compares these two models. We could define it as

$$\frac{SSE_R - SSE_C}{SSE_R},$$

where SSE_R and SSE_C are the sums of the squared errors computed, respectively, for the reduced and complete models. This would be in complete analogy to the definition of R^2 above. However, traditionally another form of the statistic is preferred, namely:

$$F = \frac{(SSE_R - SSE_C)/q}{SSE_C/(n-p-q-1)},$$

for the reason that it is useful for testing. Indeed, under the null hypotheses, that the reduced model is correct this statistic is distributed according to the Fisher distribution with q and $n-p-q-1$ degrees of freedom.

In particular the null hypothesis can be rejected at α significance level if $F > F_\alpha$. Intuitively, the reduction in the size of errors, as measured by $SSE_R - SSE_C$ is too large to be explained by pure chance.

Note, however, that the statement that F follows the Fisher distribution assumes that the errors are normal and it is quite sensitive to this assumption. (In other words, if the errors are not normal it can happen that the probability of type I error is different from α, in particular it can happen that we reject the null hypothesis too often.)

Chapter 7

Categorical data

7.1 Experiment

Here we discuss the experiment in which there are a finite number of outcomes. So we have n observation and each observation Y_i, $i = 1, \dots, n$, can belong to one of k possible categories.

You should recognize here the multinomial experiment with n trials and k possible outcomes. We assume here that there are no predictors x_i. In this case, the result of this experiment can be conveniently summarized by a list of counts. So n_j is the number of y_i that happened to be in j-th category.

Clearly we have $n_1 + \ldots + n_k = n$.

The model that we use assumes that all y_i are independent, so the counts n_j, $j = 1, \dots, k$ follow the multinomial distribution:

$$\mathbb{P}(X_1 = n_1, \dots, X_k = n_k) = \binom{n}{n_1, n_2, \dots, n_k} p_1^{n_1} p_2^{n_2} \dots p_k^{n_k},$$

where $p_1, \dots p_k$ are unknown parameters, and we used notation X_j to denote the random number of items in category j.

We are interested in statistical inferences about these parameters. For example, we might be interested if they are equal to some specific values.

The exact inference is somewhat complicated here and we are going to do the asymptotic inference which gives a good approximation if n is sufficiently

large. This can be done by using the Pearson statistic and χ^2 goodness of fit test.

7.2 Pearson's χ^2 test

Our goal is to test that the probabilities of the categories are $p_1, \ldots, p_k$. The key is the following statistic:

$$T = \sum_{j=1}^{k} \frac{(X_j - \mathbb{E}X_j)^2}{\mathbb{E}X_j} = \sum_{j=1}^{k} \frac{(n_j - np_j)^2}{np_j}$$

Pearson proved that if the null hypothesis is correct (that is, if the data indeed came from the multinomial distribution with the specified probabilities), and if n is large, then this statistic is approximately distributed as χ^2 random variable with $k-1$ degrees of freedom.

Usually the one-sided test is used so that the null hypothesis is rejected if $T > \chi^2_{\alpha,k-1}$. Alternatively, one can calculate the p-value as

$$p - \text{value} \approx \mathbb{P}(\xi^2_{k-1} \geq T$$

and compare it with α.

A usual rule of thumb is that the approximation is good if $np_j > 1$ for each n and that at least for 80% of j this product is greater than 5.

A computationally easier formula can be obtained by expanding the squares.

$$T = \sum_{j=1}^{k} \frac{n_j^2}{np_j} - n$$

Example 7.2.1 (From Ross). If a famous person is in poor health and dying, then perhaps anticipating his birthday would "cheer him up and therefore improve his health and possibly decrease the chance that he will die shortly before his birthday." The data might therefore reveal that a famous person is less likely to die in the months before his or her birthday and more likely to die in the months afterward.

To test this, a sample of 1,251 (deceased) Americans was randomly chosen from Who Was Who in America, and their birth and death days were noted. (The data are taken from D. Phillips, "Death Day and Birthday: An Unexpected Connection," in Statistics: A Guide to the Unknown, Holden-Day, 1972.)

