Chapter 3: A Review of Statistics*

Manuel V. Montesinos†

Statistics Brush-Up Course
Competition and EPP Master Programs

Fall 2020

After having refreshed some of the most important tools for describing data and explaining how these can be generated from a population, the goal of this chapter is to "invert" the process and learn how data can be used to make inference about the characteristics of the population. For that purpose, we will first study the easiest way of obtaining data, which is called ***random sampling***. Once we have data, we will be able to characterize the ***sampling distribution*** of any statistic of interest, such as the sample mean. We will learn how to do that for any sample size and, if this is not possible, how to obtain an approximation for large sample sizes. Next, we will focus on three types of statistical methods: estimation, hypothesis testing, and confidence intervals. ***Estimation*** entails computing a "best guess" numerical value for an unknown characteristic of a population distribution, such as its mean, from a sample of data. ***Hypothesis testing*** involves formulating a specific hypothesis about the population and then using sample evidence to decide whether it is true. ***Confidence intervals*** use the available data to estimate an interval or range for an unknown population charactetistic. We will conclude the chapter by providing a brief introduction of ***linear regression analysis*** with one regressor.

I. Random Sampling and the Distribution of the Sample Average

Many statistical and econometric procedures involve computing averages or weighted averages of a sample of data. Characterizing the distribution of sample averages

*These notes are partially based on James H. Stock and Mark W. Watson's textbook *Introduction to Econometrics*; Christoph Hanck, Martin Arnold, Alexander Gerber and Martin Schmelzer's textbook *Introduction to Econometrics with R*, and the Probability and Statistics courses taught by Joan Llull, Jordi Caballé and Anna Houštecká at the Universitat Autònoma de Barcelona and the Barcelona GSE. Typos, misprints, misconceptions and other errors are all mine.

†Departament d'Economia i d'Història Econòmica. Universitat Autònoma de Barcelona. Campus de Bellaterra – Edifici B, 08193, Bellaterra, Cerdanyola del Vallès, Barcelona (Spain). E-mail: manuel.montesinos@barcelonagse.eu.

therefore is an essential step toward understanding the performance of econometric procedures. In this section we will review some basic concepts about random sampling and the distributions of averages. We begin by discussing ***random sampling***. The act of random sampling, that is, randomly drawing a sample from a larger population, has the effect of making the sample average itself a random variable, and as such, it has a probability distribution.

A. Simple random sampling

To clarify the basic idea of random sampling, let us jump back to the dice rolling example. Let us assume that we are rolling a dice n times. This means that we are interested in the outcomes of the random variables Y_i, $i = 1, ..., n$ which are characterized by the same distribution. Since these outcomes are selected randomly, they are random variables themselves and their realizations will differ each time we draw a sample, i.e., each time we roll the dice n times. Furthermore, each observation is randomly drawn from the same population, that is, the numbers from 1 to 6, and their individual distribution is the same. Hence $Y_1, ..., Y_n$ are identically distributed.

Moreover, we know that the value of any of the Y_i does not provide any information about the remainder of the outcomes. In our example, rolling a 6 as the first observation in our sample does not alter the distributions of $Y_2, ..., Y_n$: all numbers are equally likely to occur. This means that all Y_i are also independently distributed. Then, we say that $Y_1, ..., Y_n$ are independently and identically distributed (***i.i.d.***).

A situation like this is called ***simple random sampling***. In simple random sampling, n objects are drawn at random from a population, and each object is equally likely to end up in the sample. We denote the value of the random variable Y for the ith randomly drawn object as Y_i. Since all objects are equally likely to be drawn and the distribution of Y_i is the same for all i, then $Y_i, ..., Y_n$ are independently and identically distributed. This means that the marginal distribution of Y_i is the same for all $i = 1, ..., n$ (*identically distributed*), and Y_1 is distributed *independently* of $Y_2, ..., Y_n$, Y_2 is distributed independently of $Y_1, Y_3, ..., Y_n$, and so forth.

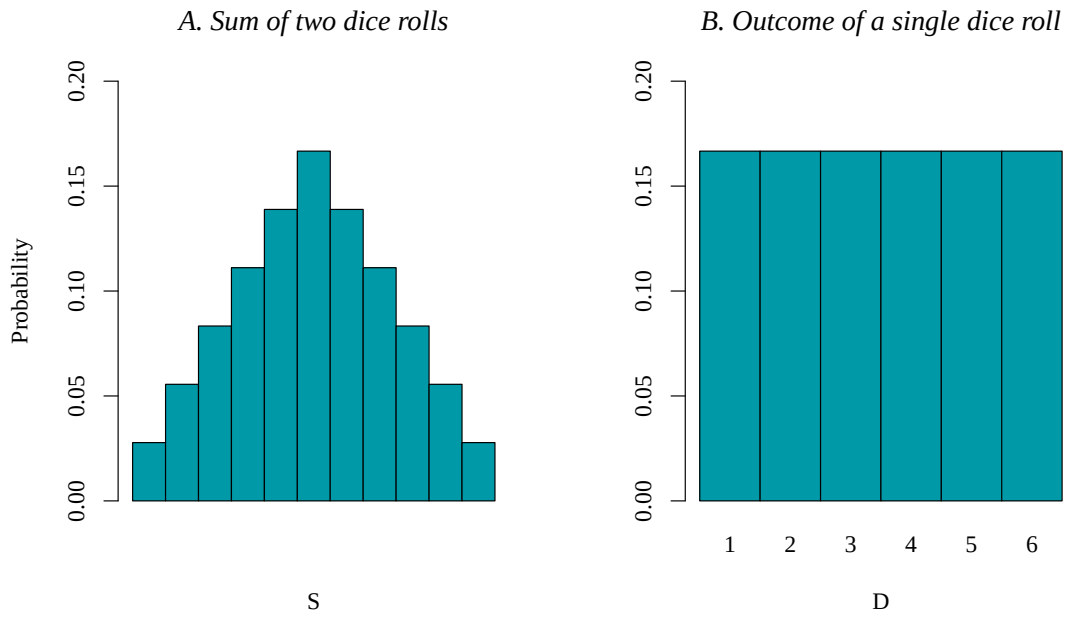
Table 1: Probability Distribution of the Sum of Two Dice Rolls

Outcome:	2	3	4	5	6	7	8	9	10	11	12
Probability:	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Now, what happens if we consider functions of the sample data? Let us consider the example of rolling a dice twice in a row. A sample now consists of two indepen-

dent random draws from the set $\{1, 2, 3, 4, 5, 6\}$. It is apparent that any function of these two random variables, e.g., their sum, is also random. Let us call this sum S. Possible outcomes of S are $\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$, and the probability distribution of S is in Table 1, which differs from the marginal distribution of a single dice roll's outcome D. We compare these two in Figure 1.

FIGURE 1: PROBABILITY DISTRIBUTIONS OF THE SUM OF TWO DICE ROLLS AND A SINGLE DICE ROLL



B. Sampling distribution of the sample average

The ***sample average*** or ***sample mean***, $\overline{Y}$, of the n observations $Y_1, ..., Y_n$ is

$$\overline{Y} = \frac{1}{n}(Y_1 + Y_2 + ... + Y_n) = \frac{1}{n}\sum_{i=1}^{n} Y_i. \tag{1}$$

An essential concept is that the act of drawing a random sample has the effect of making the sample average $\overline{Y}$ a random variable. Because the sample was drawn at random, the value of each Y_i is random, and so is the average of $Y_1, ..., Y_n$. Had a different sample been drawn, then the observations and their sample average would have been different: the value of $\overline{Y}$ differs from one random sample to the next.

Therefore, $\overline{Y}$ has a probability distribution. The distribution of $\overline{Y}$ is called the ***sampling distribution*** of $\overline{Y}$ because it is the probability distribution associated with possible values of $\overline{Y}$ that could be computed for different possible samples

$Y_1, ..., Y_n$.

Now, let us assume that the observations $Y_1, ..., Y_n$ are i.i.d. and let μ_Y and σ_Y^2 denote the mean and variance of Y_i (because the observations are i.i.d., the mean and variance is the same for all $i = 1, ..., n$). When $n = 2$, the mean of the sum $Y_1 + Y_2$ is given by $\mathbb{E}[Y_1 + Y_2] = \mu_Y + \mu_Y = 2\mu_Y$. Thus the mean of the sample average is $\mathbb{E}\left[\frac{1}{2}(Y_1 + Y_2)\right] = \mu_Y$. In general,

$$\mathbb{E}\left[\overline{Y}\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Y_i] = \mu_Y. \tag{2}$$

As for the variance of $\overline{Y}$, let us consider again $n = 2$, so that $\text{Var}\left[Y_1 + Y_2\right] = 2\sigma_Y^2$ and $\text{Var}\left[\overline{Y}\right] = \sigma_Y^2/2$. In general, because $Y_1, ..., Y_n$ are i.i.d., Y_i and Y_j are independently distributed for $i \neq j$, so $\text{Cov}\left(Y_i, Y_j\right) = 0$. Thus,

$$\begin{aligned}\text{Var}\left[\overline{Y}\right] &= \text{Var}\left[\frac{1}{n}\sum_{i=1}^{n}Y_i\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}[Y_i] + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\text{Cov}\left(Y_i, Y_j\right) = \frac{\sigma_Y^2}{n}.\end{aligned} \tag{3}$$

Equivalently, we can write

$$\text{Var}\left[\overline{Y}\right] = \text{Var}\left[\frac{1}{n}\sum_{i=1}^{n}Y_i\right] = \text{Var}\left[\frac{1}{n}Y_1 + ... + \frac{1}{n}Y_n\right] = \frac{1}{n^2}n\sigma_Y^2 = \frac{\sigma_Y^2}{n}. \tag{4}$$

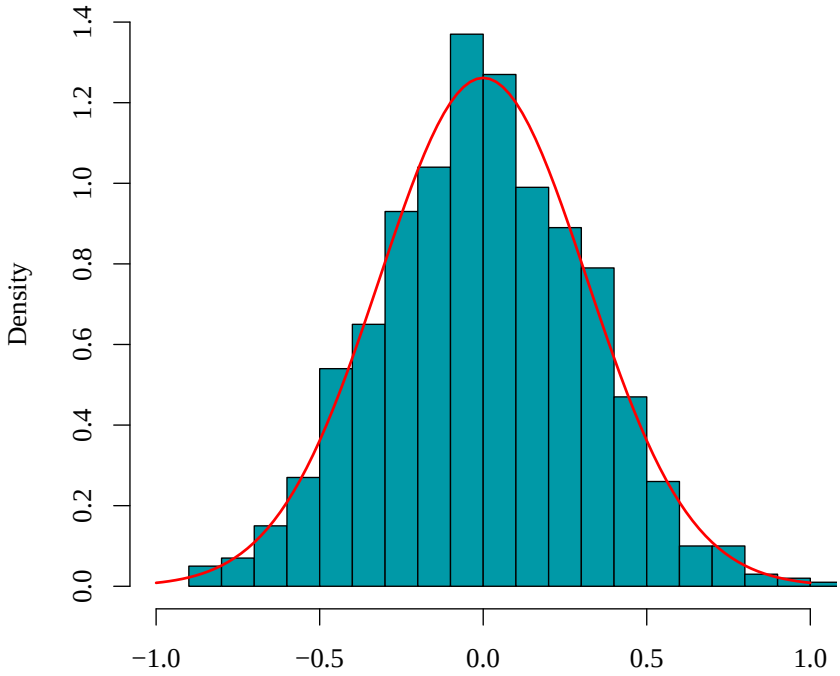
The standard deviation of $\overline{Y}$ is the square root of the variance, $\sigma_Y/\sqrt{n}$.