We choose the following categories: outcome 1 = "6, 5, 4 months before death" outcome 2 = "3, 2, 1' months before" outcome 3 = "0, 1, 2 months after death", outcome 4 = "3, 4, 5 months after death".

Outcome	Number of times occurring
1	277
2	283
3	358
4	333

The null hypothesis is that all these outcomes have equal probabilities $p_i = 1/4$ for all $i = 1, \ldots, 4$. So we calculate the Pearson statistic:

$$T = \frac{277^2 + 283^2 + 358^2 + 333^2}{1251/4} - 1251 = 14.775$$

The corresponding p-value is

$$p - \text{value} \approx \mathbb{P}(\xi_3^2 \geq 14.775) = 0.002,$$

so it appears that the null hypothesis can be rejected.

7.3 Goodness of fit tests when parameters are unspecified

The Pearson test can be applied to test whether the data came from a specified family of distribution. For example we might be interested if the data came from a Poisson, geometric, or some other more complicated distribution.

In this situation, the true values of p_i depend on other parameters, which must be estimated from data. Somewhat surprisingly, the Pearson test can still be applied if the parameters are estimated using the maximum likelihood

method. Namely, if p_i depend on m parameters and if they were estimated by maximum likelihood, then the statistic

$$T = \sum_{j=1}^{k} \frac{(n_j - n\widehat{p}_j)^2}{n\widehat{p}_j}$$

is distributed according to χ^2 distribution with $k-1-m$ degrees of freedom

Example 7.3.1. Suppose we have data on the number of accidents on an industrial plant over 30-week period:

$$\begin{bmatrix} 8 & 0 & 0 & 1 & 3 & 4 & 0 & 2 & 12 & 5 \\ 1 & 8 & 0 & 2 & 0 & 1 & 9 & 3 & 4 & 5 \\ 3 & 3 & 4 & 7 & 4 & 0 & 1 & 2 & 1 & 2 \end{bmatrix}$$

We want to test the hypothesis that number of accidents follows the Poisson distribution.

We can divide the data in 5 categories: the outcome of the number of accidents in a given week is in category 1 if there are 0 accidents, in category 2 if there is 1 accident, in category 3 if there are 2 or 3 accidents, in category 4 if there are 4 or 5 accidents, and in category 5 if there are more than 5 accidents. Then, if the parameter λ of the Poisson distribution were known, we could calculate the probability of each category:

$$\begin{aligned}
p_1 &= \mathbb{P}(Y = 0) = e^{-\lambda}, \\
p_2 &= \mathbb{P}(Y = 1) = \lambda e^{-\lambda}, \\
p_3 &= \mathbb{P}(Y = 2) + \mathbb{P}(Y = 3) = \frac{\lambda^2 e^{-\lambda}}{2} + \frac{\lambda^3 e^{-\lambda}}{6}, \\
p_4 &= \ldots, \\
p_5 &= \mathbb{P}(Y > 5) = 1 - e^{-\lambda} - \lambda e^{-\lambda} - \frac{\lambda^2 e^{-\lambda}}{2} - \frac{\lambda^3 e^{-\lambda}}{6} - \frac{\lambda^4 e^{-\lambda}}{24} - \frac{\lambda^5 e^{-\lambda}}{120}.
\end{aligned}$$

However, since it is unknown, it must be estimated from the data. The maximum likelihood estimator here is the same as the method of moments estimator:

$$\widehat{\lambda} = \overline{Y} = \frac{95}{30} = 3.16667.$$

Then a computation gives

$$\widehat{p}_1 = .04214,$$
$$\widehat{p}_2 = .13346,$$
$$\widehat{p}_3 = .43434,$$
$$\widehat{p}_4 = .28841,$$
$$\widehat{p}_5 = .10164$$

The category counts are $n_1 = 6$, $n_2 = 5$, $n_3 = 8$, $n_4 = 6$, $n_5 = 5$ and we can calculate the statistic value as

$$T = \sum_{j=1}^{k} \frac{(n_j - 30\widehat{p}_j)^2}{30\widehat{p}_j} = 21.99156$$

The number of degrees of freedom is $k - m - 1 = 5 - 1 - 1 = 3$. Then the p-value is

$$p - \text{value} \approx \mathbb{P}(\xi_3^2 \geq 21.99) = 0.000064,$$

and the hypotheses that the distribution is Poisson can be rejected.

7.4 Independence test for contingency tables

In this section we will talk about another application of the Pearson test.

We suppose now that every observation in the data has 2 characteristics, X and Y. Both are categorical, so X can take one of r values, which we code as $i = 1, \ldots, r$ and Y can take one of s values, $j = 1, \ldots, s$.

We want to check if Y depends on X. More specifically, we want to test they hypothesis that random variables Y and X are independent.

We denote the joint pmf of X and Y as p_{ij}:

$$p_{ij} := \mathbb{P}[X = i, Y = j],$$

Then the marginal pmfs of X and Y are

$$p_i := \mathbb{P}[X = i] = \sum_{j=1}^{s} p_{ij},$$
$$q_j := \mathbb{P}[Y = j] = \sum_{i=1}^{r} p_{ij},$$

and the null hypothesis states that

$$p_{ij} = p_i q_j,$$

for all i and j.

Now assume that the data also presented in the form of counts. Namely, let n_{ij} be the number of observations that have X-characterisitic equal to i and Y-characteristic equal to j. (Usually these numbers are organized as a table which is called the contingency table.)

We can also define n_i as the number of observations with X-characteristic equal to i and m_j as the number of observations with Y-characteristic equal to j. Clearly,

$$n_i := \sum_{j=1}^{s} n_{ij},$$
$$m_j := \sum_{i=1}^{r} p_{ij}.$$

If we knew p_i and q_j, then we could write the Pearson statistic as

$$T = \sum_{i,j} \frac{(n_{ij} - np_iq_j)^2}{np_iq_j}.$$

However, since we don't know these parameters we need to estimate them from data. Natural estimates are

$$\widehat{p}_i = \frac{n_i}{n} \text{ and } \widehat{q}_j = \frac{m_j}{n},$$

and then we can write the Pearson statistic as

$$T = \sum_{i,j} \frac{(n_{ij} - n\widehat{p}_i\widehat{q}_j)^2}{n\widehat{p}_i\widehat{q}_j} = \sum_{i,j} \frac{n_{ij}^2}{n_i n_j/n} - n.$$

If n is large, then the distribution of T approximately equals the χ^2 distribution. As usual the number of degrees of freedom equal the number of categories minus 1, $rs-1$, reduced by the number of estimations performed. A calculation shows that the number of calculations needed is $(r-1)+(s-1)=r+s-2$ and so the number of degrees of freedom is

$$df = rs - 1 - (r+s-2) = (r-1)(s-1).$$

So, for a given significance level α, the null hypothesis should be rejected if

$$T > \chi^2_{\alpha,(r-1)(s-1)}.$$

The p - value can be calculated as probability

$$p - \text{value} = \mathbb{P}(\chi > T),$$

where χ is a random variable that have the χ^2 distribution with $(r-1)(s-1)$ degrees of freedom. In R, this can be computed as

$$p - \text{value} = pchisq(T, df = (r-1)(s-1), lower.tail = F).$$

Example 7.4.1. A sample of 300 people was randomly chosen, and the sampled individuals were classified as to their gender and political affiliation, Democrat, Republican, or Independent. The following table displays the resulting data.

	Democrat	Republican	Independent	Total
Women	68	56	32	156
Men	52	72	20	144
Total	120	128	52	300

The null hypothesis is that a randomly chosen individual's gender and political affiliation are independent.