These results hold whatever the distribution of Y_i is; that is, the distribution of Y_i does not need to take on a specific form, such as the normal distribution, for the above results to hold. For instance, if $Y_1, ..., Y_n$ are i.i.d. draws from the $\mathcal{N}(\mu_Y, \sigma_Y^2)$ distribution, then $\overline{Y}$ is distributed $\mathcal{N}(\mu_Y, \sigma_Y^2/n)$.

We can verify this result by repeatedly drawing random samples of n observations from the $\mathcal{N}(\mu_Y, \sigma_Y^2)$ distribution and computing their respective averages. If we do it for a large number of repetitions, the simulated dataset of averages should quite accurately reflect the theoretical distribution of $\overline{Y}$ if the theoretical result holds.[1] Figure 2 represents the outcomes of this experiment for $Y \sim \mathcal{N}(0, 1)$, where $\overline{Y} \sim \mathcal{N}(0, 0.1)$.

[1] This approach is commonly known as a Monte Carlo Experiment.

FIGURE 2: HISTOGRAM OF $\overline{Y}$ WHEN $Y \sim \mathcal{N}(0,1)$



II. Large-Sample Approximations to Sampling Distributions

Sampling distributions as considered in the previous section play an important role in the development of econometric methods, so it is important to know how to characterize these distributions. There are two main approaches to do that: an *exact* approach and an *approximate* approach.

The *exact* approach entails deriving a formula for the sampling distribution that holds exactly for any sample size n. The sampling distribution that exactly describes the distribution of $\overline{Y}$ for any n is called the ***exact distribution*** or ***finite-sample distribution*** of $\overline{Y}$. In the previous examples of dice rolling and normal variates, we have dealt with functions of random variables whose sample distributions are exactly known in the sense that we can write them down as analytic expressions. However, this is not always possible. As we said before, normality of Y_i implies normality of $\overline{Y}$, but the exact distribution of $\overline{Y}$ is generally unknown and often hard to derive (or even untraceable) if we drop the assumption that Y_i has a normal distribution.

By contrast, the *approximate* approach uses approximations to the sampling distribution that rely on the sample size being large. The large-sample approximation to the sampling distribution is often called the ***asymptotic distribution***. We call it "asymptotic" because they become exact in the limit when $n \to \infty$. These approximations can be very accurate even if the sample size is only $n = 30$ observations. Because sample sizes used in practice in econometrics typically number in the hundreds or thousands, these asymptotic distributions can be counted on to provide very good approximations to the exact sampling distribution.

In this section we will discuss two well-known results used to approximate sampling distributions when the sample size is large: the *law of large numbers* and the *central limit theorem*. The law of large numbers states that in large samples, $\overline{Y}$ will be close to μ_Y with very high probability. The central limit theorem says that, when the sample size is large, the sampling distribution of the standardized sample average, $(\overline{Y} - \mu_Y)/\sigma_{\overline{Y}}$, is approximately normal.

Although exact sampling distributions are complicated and depend on the distribution of Y, the asymptotic distributions are simple. Moreover, the asymptotic normal distribution of $(\overline{Y} - \mu_Y)/\sigma_{\overline{Y}}$ does not depend on the distribution of Y. This normal approximate distribution simplify the development and applicability of econometric procedures enormously, and is a key component underlying the theory of statistical inference for regression models.

A. The law of large numbers and consistency

The ***law of large numbers*** states that, under general conditions, $\overline{Y}$ will be near μ_Y with very high probability when n is large. Intuitively, when a large number of random variables with the same mean are averaged together, the large values balance the small values and their sample average is close to their common mean. Formally, we say that the sample average $\overline{Y}$ converges in probability to μ_Y (or, equivalently, $\overline{Y}$ is *consistent* for μ_Y) if the probability that $\overline{Y}$ is in the range $\mu_Y - c$ to $\mu_Y + c$ becomes arbitrarily close to 1 as n increases for any constant $c > 0$. The convergence of $\overline{Y}$ to μ_Y in probability is written as $\overline{Y} \xrightarrow{p} \mu_Y$. The law of large numbers says that if Y_i for $i = 1, ..., n$ are independently and identically distributed with $\mathbb{E}[Y_i] = \mu_Y$, and if large outliers are unlikely (technically, if $\text{Var}[Y_i] = \sigma_Y^2 < \infty$), then $\overline{Y} \xrightarrow{p} \mu_Y$.

We can illustrate this result by the following example. Let us consider the example of repeatedly tossing a coin, where Y_i is the result of the ith coin toss. Y_i is a Bernoulli distributed random variable with p being the probability of observing head

$$\Pr(Y_i) = \begin{cases} p, \ Y_i = 1 \\ 1 - p, \ Y_i = 0, \end{cases} \tag{5}$$

where $p = 0.5$ if we assume a fair coin. We know that $\mu_Y = p = 0.5$. Let R_N denote the proportion of heads in the first n tosses,

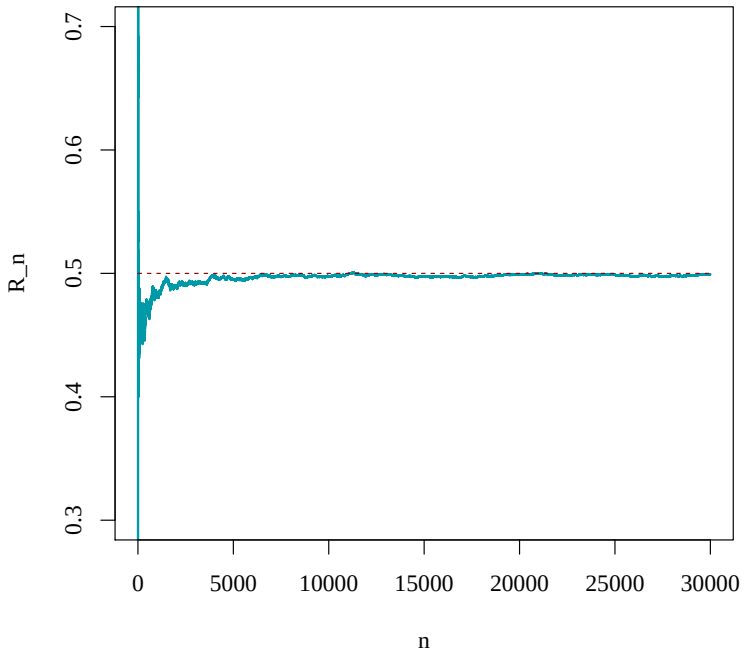
$$R_n = \frac{1}{n} \sum_{i=1}^{n} Y_i. \tag{6}$$

According to the law of large numbers, the observed proportion of heads converges in probability to $\mu_Y = 0.5$, which is the probability of tossing head in a single coin toss:

$$R_n \xrightarrow{p} \mu_Y = 0.5 \text{ as } n \to \infty. \tag{7}$$

In order to check whether this result holds, we can sample n observations from the Bernoulli distribution and calculate the proportion of heads R_n. The outcome of this experiment for $n = 30,000$ is represented in Figure 3. As we can see, if the number of coin tosses is small, the proportion of heads may be anything but close to its theoretical value of $\mu_Y = 0.5$. However, as more and more observations are included in the sample, we find that the path stabilizes in the neighborhood of 0.5. The average of multiple trials shows a clear tendency to converge to its expected value as the sample size increases, just as claimed by the law of large numbers.

FIGURE 3: CONVERGING SHARE OF HEADS IN REPEATED COIN TOSSING



B. The central limit theorem

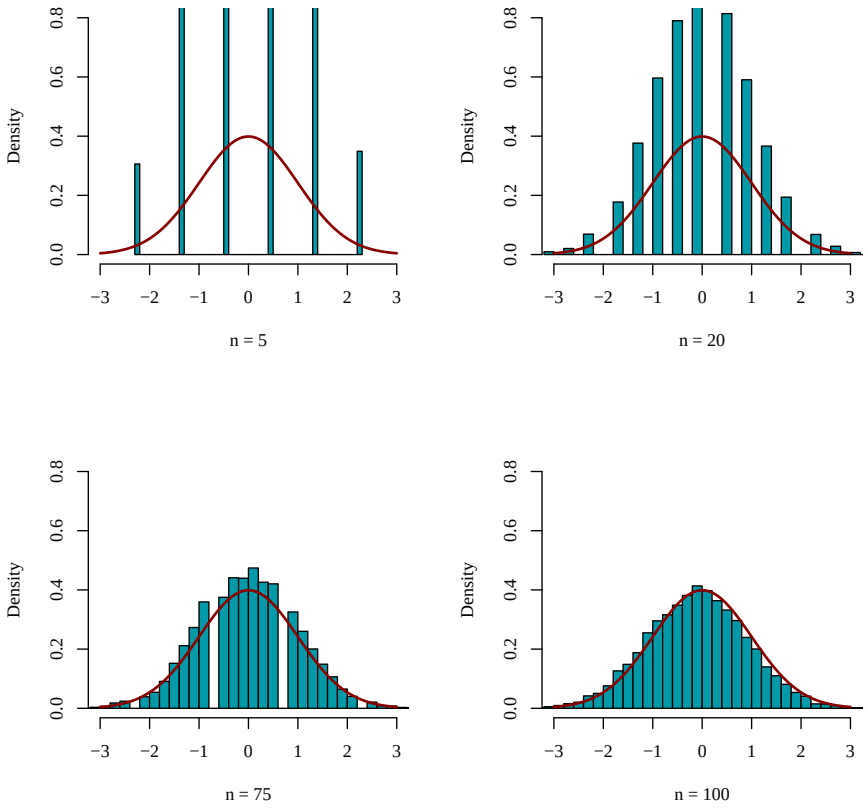
The ***central limit theorem*** says that, under general conditions, the distribution of $\overline{Y}$ is well approximated by a normal distribution when the sample size is large. Formally, suppose that $Y_1, ..., Y_n$ are i.i.d. with $\mathbb{E}[Y_i] = \mu_Y$ and $\text{Var}[Y_i] = \sigma_Y^2$, such that $0 < \sigma_Y^2 < \infty$. As $n \to \infty$, the distribution of $(\overline{Y} - \mu_Y)/\sigma_{\overline{Y}}$, with $\sigma_{\overline{Y}} = \sigma_Y^2/n$ becomes arbitrarily well approximated by the standard normal distribution.