We calculate:

$$n\widehat{p}_1\widehat{q}_1 = \frac{n_1 m_1}{n} = \frac{156 \times 120}{300} = 62.40$$
$$n\widehat{p}_1\widehat{q}_2 = \frac{n_1 m_2}{n} = \frac{156 \times 128}{300} = 66.56$$
$$n\widehat{p}_1\widehat{q}_3 = \frac{n_1 m_3}{n} = \frac{156 \times 52}{300} = 27.04$$
$$n\widehat{p}_2\widehat{q}_1 = \frac{n_2 m_1}{n} = \frac{144 \times 120}{300} = 57.60$$
$$n\widehat{p}_2\widehat{q}_2 = \frac{n_2 m_2}{n} = \frac{144 \times 128}{300} = 61.44$$
$$n\widehat{p}_2\widehat{q}_3 = \frac{n_2 m_3}{n} = \frac{144 \times 52}{300} = 24.96$$

Therefore, the test statistic is

$$TS = \frac{68^2}{62.40} + \frac{56^2}{66.56} + \frac{32^2}{27.04} + \frac{52^2}{57.60} + \frac{72^2}{61.44} + \frac{20^2}{24.96} - 300 = 6.432857$$

The number of degrees of freedom is $(3-1)(2-1) = 2$, so we look at the χ^2 distribution with 2 degrees of freedom. At $\alpha = 0.05$ we have

$$\chi^2_{0.05,2} = 5.991,$$

so we can reject the null hypothesis that gender and political affiliation are independent. We can calculate p - value as

$$pchisq(6.433, df = 2, lower.tail = F) \approx 0.04,$$

so at $\alpha = 0.01$ we cannot reject this hypothesis.

In R if we have two variables X and Y as factors in a database "data", then one can build the contingency table by using the function "table()":

$$tb \leftarrow table(data\$X, data\$Y),$$

and the Pearson table of independence can be done using function "chisq.test()":

$$chisq.test(tb)$$

Chapter 8

Bayesian Inference

8.1 Estimation

The Bayesian inference is a collection of statistical methods based on a different statistical philosophy. The statistical model is still consist of observations $X_1, \ldots, X_n$ which are random with the distribution $f(\vec{X}|\theta)$ that depend on a vector of parameters θ. However, while the classical statistic treats the parameters as fixed and unknown quantities, the Bayesian statistic models researchers' beliefs about the parameters by using probability theory. This adds a second layer of randomness: now the parameters θ of the data-generating distribution $f(\vec{X}|\theta)$ have their own probability distributions which model our beliefs about them.

In fact, the parameters are treated as random variables that have two probability distributions: before and after the data is observed. Their distribution before the data is observed is described by a *prior distribution* with density (or mass) function $p(\theta)$. The *posterior distribution* is the distribution of parameters after the data is observed. It captures our beliefs after they were modified by the observed data. The density (or mass) function of the posterior distribution is the conditional density $p(\theta|x_1, \ldots, x_n)$. It can be calculated from the prior distribution and the data by using Bayes' formula:

$$p(\theta|x_1,\ldots,x_n) = \frac{f(x_1,\ldots,x_n|\theta)p(\theta)}{\int f(x_1,\ldots,x_n|\varphi)p(\varphi)\,d\varphi}.$$

The integral in the denominator is a normalizing constant – it does not depend on θ. Often, it is not written explicitly, and the formula is written as

$$p(\theta|x_1,\ldots,x_n) \propto f(x_1,\ldots,x_n|\theta)p(\theta).$$

The symbol $\propto$ means "proportional to".

Example 8.1.1. A biased coin is tossed n times. Let X_i be 1 or 0 as the i-th toss is or is not a head. The probability of a head is θ. Suppose we have no idea how biased the coin is, and we place a uniform prior distribution on θ, to give a so-called "non-informative prior" of

$$p(\theta) = 1, \quad 0 \le \theta \le 1.$$

Let t be the number of heads. Then, the posterior distribution of θ is

$$p(\theta|x_1,\ldots,x_n) \propto \theta^t(1-\theta)^{n-t}$$

By inspection we realize that if the appropriate constant on the right-hand side, then we have the density of the Beta distribution with parameters $(t+1, n-t+1)$. This is the posterior distribution of θ given x.