We said before that the mean of $\overline{Y}$ is μ_Y and its variance is $\sigma_{\overline{Y}}^2 = \sigma_Y^2/n$. By the central limit theorem, when n is large, the distribution of $\overline{Y}$ is approximately $\mathcal{N}(\mu_Y, \sigma_{\overline{Y}}^2)$. Note that the distribution of $\overline{Y}$ is exactly $\mathcal{N}(\mu_Y, \sigma_{\overline{Y}}^2)$ when the sample is drawn from a population with normal distribution $\mathcal{N}(\mu_Y, \sigma_Y^2)$. The central limit theorem says that this same result is approximately true when n is large even if $Y_1, ..., Y_n$ are not themselves normally distributed.

At this point, one might wonder how large must n be for the distribution of $\overline{Y}$ to be approximately normal. The answer is that it depends on the distribution of the underlying Y_i that make up the average. At one extreme, if $Y_1, ..., Y_n$ are themselves normally distributed, then $\overline{Y}$ is exactly normally distributed for all n. In contrast, when the underlying $Y_1, ..., Y_n$ themselves have a distribution that is far from normal, the the approximation can require $n = 30$ or even more to be good enough.

This point is illustrated in Figure 4 for the Bernoulli distribution. By the central limit theorem, the distribution of the sample mean $\overline{Y}$ of the Bernoulli distributed random variables Y_i for $i = 1, ..., n$ is well approximated by the normal distribution with parameters $\mu_Y = p = 0.5$ and $\sigma_Y^2 = p(1-p) = 0.25$ for large n. Consequently, for the standardized sample mean we conclude that $(\overline{Y} - 0.5)/(0.5/\sqrt{n})$ should be well approximated by the standard normal distribution $\mathcal{N}(0, 1)$. In order to check whether this is true, we can draw a large number of random samples, say 10,000, of size n from the Bernoulli distribution, compute the sample averages and standardize them, and repeat the experiment for sample sizes of 5, 20, 75 and 100. We will see that the simulated sampling distribution of the standardized average tends to deviate strongly from the standard normal distribution if the sample size is small (e.g., for $n = 5$ and $n = 20$). However, as n grows, the histograms approach the standard normal distribution (the approximation works quite well for $n = 100$).

FIGURE 4: CONVERGING STANDARDIZED SAMPLE MEAN FOR THE BERNOULLI DISTRIBUTION



In sum, the central limit theorem is a remarkable result. While the "small n" distribution of $\overline{Y}$ can be difficult to obtain and depends on the underlying distribution of Y_i, the "large n" distribution is quite simple. Because the distribution of $\overline{Y}$

approaches the normal as n grows large, $\overline{Y}$ is said to have an ***asymptotic normal distribution***. The convenience of the normal approximation, combined with its wide applicability because of the central limit theorem, makes it a key underpinning of modern applied econometrics.

III. Estimation

A. Estimators and their properties

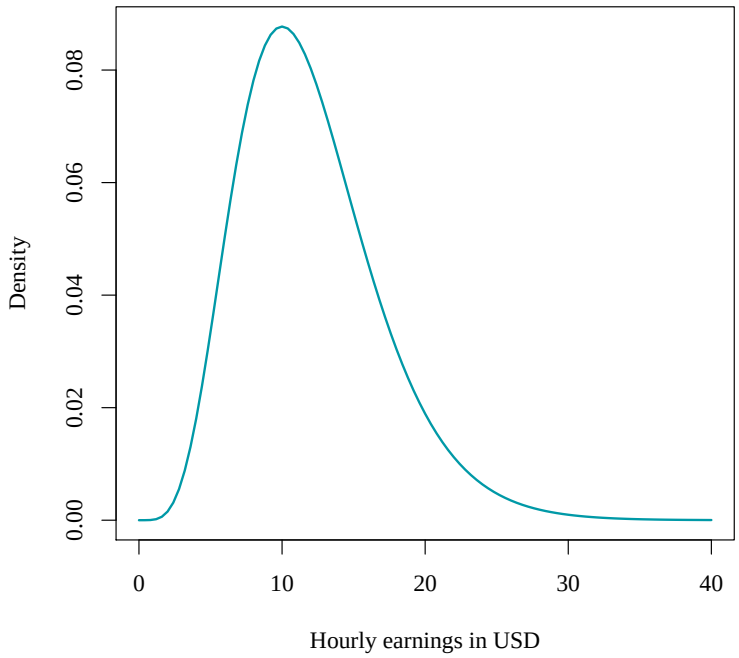
To get started, let us think of some economic variable in a population, for example hourly earnings of college graduates, denoted by Y, and let us assume that we are interested in its mean value, that is, μ_Y. Instead of conducting an interview with every working graduate and asking about hourly earnings (which is unfeasible), we can opt for a natural way to estimate this mean, which is to compute the sample average $\overline{Y}$ from a sample of n independently and identically distributed observations $Y_1, Y_2, ..., Y_n$ (they will be i.i.d. if they are collected by simple random sampling). In this context, we say that the sample mean is an ***estimator*** of the population mean, i.e., a function of a sample of data to be drawn randomly from a population which provides with an ***estimate***, a numerical value of the population characteristic of interest (in our case, the population mean μ_Y). An estimator is a random variable because it depends on the sample used (takes different values from one sample to the next), which is selected at random, while an estimate is a non-random number.

However, using the sample mean is not the only way of estimating μ_Y. Why not using an even simpler estimator, such as the very first observation Y_1 in the sample? Is Y_1 a good estimator? For now, let's assume that Y follows a chi-squared distribution with 12 degrees of freedom χ^2_{12}, which might be plausible given that hourly income is non-negative and we expect many hourly earnings to be in a range of \$5 to \$15. Moreover, it is common for income distributions to be skewed to the right, which is a property of the chosen distribution.

If we now draw a sample of $n = 100$ observations, take the first observation Y_1 as an estimate for μ_Y, and Y_1 turns out to be, say, 13.91, this value is near from the population mean, which is $\mu_Y = 12$. Nevertheless, it is intuitive that we could do better. After all, Y_1, as an estimator of μ_Y, discards a lot of information, and its variance is the population variance, which is big: $\text{Var}[Y_1] = \text{Var}[Y] = 2 \times 12 = 24$.

But then, if there are many possible estimators, what makes one estimator "better" than another? What is a "good" estimator of an unknown parameter? Because estimators are random variables, these questions can be phrased more precisely: what are desirable characteristics of the sampling distribution of an estimator? In

FIGURE 5: HYPOTHETICAL DISTRIBUTION OF HOURLY EARNINGS ($Y \sim \chi^2_{12}$)



general, we would like an estimator that gets as close as possible to the unknown true value, at least in some average sense; in other words, we would like the sampling distribution of an estimator to be as tightly centered on the unknown value as possible. This observation leads to three specific desirable characteristics of an estimator: unbiasedness (lack of bias), consistency, and efficiency.

First, let's suppose that we evaluate an estimator many times over repeated randomly drawn samples. It is reasonable to hope that, on average, we would get the right answer. Thus, a desirable property of an estimator is that the mean of its sampling distribution equals μ_Y. If that is the case, we say that the estimator is ***unbiased***. Formally, let $\hat{\mu}_Y$ denote some estimator of μ_Y, such as $\overline{Y}$ or Y_1. The estimator $\hat{\mu}_Y$ is unbiased if

$$\mathbb{E}[\hat{\mu}_Y] = \mu_Y, \tag{8}$$

where $\mathbb{E}[\hat{\mu}_Y]$ is the mean of the sampling distribution of $\hat{\mu}_Y$. If this equality does not hold, $\hat{\mu}_Y$ is biased.

A second desirable property of an estimator $\hat{\mu}_Y$ is that it is ***consistent***, what entails that, when the sample size is large, the uncertainty about the value of μ_Y arising from random variations in the sample is very small. More precisely, $\hat{\mu}_Y$ is a consistent estimator of μ_Y if $\hat{\mu}_Y \xrightarrow{p} \mu_Y$, meaning that the probability that $\hat{\mu}_Y$ is within a small interval surrounding the true value of μ_Y approaches 1 as the sample

size increases.

Having said that, let us assume that we have two candidate estimators for μ_Y, which are $\hat{\mu}_Y$ and $\tilde{\mu}_Y$, both of which are unbiased and consistent. How can we choose between them? One way to do so is to choose the estimator with the tightest sampling distribution. This suggests choosing between $\hat{\mu}_Y$ and $\tilde{\mu}_Y$ by picking the estimator with the smallest variance. If $\hat{\mu}_Y$ has a smaller variance than $\tilde{\mu}_Y$, then we say that $\hat{\mu}_Y$ is more efficient than $\tilde{\mu}_Y$. The terminology ***efficiency*** stems from the notion that if $\hat{\mu}_Y$ has a smaller variance than $\tilde{\mu}_Y$, then it uses the information in the data more efficiently than does $\tilde{\mu}_Y$.