After a bit of reflection, we realize that if we start with the prior distribution which is the beta distribution with parameters α_1, α_2, that is, if $p(\theta) \propto \theta^{\alpha_1-1}(1-\theta)^{\alpha_2-1}$, the the posterior distribution is

$$p(\theta|x_1,\ldots,x_n) \propto \theta^{t+\alpha_1-1}(1-\theta)^{n-t+\alpha_2-1},$$

that is, it is still the beta distribution but with updated parameters $t+\alpha_1$ and $n-t+\alpha_2$. Note that here α_1 and α_2 are parameters for the prior distribution of the parameter θ! Sometimes they are called hyper-parameters.

Definition 8.1.2. Let $f(x|\theta)$ be the distribution of data point x given the parameter θ. Let $p(\theta)$ be a prior distribution for the parameter. If the posterior distribution $p(\theta|x_1,\ldots,x_n)$ has the same functional form as the prior but with altered parameter values, then the prior $p(\theta)$ is said to be *conjugate* to the distribution $f(x|\theta)$.

The conjugate priors are very convenient in modeling and are often used in practice. Here is another example.

Exercise 8.1.3. Suppose $X_1, \ldots, X_n$ are normally distributed $X_i \sim N(\mu, \sigma)$ with unknown parameter μ and known $\sigma = 1$. Let the prior distribution for μ be $N(0, \tau^{-2})$ for known τ^{-2}. Then the posterior distribution for μ is

$$N\Big(\frac{\sum_{i=1}^n x_i}{n+\tau^2}, \frac{1}{n+\tau^2}\Big).$$

In many cases, in practical applications we need a point estimator and not a posterior distribution of the parameter. Bayesian statistic addresses this concern with the concept of the loss function. A *loss function* $L(\theta, a)$ is a measure of the loss incurred by estimating the value of the parameter to be a when its true value is θ. The estimator $\widehat{\theta}$ is chosen to minimize the expected loss $\mathbb{E}L(\theta, \widehat{\theta})$, where the expectation is taken over θ with respect to the posterior distribution $p(\theta|\vec{x})$,

$$\widehat{\theta} = \arg\min_a \mathbb{E}\big[L(\theta, a)\big]$$

Theorem 8.1.4. *(a) Suppose that the loss function is quadratic in error: $L(\theta, a) = (\theta - a)^2$. Then the expected loss is minimized by taking $\widehat{\theta}$ to be the mean of the posterior distribution:*

$$\widehat{\theta} = \int \theta p(\theta|x_1, \ldots, x_n)\, d\theta.$$

(b) Suppose the loss function is the absolute value of the error: $L(\theta, a) = |\theta - a|$. Then the expected loss is minimized by taking $\widehat{\theta}$ to be the median of the posterior distribution.

Example 8.1.5 (Coin Tosses). Consider the setting of Example 8.1.1. The posterior distribution is Beta distribution with parameters $t+1$ and $n-t+1$. So, by properties of Beta distribution, the *posterior mean* estimator is

$$\widehat{\theta} = \frac{t+1}{n+2},$$

and the *posterior median* estimator needs to be calculated numerically. Note that both are different from the standard estimator $\overline{x} = t/n$.

Note that both posterior mean and posterior median estimators depends on our choice of the prior distribution.

The Bayesian analogue of confidence intervals is *credible sets*. A set A is a *credible set* with probability $1-\alpha$ if the probability that the parameter belongs to the set is α when the probability is calculated with respect to the posterior distribution. That is,

$$\mathbb{P}\theta \in A = \int_A p(\theta|x_1, \ldots, x_n)\, d\theta = 1 - \alpha.$$

Calculation of the credible sets is straightforward when we know the posterior distribution. It is important to understand that confidence sets depend on the choice of prior distribution.