B. Properties of the sample mean

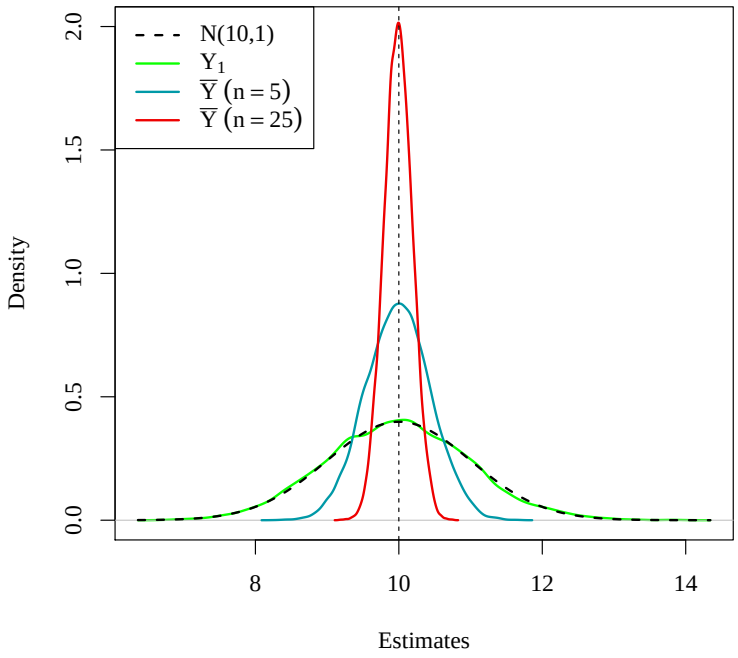
After discussing the desirable properties of an estimator, how does the sample mean $\overline{Y}$ fare as an estimator of μ_Y when judged by the three criteria of bias, consistency and efficiency? To begin with, from the previous chapter we know that $\mathbb{E}\left[\overline{Y}\right] = \mu_Y$, so $\overline{Y}$ is an unbiased estimator of μ_Y. Besides, the law of large numbers states that $\overline{Y} \xrightarrow{p} \mu_Y$, that is, $\overline{Y}$ is consistent.

What can be said about efficiency? In this case, we need to compare $\overline{Y}$ with other estimators. Let's start by comparing the efficiency of $\overline{Y}$ to the one of Y_1. Because $Y_1, ..., Y_n$ are i.i.d., the mean of the sampling distribution of Y_1 is $\mathbb{E}[Y_1] = \mu_Y$. Thus Y_1 is an unbiased estimator of μ_Y. Its variance is equal to the variance of the population, σ_Y^2. Meanwhile, the variance of $\overline{Y}$ is σ_Y^2/n. Thus, for $n \geq 2$, the variance of $\overline{Y}$ is lower than the variance of Y_1, so $\overline{Y}$ is a more efficient estimator than Y_1. Of course, Y_1 might strike as an obviously poor suggestion for an estimator –why would we go to the trouble of collecting a sample of n observations only to throw away all but the first?–, and the concept of efficiency provides a formal way to show that $\overline{Y}$ is a better estimator than Y_1.

In Figure 6 we compare these two estimators by generating a population of 10,000 observations from a $\mathcal{N}(10, 1)$ distribution and drawing 25,000 samples of sizes $n = 5$ and $n = 25$. As we would expect, the sampling distribution of Y_1 tracks the density of the $\mathcal{N}(10, 1)$ distribution pretty closely, since Y_1 is nothing but an observation that is randomly selected from a population with the $\mathcal{N}(10, 1)$ distribution. However, the sampling distribution of $\overline{Y}$ shows less dispersion than the sampling distribution of Y_1, because $\overline{Y}$ has a lower variance than Y_1, and the density of the former is more concentrated around the population mean than the latter, an effect that becomes stronger as the the sample size increases.

What about a less obviously poor estimator? Let's consider the weighted average

FIGURE 6: SAMPLING DISTRIBUTIONS OF UNBIASED ESTIMATORS



in which the observations are alternatively weighted by 1/2 and 3/2:

$$\tilde{Y} = \frac{1}{n}\left(\frac{1}{2}Y_1 + \frac{3}{2}Y_2 + \frac{1}{2}Y_3 + ... + \frac{1}{2}Y_{n-1} + \frac{3}{2}Y_n\right) \tag{9}$$

where the number of observations n is assumed to be even for convenience. The mean of $\tilde{Y}$ is μ_Y, and its variance is $\text{Var}\left[\tilde{Y}\right] = 1.25\sigma_Y^2/n$. Then, $\tilde{Y}$ is unbiased, and because $\text{Var}[\tilde{Y}] \to 0$ as $n \to \infty$, $\tilde{Y}$ is consistent. However, $\tilde{Y}$ has a larger variance than $\overline{Y}$, so $\overline{Y}$ is more efficient than $\tilde{Y}$.

Note that estimators $\overline{Y}$, Y_1 and $\tilde{Y}$ are all weighted averages of $Y_1, ..., Y_n$. We have concluded that Y_1 and $\tilde{Y}$ have larger variances than $\overline{Y}$, what reflects a more general result: $\overline{Y}$ is the most efficient estimator of all unbiased estimator that are weighted averages of $Y_1, ..., Y_n$. Said differently, $\overline{Y}$ is the ***Best Linear Unbiased Estimator (BLUE)***, i.e., the most efficient (best) estimator among all estimators that are unbiased and linear functions of $Y_1, ..., Y_n$.

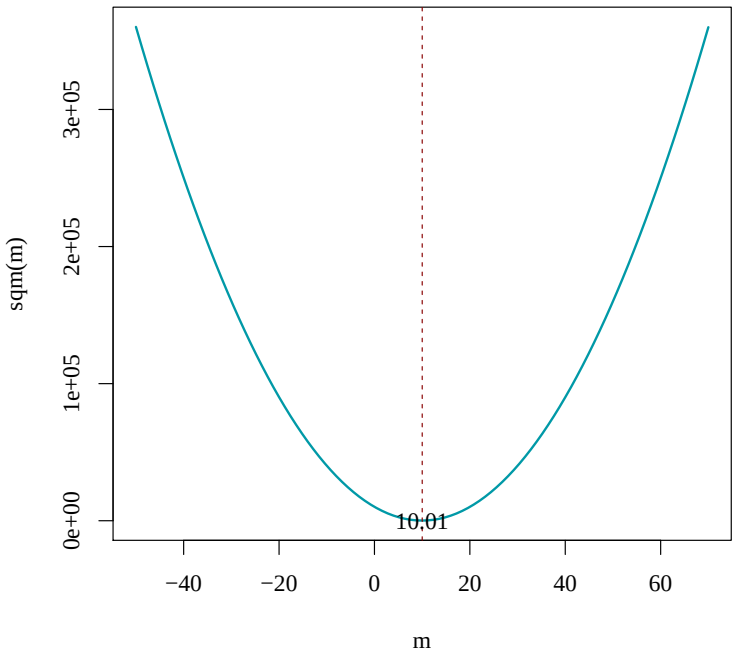
Another advantage of the sample mean $\overline{Y}$ as an estimator of the population mean μ_y is that it provides the *best* fit to the data, in the sense that it is the solution to the problem of finding the estimator m that minimizes the sum of squared deviations

between observed and predicted values:

$$\sum_{i=1}^{n} (Y_i - m)^2 . \tag{10}$$

We can think of the difference between Y_i and m as the mistake made when predicting Y_i using m, so that we can interpret the above expression as the sum of squared prediction mistakes. It is possible to show that the sample mean $\overline{Y}$ is the estimator m that solves this problem. In other words, $\overline{Y}$ is the ***least squares estimator***. Figure 7 illustrates this point.

FIGURE 7: THE SAMPLE MEAN AS THE LEAST SQUARES ESTIMATOR



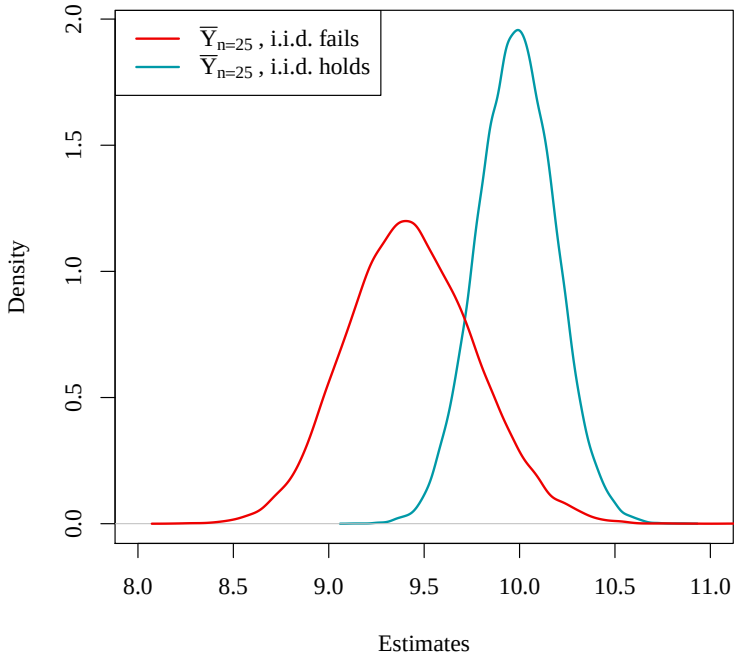
C. The importance of random sampling

So far, we have assumed that the observed data $Y_1, ..., Y_n$ are i.i.d. draws, such that these are the result of a sampling process that satisfies the assumption of simple random sampling. This assumption is important, because non-random sampling can result in the sample mean $\overline{Y}$ being biased. For instance, carrying out a poll only among individuals of certain socioeconomic characteristics can be very misleading, since the voters of some political parties might be overly represented in the sample.

In order to illustrate this idea, let's suppose that we have a population of 10,000 individuals and are interested in computing the mean μ_Y. We could draw random

samples of size 10 from the population and estimate μ_Y using $\overline{Y}$ for each of them. The results can be compared to those obtained under a different sampling scheme that deviates from simple random sampling. In such a case, instead of ensuring that each population member has the same chance to end up in a sample, we could assign a higher probability of being sampled to the 2,500 smallest observations of the population. Figure 8 plots the result of this exercise. As we can see, when the i.i.d. assumption fails, μ_Y is underestimated on average, that is, $\overline{Y}$ becomes a biased estimator for μ_Y.

FIGURE 8: WHEN THE *i.i.d.* ASSUMPTION FAILS ($Y \sim \mathcal{N}(10, 1)$)



IV. Hypothesis Testing

Many hypotheses about the world around us can be phrased as yes/no questions. Do the mean hourly earnings of recent U.S. college graduates equal \$20 per hour? Are mean earnings the same for male and female college graduates? Both questions embody specific hypotheses about the population distribution of earnings, and we are interested in finding an answer based on a sample of evidence.

A. Null and alternative hypotheses

The starting point of statistical hypotheses testing is specifying the hypothesis to be tested, called the ***null hypothesis***. Hypothesis testing entails using data to com-

pare the null hypothesis to a second hypothesis, called the ***alternative hypothesis***, that holds if the null does not.

We are going to focus on hypothesis tests concerning the population mean $\mathbb{E}[Y]$. In this case, the null hypothesis, denoted H_0, is that $\mathbb{E}[Y]$ takes on a specific value $\mu_{Y,0}$:

$$H_0 : \mathbb{E}[Y] = \mu_{Y,0}. \tag{11}$$

For example, the conjecture that, on average in the population, college graduates earn \$20 per hour constitutes a null hypothesis about the population distribution of hourly earnings. Stated formally, if Y represents the hourly earnings of a randomly selected college graduate, then the null hypothesis is that $\mathbb{E}[Y] = 20$, that is, $\mu_{Y,0} = 20$ in (11).

The alternative hypothesis specifies what is true if the null hypothesis is not. The most general alternative hypothesis is that $\mathbb{E}[Y] \neq \mu_{Y,0}$, which is called a ***two-sided alternative hypothesis*** because it allows $\mathbb{E}[Y]$ to be either less or greater than $\mu_{Y,0}$. The two-sided alternative is written as

$$H_1 : \mathbb{E}[Y] \neq \mu_{Y,0}. \tag{12}$$

One-sided alternatives are also possible, and we will discuss them later. In any case, the problem we face as statisticians is to use the evidence in a randomly selected sample of data to decide whether to accept the null hypothesis H_0 or to reject it in favor of the alternative hypothesis H_1. If the null hypothesis is "accepted", this does not mean that we declare it to be true; rather, it is accepted tentatively with the recognition that it might be rejected later based on additional evidence. For this reason, the result of hypothesis testing can be posed as either rejecting the null hypothesis or failing to do so.

B. The p-value

In any given sample, the sample average $\overline{Y}$ will rarely be exactly equal to the hypothesized value $\mu_{Y,0}$. Differences between $\overline{Y}$ and $\mu_{Y,0}$ can arise because the true mean in fact does not equal $\mu_{Y,0}$ (the null hypothesis is false) or because the true mean equals $\mu_{Y,0}$ (the null hypothesis is true) but $\overline{Y}$ differs from $\mu_{Y,0}$ because of random sampling. It is impossible to distinguish between these two possibilities with certainty. Although a sample of data cannot provide conclusive evidence about the null hypothesis, it is possible to do a probabilistic calculation that permits testing the null hypothesis in a way that accounts for sampling uncertainty. This calculation involves using the data to compute the p-value of the null hypothesis.

The ***p-value***, also called the ***significance probability***, is the probability of drawing a statistic at least as adverse to the null hypothesis as the one we can actually compute in our sample, assuming the null hypothesis is correct. In our case, the p-value is the probability of drawing $\overline{Y}$ at least as far in the tails of its distribution under the null hypothesis as the sample average we can compute.

For example, let's assume that in our sample of college graduates the average wage is \$22.64. The p-value is the probability of observing a value of $\overline{Y}$ at least as different from \$20 (the population mean under the null) as the observed value of \$22.64 by pure random sampling variation, assuming that the null hypothesis is true. If this p-value is small, say 0.5%, then it is very unlikely that this sample would have been drawn if the null hypothesis is true. Thus, it would be reasonable to conclude that the null hypothesis is not true. By contrast, if this p-value is large, say 40%, then it is quite likely that the observed sample average of \$22.64 could have arisen just by random sampling variation if the null hypothesis is true. Accordingly, the evidence against the null hypothesis would be weak in a probabilistic sense, and it would be reasonable not to reject the null hypothesis.

Formally, let $\overline{Y}^{act}$ denote the value of the sample average actually computed in the dataset at hand, and let $\Pr_{H_0}$ denote the probability computed under the null hypothesis (that is, computed assuming that $\mathbb{E}[Y_i] = \mu_{Y,0}$). The p-value is

$$p\text{-value} = \Pr_{H_0}\left[|\overline{Y} - \mu_{Y,0}| > |\overline{Y}^{act} - \mu_{Y,0}|\right]. \tag{13}$$

That is, the p-value is the area in the tails of the distribution of $\overline{Y}$ under the null hypothesis that is beyond $|\overline{Y}^{act} - \mu_{Y,0}|$. If the p-value is large, then the observed value $\overline{Y}^{act}$ is consistent with the null hypothesis, but if the p-value is small, it is not.

To compute the p-value, it is necessary to know the sampling distribution of $\overline{Y}$ under the null hypothesis. Fortunately, as stated by the central limit theorem, when the sample size is large, the sampling distribution of $\overline{Y}$ is well approximated by a normal distribution. Under the null hypothesis the mean of this normal distribution is $\mu_{Y,0}$, so under the null hypothesis $\overline{Y}$ is distributed $\mathcal{N}\left(\mu_{Y,0}, \sigma^2_{\overline{Y}}\right)$, where $\sigma^2_{\overline{Y}} = \sigma^2_Y/n$. This large-sample normal approximation makes it possible to compute the p-value without the need to know the population distribution of Y, as long as the sample size is large. The details of the calculation, however, depend on whether the variance σ^2_Y is known.

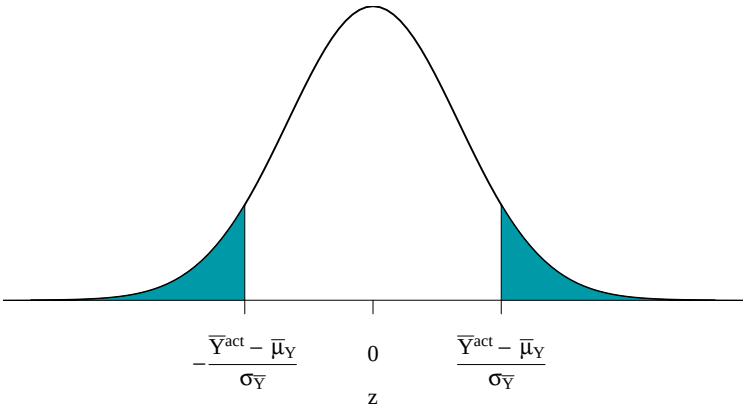
C. Calculating the p-value when the variance is known

The calculation of the p-value when σ_Y is known is summarized in Figure 9. If the sample size is large, under the null hypothesis the sampling distribution of $\overline{Y}$ is $\mathcal{N}(\mu_{Y,0}, \sigma^2_{\overline{Y}})$, where $\sigma^2_{\overline{Y}} = \sigma^2_Y/n$. Thus, under the null hypothesis, the standardized version of $\overline{Y}$, $(\overline{Y} - \mu_{Y,0})/\sigma_{\overline{Y}}$, has a standard normal distribution. The p-value is the probability of obtaining a value of $\overline{Y}$ farther from $\mu_{Y,0}$ than $\overline{Y}^{act}$ under the null hypothesis, or equivalently, is the probability of obtaining $(\overline{Y} - \mu_{Y,0})/\sigma_{\overline{Y}}$ greater than $(\overline{Y}^{act} - \mu_{Y,0})/\sigma_{\overline{Y}}$ in absolute value. This probability is the shaded area shown in Figure 9. Formally, this is

$$p\text{-value} = \Pr_{H_0}\left(\left|\frac{\overline{Y} - \mu_{Y,0}}{\sigma_{\overline{Y}}}\right| > \left|\frac{\overline{Y}^{act} - \mu_{Y,0}}{\sigma_{\overline{Y}}}\right|\right) = 2\Phi\left(-\left|\frac{\overline{Y}^{act} - \mu_{Y,0}}{\sigma_{\overline{Y}}}\right|\right), \tag{14}$$

where Φ is the standard normal cumulative distribution function. That is, the p-value is the area in the tails of a standard normal distribution outside $\pm(\overline{Y}^{act} - \mu_{Y,0})/\sigma_{\overline{Y}}$. The above equation depends on the variance of the population distribution, σ^2_Y. In practice, this variance is often unknown.

FIGURE 9: CALCULATION OF THE P-VALUE



D. Sample variance, sample standard deviation, and standard error

Because in general σ^2_Y must be estimated before the p-value can be computed, we now turn to the problem of estimating σ^2_Y. For this purpose, we use the ***sample variance*** s^2_Y

$$s^2_Y = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2. \tag{15}$$

The ***sample standard deviation*** s_Y is the square root of the sample variance. Note that the divisor used here is $n-1$ instead of n. The reason for this modification is that estimating μ_Y by $\overline{Y}$ introduces a small downward bias in $\left(Y_i - \overline{Y}\right)^2$. Specifically,

$$\mathbb{E}\left[\left(Y_i - \overline{Y}\right)^2\right] = \left[(n-1)/n\right]\sigma_Y^2. \tag{16}$$

Thus,

$$\mathbb{E}\left[\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2\right] = n\mathbb{E}\left[\left(Y_i - \overline{Y}\right)^2\right] = (n-1)\sigma_Y^2. \tag{17}$$

Dividing by $n-1$ instead of n corrects for this downward bias, and as a result s_Y^2 is unbiased.[2]

The sample variance is also a consistent estimator of the population variance: $s_Y^2 \xrightarrow{p} \sigma_Y^2$. In other words, the sample variance is close to the population variance with high probability when the sample size is large. We prove this result formally below, under the assumptions that $Y_1, ..., Y_n$ are i.i.d. and Y_i has a finite fourth moment, i.e., $\mathbb{E}[Y_i^4] < \infty$.