8.2 Hypothesis testing

Bayesian inference is also different in its approach to hypothesis testing from the approach of the classical statistics. In fact, the hypothesis testing plays less significant role in Bayesian inference simply because the idea that a continuous parameter can be for sure equal to a specific value is at odds with main idea of Bayesian inference to model beliefs about parameters with probabilities. Still it is possible to evaluate two competing hypothesis using Bayesian methods.

In the classical case, the setup consists of a pair of hypotheses: the null hypothesis H_0 and the alternative hypotheses H_a. The rejection region is selected as a set of data samples for which we reject of H_0 in favor of H_a. This set is selected in such a way that the probabilities of making the wrong decisions, - the probabilities of type I and type II errors, α and β, – are small.

The deficiency of this method is that it works really well only if both the null and alternative are simple hypothesis, so that we can calculate α and β unambiguously. In this case, the Neyman-Pearson lemma provides us with the most powerful test, that is the test, that has the smallest β for a given α. This test is based on the ratio of likelihood functions, that is on the ratio

of data densities for the parameters θ_0 and θ_a:

$$\frac{L(\theta_0|x_1,\ldots,x_n)}{L(\theta_a|x_1,\ldots,x_n)} \equiv \frac{f(x_1,\ldots,x_n|\theta_0)}{f(x_1,\ldots,x_n|\theta_a)}.$$

Sometimes, it is possible to develop the best test (Uniformly Most Powerful Test) even when the alternative hypothesis is composite. However, in many cases, for example, for the two-sided alternative hypothesis there are no UMP tests. In addition, in order to search for UMP tests, we have to assume that the null hypothesis is simple. This is somewhat unsatisfactory since in many practical situations it is difficult to justify a specific value for the null hypothesis.

In practice, statisticians are satisfied with reasonable, although not UMP, tests, which would allow us to do testing even if the null hypothesis is not simple. One of this tests, the likelihood ratio test is based on the test statistic:

$$\lambda(x_1,\ldots,x_n) = \frac{\max_{\theta\in\Omega_0} L(\theta|x_1,\ldots,x_n)}{\max_{\theta\in\Omega_0\cup\Omega_a} L(\theta|x_1,\ldots,x_n)} \equiv \frac{\max_{\theta\in\Omega_0} f(x_1,\ldots,x_n|\theta)}{\max_{\theta\in\Omega_0\cup\Omega_a} f(x_1,\ldots,x_n|\theta)}$$

In other words, we choose the value θ_0 in the null hypothesis set Ω_0, which gives the largest probability density of the data, and compare this density with the maximum of the data probability density when the parameter is allowed to vary over both the null (Ω_0) and alternative (Ω_a) hypothesis sets. We reject the null if the ratio of these two probabilities is smaller than a threshold.

While this procedure is very reasonable, it does not have a clear probabilistic justification.

In contrast, in the Bayesian inference, the null hypothesis is typically not a single value but a big set of parameters $H_0 : \theta \in \Omega_0$ and the alternative is the complement of this set, $H_a : \theta \in \Omega_a = \Omega_0^c$. For example, we can have the null hypothesis $H_0 : \theta \leq \theta_0$ and the alternative $H_a : \theta > \theta_0$.

The null hypothesis is rejected by the Bayesian test, if the ratio of pos-

terior probabilities of hypotheses:

$$\lambda_B(x_1,\dots,x_n) = \frac{\mathbb{P}[\theta\in\Omega_0]}{\mathbb{P}[\theta\in\Omega_a]} = \frac{\int_{\Omega_0} p(\theta|x_1,\dots,x_n)\,d\theta}{\int_{\Omega_a} p(\theta|x_1,\dots,x_n)\,d\theta} = \frac{\int_{\Omega_0} f(x_1,\dots,x_n|\theta)p(\theta)\,d\theta}{\int_{\Omega_a} f(x_1,\dots,x_n|\theta)p(\theta)\,d\theta}$$

is smaller than a certain threshold t, which measures the degree of our conservatism. For example, if the threshold is set to $1/3$, then we reject the null hypothesis H_0 only if the posterior probability of H_0 is three times smaller than the posterior probability of H_a.