Proof. First, add and subtract μ_Y to write

$$\begin{aligned}\left(Y_i - \overline{Y}\right)^2 &= \left[(Y_i - \mu_Y) - \left(\overline{Y} - \mu_Y\right)\right]^2 \\ &= (Y_i - \mu_Y)^2 - 2\left(Y_i - \mu_Y\right)\left(\overline{Y} - \mu_Y\right) + \left(\overline{Y} - \mu_Y\right)^2. \end{aligned} \tag{18}$$

Substituting this expression for $(Y_i - \overline{Y})^2$ into (15), we have

$$\begin{aligned} s_Y^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \mu_Y)^2 - \frac{2}{n-1}\sum_{i=1}^{n}(Y_i - \mu_Y)\left(\overline{Y} - \mu_Y\right) + \frac{1}{n-1}\sum_{i=1}^{n}\left(\overline{Y} - \mu_Y\right)^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \mu_Y)^2 - \frac{2}{n-1}\left(\overline{Y} - \mu_Y\right)\sum_{i=1}^{n}(Y_i - \mu_Y) + \frac{n}{n-1}\left(\overline{Y} - \mu_Y\right)^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \mu_Y)^2 - \frac{2}{n-1}\left(\overline{Y} - \mu_Y\right)\left(n\overline{Y} - n\mu_Y\right) + \frac{n}{n-1}\left(\overline{Y} - \mu_Y\right)^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \mu_Y)^2 - \frac{2n}{n-1}\left(\overline{Y} - \mu_Y\right)^2 + \frac{n}{n-1}\left(\overline{Y} - \mu_Y\right)^2 \\ &= \left(\frac{n}{n-1}\right)\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu_Y)^2\right] - \left(\frac{n}{n-1}\right)\left(\overline{Y} - \mu_Y\right)^2. \end{aligned} \tag{19}$$

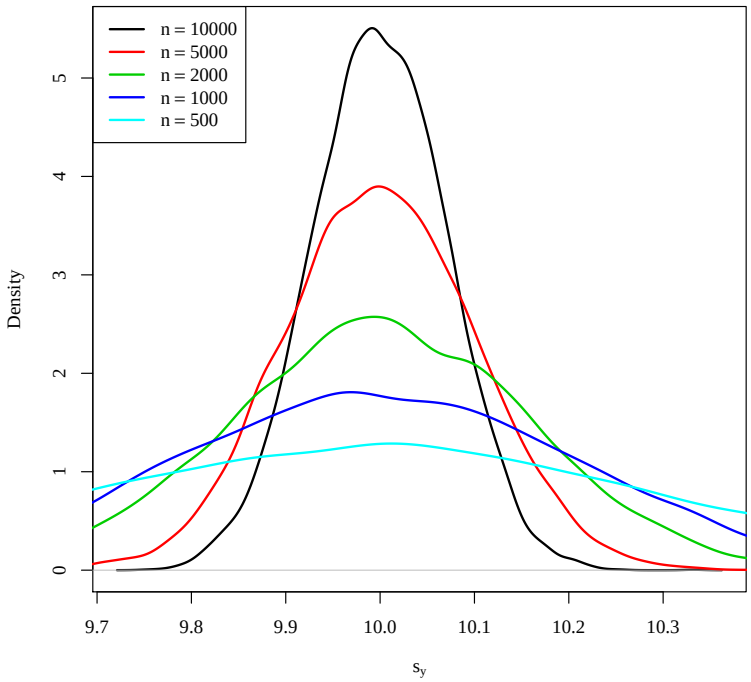
The law of large numbers can now be applied. Define $W_i = (Y_i - \mu_Y)^2$. Note that

[2] This is a *degrees of freedom correction*. Estimating the mean uses up one degree of freedom in the data, so that only $n-1$ degrees of freedom remain.

$\mathbb{E}[W_i] = \sigma_Y^2$. Because the random variables $Y_i, ..., Y_n$ are i.i.d., the random variables $W_1, ..., W_n$ are also i.i.d. In addition, $\mathbb{E}[W_i^2] = \mathbb{E}\left[(Y_i - \mu_Y)^4\right] < \infty$, because we have assumed that $\mathbb{E}[Y_i^4] < \infty$. Thus, $W_1, ..., W_n$ are i.i.d., and $\text{Var}[W_i] < \infty$, so $\overline{W}$ satisfies the conditions for the law of large numbers to be applied. Therefore, $\overline{W} \xrightarrow{p} \mathbb{E}[W_i]$. Besides, $\overline{W} = (1/n)\sum_{i=1}^n (Y_i - \mu_Y)^2$ and $\mathbb{E}[W_i] = \sigma_Y^2$, so $(1/n)\sum_{i=1}^n (Y_i - \mu_Y)^2 \xrightarrow{p} \sigma_Y^2$. Also, $n/(n-1) \to 1$, so the first term in the above equation converges in probability to σ_Y^2. Finally, note that $\overline{Y} \xrightarrow{p} \mu_Y$ and $(\overline{Y} - \mu_Y)^2 \xrightarrow{p} 0$, so the second term converges in probability to zero. Combining these results yields $s_Y^2 \xrightarrow{p} \sigma_Y^2$. □

Intuitively, the reason why s_Y^2 is consistent is that it is a sample average, so it obeys the law of large numbers as long as $(Y_i - \mu_Y)^2$ has a finite variance, which in turn means that $\mathbb{E}[Y_i^4]$ must be finite. Figure 10 illustrates the consistency of s_Y^2 by plotting the results of generating a large number of samples from a $\mathcal{N}(10, 10)$ distribution and computing s_Y for different sample sizes. As we can see, the distribution of s_Y tightens around the true value $\sigma_Y = 10$ as the sample size increases.

FIGURE 10: SAMPLING DISTRIBUTIONS OF s_Y



Because the standard deviation of the sampling distribution of $\overline{Y}$ is $\sigma_{\overline{Y}} = \sigma_Y/\sqrt{n}$, we can use $s_Y/\sqrt{n}$ as an estimator of $\sigma_{\overline{Y}}$. This estimator is called the ***standard error*** of $\overline{Y}$ and is denoted $SE(\overline{Y})$ or $\hat{\sigma}_{\overline{Y}}$.

As an example of standard error, let's consider a sample of 100 i.i.d. observations of the Bernoulli distributed variable Y with success probability $p = 0.1$. Then

$\mathbb{E}[Y] = p = 0.1$ and $\text{Var}[Y] = p(1-p) = 0.09$. $\mathbb{E}[Y]$ can be estimated by $\overline{Y}$, which has variance

$$\sigma^2_{\overline{Y}} = \frac{p(1-p)}{n} = 0.0009 \tag{20}$$

and standard deviation

$$\sigma_{\overline{Y}} = \sqrt{\frac{p(1-p)}{n}} = 0.03. \tag{21}$$

In this case, the standard error of $\overline{Y}$ can be estimated by

$$SE\left(\overline{Y}\right) = \sqrt{\frac{\overline{Y}\left(1-\overline{Y}\right)}{n}}. \tag{22}$$

E. Calculating the p-value when the variance is unknown

When σ_Y is unknown and $Y_1, ..., Y_n$ are i.i.d., the p-value can be computed by replacing $\sigma_{\overline{Y}}$ in Equation (14) by the standard error $SE(\overline{Y}) = \hat{\sigma}_{\overline{Y}}$. That is, when σ_Y is unknown and $Y_1, ..., Y_n$ are i.i.d., the p-value is calculated using

$$p\text{-value} = 2\Phi\left(-\left|\frac{\overline{Y}^{act} - \mu_{Y,0}}{SE(\overline{Y})}\right|\right). \tag{23}$$

F. The t-statistic

The standardized sample mean $(\overline{Y} - \mu_{Y,0})/SE(\overline{Y})$ plays a central role in testing statistical hypothesis and has a special name, the ***t-statistic*** or ***t-ratio***:

$$t = \frac{\overline{Y} - \mu_{Y,0}}{SE(\overline{Y})}. \tag{24}$$

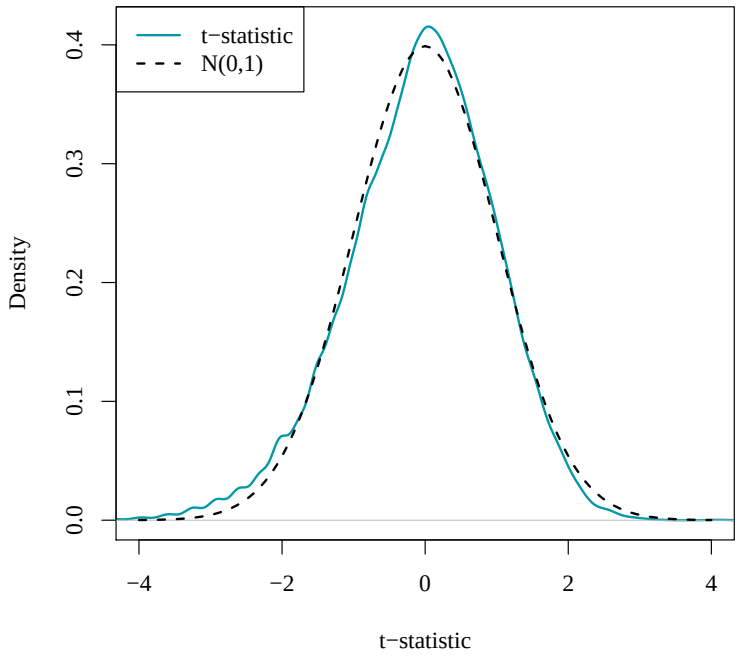
In general, a ***test statistic*** is a statistic used to perform a hypothesis test. The t-statistic is an important example of a test statistic.

Note that when the sample size is large, s^2_Y is close to σ^2_Y with high probability. Thus the distribution of the t-statistic is approximately the same as the distribution of $(\overline{Y} - \mu_{Y,0})/\sigma_{\overline{Y}}$, which in turn is well approximated by the standard normal distribution because of the central limit theorem. Accordingly, under the null hypothesis, t is approximately distributed $\mathcal{N}(0,1)$ for large n. Figure 11 illustrates this idea.

Equation (23) can be rewritten in terms of the t-statistic. Let t^{act} denote the value of the t-statistic actually computed:

$$t^{act} = \frac{\overline{Y}^{act} - \mu_{Y,0}}{SE(\overline{Y})}. \tag{25}$$

FIGURE 11: ESTIMATED DISTRIBUTION OF THE t-STATISTIC WHEN $n = 300$



Accordingly, when n is large, the p-value can be calculated using

$$p\text{-value} = 2\Phi\left(- \mid t^{act} \mid\right). \tag{26}$$

As a hypothetical example, suppose that a sample of 200 recent college graduates is used to test the null hypothesis that the mean wage $\mathbb{E}[Y]$ is \$20 per hour. Suppose that the sample average wage is $\overline{Y}^{act} = \$22.64$ and the sample standard deviation is $s_Y = \$18.14$. Then the standard error of $\overline{Y}$ is $s_Y/\sqrt{n} = 18.14/\sqrt{200} = 1.28$. The value of the t-statistic is $t^{act} = (22.64 - 20)/1.28 = 2.06$. Looking at the tables of the cumulative standardized normal distribution, the p-value is $2 \times \Phi(-2.06) = 2 \times 0.0197 = 0.039$, or 3.9%. That is, under the assumption that the null hypothesis is true, the probability of obtaining a sample average at least as different from the null as the one actually computed is 3.9%.

G. Hypothesis testing with a prespecified significance level

When we undertake a hypothesis test, we can make two types of mistakes: a ***type-I error***, in which the null hypothesis is rejected when in fact it is true; and a ***type-II error***, in which the null hypothesis is not rejected when in fact it is false. Hypothesis tests can be performed without computing the p-value if we are willing to specify in advance the probability of making the type-I error that we are willing

to tolerate, i.e., of incorrectly rejecting the null hypothesis when it is true. If we choose a prespecified probability of rejecting the null hypothesis when it is true of, for example, 5%, then we will reject the null hypothesis if and only if the p-value is less than 0.05. This prespecified probability of committing a type-I error is called the ***significance level*** of the test. Meanwhile, the actual probability that the test rejects the true null hypothesis is called the ***size*** of the test. In an ideal setting, the size equals the significance level. Apart from them, sometimes we are also interested in the ***power*** of the test, which is the probability that the test correctly rejects a false null hypothesis.

If we are conducting a hypothesis test on the population mean with $H_0 : \mu_Y = \mu_{Y,0}$ and $H_1 : \mu_Y \neq \mu_{Y,0}$, and have decided that the null hypothesis will be rejected if the p-value is less than 5%, we can follow a simple rule:

$$\text{Reject } H_0 \text{ if } \mid t^{act} \mid > 1.96, \tag{27}$$

that is, we reject the null hypothesis if the absolute value of the t-statistic computed from the sample is greater than 1.96. We know that if n is large enough, the t-statistic has a $\mathcal{N}(0,1)$ distribution under the null hypothesis, and we use this number (1.96) because the area under the tails of the normal distribution and outside ± 1.96 is 5%. This value, for which the test just rejects the null hypothesis at the given significance level, is called the ***critical value*** of the test. In other words, the critical value of the test draws a line between the ***rejection region*** and the ***acceptance region*** of the test. The former is the set of values for which the test rejects the null hypothesis (in our case, the values of the t-statistic outside ± 1.96), and the latter is the complementary, i.e., the set of values for which the test does not reject the null hypothesis. If the test rejects the null hypothesis at the 5% significance level, we say that the population mean μ_Y is ***statistically different*** from $\mu_{Y,0}$ at the 5% significance level.

At this point, we might wonder what significance level to use in practice. In many cases, statisticians and econometricians use a 5% significance level, but this implies that if we test many hypothesis, we will incorrectly reject the null on average 5% of the times. Sometimes a more conservative significance level is more convenient, especially in the case that we want to be quite sure that a rejection of the null hypothesis is not just a result of random sample variation. The more conservative we want to be in this sense, the lower the significance level must be. Yet, it is important to notice that the lower is the significance level, the lower is the power of the test, i.e., the more difficult it becomes to reject the null when it is actually false.

H. One-sided alternatives

In some circumstances, the alternative hypothesis might be that the mean exceeds $\mu_{Y,0}$. For example, if we hope that education helps in the labor market, the relevant alternative to the null hypothesis that earnings are the same for college graduates and non-college graduates is not just that their earnings differ, but rather that graduates earn more than non-graduates. This is called a ***one-sided alternative hypothesis*** and can be written

$$H_1 : \mu_Y > \mu_{Y,0}. \tag{28}$$

The general approach to computing p-values and to hypothesis testing is the same for one-sided alternatives as it is for two-sided alternatives, with the modification that only large positive values of the t-statistic reject the null hypothesis rather than values that are large in absolute value. Specifically, to test the one-sided hypothesis above, we construct the t-statistic as before, but the p-value is the area under the standard normal distribution to the right of the calculated t-statistic:

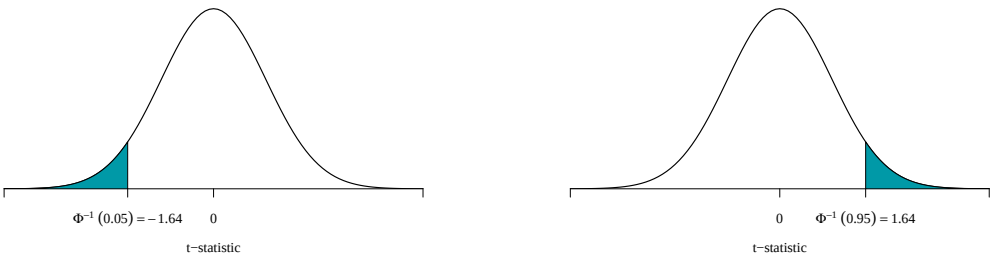
$$p\text{-value} = \Pr_{H_0}\left(Z > t^{act}\right) = 1 - \Phi(t^{act}). \tag{29}$$

The $\mathcal{N}(0,1)$ critical value for a one-sided test with a 5% significance level is 1.64, so the rejection region for this test is composed of all values of the t-statistic exceeding 1.64.

If instead the alternative hypothesis is that $\mu_Y < \mu_{Y,0}$, then the above discussion applies except that the signs are switched: for example, the 5% rejection region consists of values of the t-statistic less than -1.64. We illustrate the two types of one-sided tests in Figure 12.

Figure 12: Rejection Region of a One-Sided Test

(a) Left-sided test ($\mu_Y < \mu_{Y,0}$) (b) Right-sided test ($\mu_Y > \mu_{Y,0}$)



V. Confidence Intervals

Because of random sampling error, we will never be able to estimate the exact value of the population mean of Y using only the information in a sample. However, it is possible to use data from a random sample to construct a set of values that contains the true population mean μ_Y with a certain prespecified probability. Such a set is called a ***confidence set***, and the prespecified probability that μ_Y is contained in this set is called the ***confidence level***. The confidence set for μ_Y turns out to be all the possible values of the mean between a lower and an upper limit, so that the confidence set is an interval called ***confidence interval***.

If we are interested in constructing a 95% confidence set for the population mean, we must first pick some arbitrary value for the mean, let's call it $\mu_{Y,0}$. Next we can test the null hypothesis that $\mu_Y = \mu_{Y,0}$ against the alternative that $\mu_Y \neq \mu_{Y,0}$ by computing the t-statistic. If it is less than 1.96, this hypothesized value $\mu_{Y,0}$ is not rejected at the 5% level. Now we can pick another arbitrary value for $\mu_{Y,0}$ and test it, and so on. Continuing this process yields the set of all values of the population mean that cannot be rejected at the 5% level by a two-sided hypothesis test based on our data. We can notice that this set of values has a remarkable property: the probability that it contains the true value of the population mean is 95%. This probability (95%) that the confidence interval computed over all possible random samples contains the true population mean is called the ***coverage probability***.

For instance, if the true value of μ_Y is 21.5, then $\overline{Y}$ has a normal distribution centered on 21.5, and the t-statistic testing the null hypothesis $\mu_Y = 21.5$ has a $\mathcal{N}(0,1)$ distribution. Thus, if n is large, the probability of rejecting the null hypothesis $\mu_Y = 21.5$ at the 5% level is 5%. This implies that in 95% of all samples, the confidence set will contain the true value of μ_Y.

Nevertheless, constructing a confidence set by testing all possible values of μ_Y as null hypothesis is not practical, and there is a much easier approach. Using the definition of the t-statistic in (25), a trial value of $\mu_{Y,0}$ is rejected at the 5% level if it is more than 1.96 standard errors away from $\overline{Y}$. Thus, the set of values of μ_Y that are not rejected at the 5% level consists of those values within $\pm 1.96 \times SE(\overline{Y})$ of $\overline{Y}$. That is, a 95% confidence interval for μ_Y is

$$\overline{Y} - 1.96 \times SE(\overline{Y}) \leq \mu_Y \leq \overline{Y} + 1.96 \times SE(\overline{Y}). \tag{30}$$

Analogously, the 90% confidence interval for μ_Y is $\{\overline{Y} \pm 1.64 \times SE(\overline{Y})\}$, and the 99% confidence interval is $\{\overline{Y} \pm 2.58 \times SE(\overline{Y})\}$.

As an example, let's consider the problem of constructing a 95% confidence

interval for the mean hourly earnings of recent college graduates using a random sample of 200 recent college graduates where $\overline{Y} = \$22.64$ and $SE(\overline{Y}) = 1.28$. The 95% confidence interval for mean hourly earnings is

$$22.64 \pm 1.96 \times 1.28 = 22.64 \pm 2.51 = [\$20.13, \$25.15]\,. \tag{31}$$

Now, let's consider the following statements:

1. In repeated sampling, the interval $\{\overline{Y} \pm 1.96 \times SE(\overline{Y})\}$ covers the true value of μ_Y with a probability of 95%.

2. The interval $22.64 \pm 1.96 \times 1.28 = [20.13, 25.15]$ covers the true value of μ_Y with a probability of 95%.

While the first statement is right, the second is wrong. The difference is that while in the first statement we define a random variable (the bounds of the interval depend on $\overline{Y}$, which is random), the second presents only one possible outcome of this random variable, so we cannot make any probabilistic statement about it. Either the computed interval does cover μ_Y or does not.

VI. Comparing Means from Different Populations

Some questions in statistics and econometrics involve comparing the means of two different population distributions. This is what we need to do if we are interested, for instance, on whether male and female college graduates earn the same amount of money on average. In this section we are going to learn how to test hypotheses and how to construct confidence intervals for the difference in the means from two different populations.

A. Hypothesis tests for the difference between two means

To illustrate a ***test for the difference between two means***, let μ_w be the mean hourly earnings in the population of women recently graduated from college and let μ_m be the population mean for recently graduated men. Let's consider the null hypothesis that mean earnings for these two populations differ by a certain amount, say d_0. Then the null hypothesis and the two-sided alternative hypothesis are

$$H_0 : \mu_m - \mu_w = d_0 \text{ vs. } H_1 : \mu_m - \mu_w \neq d_0. \tag{32}$$

The null hypothesis that men and women in these populations have the same mean earnings corresponds to H_0 in the above expression, with $d_0 = 0$.

These population means are unknown, so they must be estimated from samples of men and women. Let's suppose that we have samples of n_m men and n_w women drawn at random from their populations. Let the sample average annual earnings be $\overline{Y}_m$ for men and $\overline{Y}_w$ for women. Then an estimator of $\mu_m - \mu_w$ is $\overline{Y}_m - \overline{Y}_w$.

To test the null hypothesis that $\mu_m - \mu_w = d_0$ using $\overline{Y}_m - \overline{Y}_w$, we need to know the distribution of $\overline{Y}_m - \overline{Y}_w$. According to the central limit theorem, $\overline{Y}_m$ is approximately distributed $\mathcal{N}(\mu_m, \sigma_m^2/n_m)$, where σ_m^2 is the population variance of earnings for men. Similarly, $\overline{Y}_w$ is approximately distributed $\mathcal{N}(\mu_w, \sigma_w^2/n_w)$, where σ_w^2 is the population variance of earnings for women. Besides, we know that the weighted average of two normal random variables is itself normally distributed. On top of that, $\overline{Y}_m$ and $\overline{Y}_w$ are constructed from different randomly selected samples, so they are independent random variables. Therefore, $\overline{Y}_m - \overline{Y}_w$ is distributed $\mathcal{N}\left(\mu_m - \mu_w, (\sigma_m^2/n_m) + (\sigma_w^2/n_w)\right)$.