This resembles the likelihood ratio test, except instead of maximizing the data density over the set of parameters Ω_0 and Ω_a, we take the average of the data densities by using the prior probability distribution $p(\theta)$.

It is in principle possible to define α and β of the Bayesian test as the average probabilities of making type I and type II errors, where the average is calculated with respect to the prior distribution. However, the definition is more complicated. In addition, the Bayesian analysis is most useful when the amount of the data is not overwhelmingly large compared to our prior beliefs. In this situation, there is no analogue of Wilkes' theorem for the likelihood ratio, and so it is significantly more difficult to develop a test with a given α. For this reason α and β are very rarely used in Bayesian inference.

Here is an example, how a Bayesian test applies in practice.

Example 8.2.1. Let $X_1,\dots,X_n$ be a sample form an exponentially distributed population with density $f(x|\theta) = \theta e^{-\theta x}$. (Note that this is a slightly different parameterization of the exponential distribution. The mean of the distribution is $\mu = 1/\theta$. Suppose the prior distribution is Gamma distribution with parameters α and β. Test the hypothesis $H_0 : \theta \le \theta_0$ versus $H_a : \theta > \theta_0$.

The density of the data sample is

$$f(x_1,\dots,x_n|\theta) = \theta^n e^{-\theta\sum_{i=1}^n x_i}.$$

The prior is

$$p(\theta) \propto \theta^{\alpha-1} e^{-\theta/\beta}.$$

The posterior distribution is

$$p(\theta|x_1, \ldots, x_n) \propto \theta^{n+\alpha-1} e^{-\theta(\sum_{i=1}^n x_i + 1/\beta)}$$

So the posterior distribution is the Gamma distribution with parameters

$$\begin{aligned} \alpha' &= n + \alpha, \\ \beta' &= \frac{1}{\sum_{i=1}^n x_i + 1/\beta} = \frac{\beta}{\beta \sum_{i=1}^n x_i + 1} \end{aligned}$$

In particular we showed that the Gamma distribution is the conjugate prior for the exponential distribution. We reject the null hypothesis only if

$$\mathbb{P}[\theta \leq \theta_0] < \frac{t}{1+t}.$$

In R, this can be solved by checking if

$$\texttt{pgamma}(\theta_0, \alpha', 1/\beta') < \frac{t}{1+t}.$$

For example, let $n = 10$, $\sum x_i = 1.26$, $\alpha = 3$ and $\beta = 5$. (These are perhaps are obtained by reviewing prior studies about θ.) We want to test the null hypothesis that $H_0 : \mu > .12$ against $H_a : \mu \leq .12$ using $t = 1$ (This not a very conservative test. We reject null hypothesis if its probability is smaller than the probability of the alternative.) In terms of the parameter θ, the hypotheses are

$$H_0 : \theta < 1/(.12) = 8.333 \text{ against } H_0 :\geq 8.333.$$

We calculate

$$\begin{aligned} \alpha' &= n + \alpha = 10 + 3 = 13 \\ 1/\beta' &= \sum_{i=1}^n x_i + 1/\beta = 1.26 + 1/5 = 1.46, \end{aligned}$$

then

$$\texttt{pgamma}(\theta_0, \alpha', 1/\beta') = \texttt{pgamma}(8.333, 13, 1.46) = 0.4430332$$

Since this is smaller than $1/(1+t) = 1/2$, we can reject the null hypothesis.

One observation about the Bayesian hypothesis testing is that the results of the tests depend on the choice of the prior distribution and this choice should be careful and well-justified. The second observation is that in practice it is sometimes difficult to calculate the probabilities under the posterior distribution. This calculation may involve difficult integrations. In this respect, the classical approach is often computationally easier.