If σ_m^2 and σ_w^2 are known, then this approximate normal distribution can be used to compute p-values for the test of the null hypothesis that $\mu_m - \mu_w = d_0$. In practice, however, these population variances are typically unknown so they must be estimated. As before, they can be estimated using the sample variances s_m^2 and s_w^2 of mean earnings for the samples of men and women, respectively. Thus, the standard error of $\overline{Y}_m - \overline{Y}_w$ is

$$SE\left(\overline{Y}_m - \overline{Y}_w\right) = \sqrt{\frac{s_m^2}{n_m} + \frac{s_w^2}{n_w}}. \tag{33}$$

The t-statistic for testing the null hypothesis is constructed analogously to the t-statistic for testing a hypothesis about a single population mean, by subtracting the null hypothesized value of $\mu_m - \mu_w$ from the estimator $\overline{Y}_m - \overline{Y}_w$ and dividing the result by the standard error of $\overline{Y}_m - \overline{Y}_w$:

$$t = \frac{\left(\overline{Y}_m - \overline{Y}_w\right) - d_0}{SE\left(\overline{Y}_m - \overline{Y}_w\right)}. \tag{34}$$

If both n_m and n_w are large, then the distribution of this t-statistic can be approximated by the standard normal distribution, and the p-value of the two-sided test is computed exactly as it was in the case of a single population, as in (26).

To conduct a test with a prespecified significance level, we can simply calculate the t-statistic in (34) and compare it to the appropriate critical value. For example, the null hypothesis is rejected at the 5% significance level if the absolute value of the t-statistic exceeds 1.96.

If the alternative is one-sided rather than two-sided (that is, if the alternative is

that $\mu_m - \mu_w > d_0$), then the test can be modified as for one-sided alternatives in Section IV.H.

B. Confidence intervals for the difference between two population means

Similarly, the method for constructing confidence intervals summarized in Section V extends to the difference between the means, $d = \mu_m - \mu_w$. Because the hypothesized value d_0 is rejected at the 5% level if $\mid t \mid > 1.96$, d_0 will be in the confidence set if $\mid t \mid \leq 1.96$. Note that $\mid t \mid \leq 1.96$ means that the estimated difference $\overline{Y}_m - \overline{Y}_w$ is less than 1.96 standard errors away from d_0. Thus, the 95% two-sided confidence interval for d consists of those values of d within ± 1.96 standard errors of $\overline{Y}_m - \overline{Y}_w$:

$$\{\overline{Y}_m - \overline{Y}_w \pm 1.96 \times SE\left(\overline{Y}_m - \overline{Y}_w\right)\}.$$

VII. Linear Regression with One Regressor

In this section we introduce the linear regression model relating one variable X to another Y. For example, if a school district cuts the size of its elementary school classes, what is the effect on its students' standardized test scores? If a person completes one more year of college, what is the effect on her future earnings? Both questions are about the unknown effect of changing one variable X (class size, years of education) on another variable Y (student test scores, earnings). The linear regression model postulates a linear relationship between X and Y, with the slope of the line relating these variables representing the effect that a one-unit change in X has in Y. Just as the mean of Y is an unknown characteristic of the population distribution of Y, the slope of the line relating X and Y is an unknown characteristic of the population joint distribution of X and Y. The problem that econometricians care about is to estimate this slope using a sample of data on these two variables.

A. The linear regression model

Let's go back to one of the questions posed before: if school class sizes are cut in a district, what will the effect be on standardized test scores? Test scores are just a way of measuring students' performance, and as such, policy-makers, school boards and families might be interested in this question. In order to come up with a precise answer, we need to make a quantitative statement about changes: if the class size is changed by a certain amount, what would the change in standardized test scores

be? Formally, this effect would be

$$\beta_1 = \frac{\Delta \text{TestScore}}{\Delta \text{ClassSize}}, \tag{35}$$

where Δ stands for "change in". That is, β_1 is the change in test scores that results from changing the class size by one student. For instance, if $\beta_1 = -0.6$, then a reduction in the class size of two students per class would yield a predicted change in test scores of $(-0.6) \times (-2) = 1.2$, so test scores would rise by 1.2 points as a result of the reduction in the class size by two students per class.

We can note that Equation (35) is nothing but the definition of the slope of a straight line relating test scores and class size. This straight line can be written as

$$\text{TestScore} = \beta_0 + \beta_1 \times \text{ClassSize}, \tag{36}$$

where β_0 is the intercept of the line and β_1 is the slope. According to this equation, if we knew β_0 and β_1, we would not only be able to determine the change in test scores associated with a change in class size, but we would also be able to predict the average test score itself for a given class size.

Of course, class size is just one of many facets of elementary education. Two districts with the same class sizes will have different test scores for many reasons (teachers quality, textbooks, family income, etc.), and even if the districts were comparable in all aspects, they might still have different test scores for essentially random reasons having to do with the performance of each student on the test day. That is why Equation (36) will not hold exactly for all districts. Instead, it should be viewed as a statement about a relationship that holds *on average* across the population of districts. If we want to write a version of this linear relationship that holds for each district, we must incorporate these other factors influencing test scores, including the unique characteristics of each district. For now, we can simply group these "other factors" and rewrite the relationship as

$$\text{TestScore} = \beta_0 + \beta_1 \times \text{ClassSize} + \text{OtherFactors}. \tag{37}$$

In general, let's suppose that we have a sample of n districts. Let Y_i be the average test score in the ith district, X_i be the average class size in the ith district, and u_i denote other factors influencing the test score in the ith district. Then, we can write the previous equation more generally as

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \tag{38}$$

for each district $i = 1, ..., n$. This is the ***linear regression model with a single regressor***, in which Y is the ***dependent variable*** and X is the ***independent variable*** or the ***regressor***.

The first part of (38), $\beta_0 + \beta_1 X_i$ is the ***population regression line*** or the *population regression function*. This is the relationship that holds between Y and X on average over the population. Thus, if we knew the value of X, we could predict the value of the dependent variable Y as $\beta_0 + \beta_1 X$.

The ***intercept*** β_0 and the ***slope*** β_1 are the ***coefficients*** (a.k.a. the *parameters*) of the population regression line. The slope β_1 is the change in Y associated with a unit change in X. The intercept β_0 is the value of the population regression line when $X = 0$, i.e., the point at which the population regression line intersects the Y axis. In some contexts, the intercept has a meaningful economic interpretation, but in others, it has no real-world meaning. For example, when X is the class size, strictly speaking the intercept is the predicted value of test scores when there are no students in the class. In a case like this, it is best to think of the intercept just as the coefficient that determines the level of the regression line.

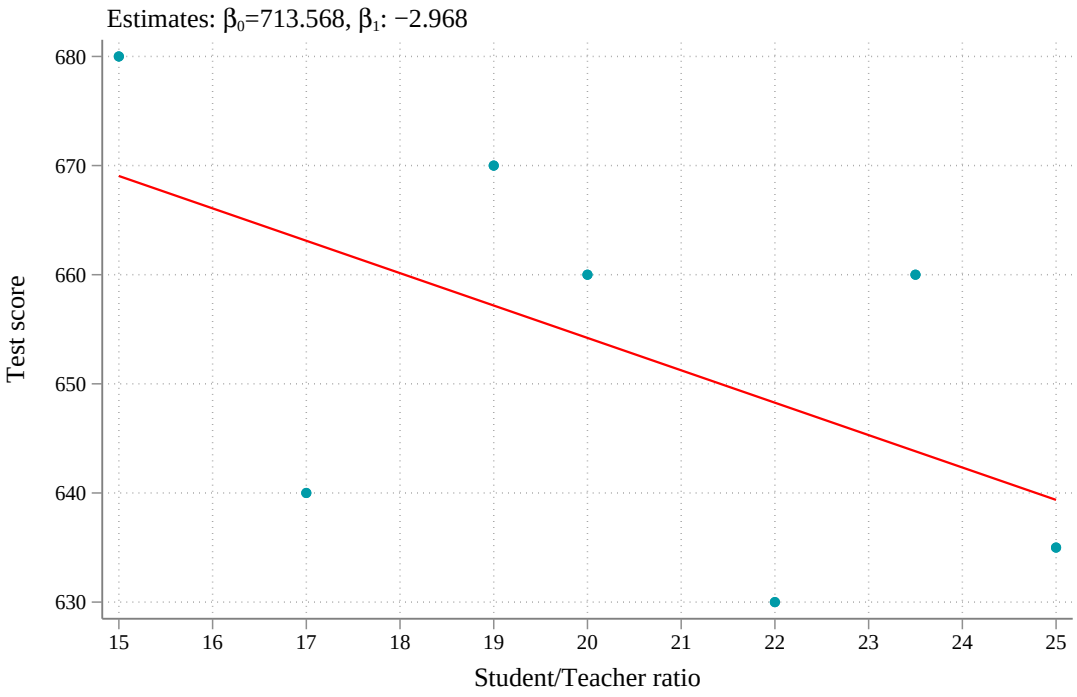
The term u_i is the ***error term***. The error term incorporates all the factors responsible for the difference between the ith district's average test score and the value predicted by the population regression line, that is, all the factors besides X that determine the value of the dependent variable Y for a specific observation i. In the class size example, these other factors include all the unique features of the ith district that we are not modeling explicitly but affect the performance of its students on the test.

Figure 13 summarizes the linear regression model with a single regressor for seven hypothetical observations on test scores (Y) and class size (X). The population regression line is $\beta_0 + \beta_1 X$, which slopes down ($\beta_1 < 0$), meaning that districts with lower student-teacher ratios (smaller classes) tend to have higher test scores. Because of other factors that determine test performance but are not modeled explicitly, the hypothetical observations in the plot do not fall exactly on the population regression line. For those values of Y that are above the population regression line, test scores are better than predicted by the regression line, so the error term for those districts is positive. The opposite happens for those districts with an observed test score that is lower than predicted ($u_i < 0$).

B. Estimating the coefficients of the linear regression model

In a practical situation such as the application to class size and test scores, the intercept β_0 and slope β_1 of the population regression line are unknown. Therefore,

FIGURE 13: TEST SCORE VS. STUDENT-TEACHER RATIO



we must use data to estimate these parameters. This estimation problem is similar to others we face in statistics. For example, before we talked about comparing the mean earnings of men and women who recently graduated from college. As we said, although the population mean earnings are unknown, we can estimate the population means using a random sample of male and female college graduates. Then the natural estimator of the unknown population mean earnings for women is the average earnings of the female college graduates in the sample. The same idea extends to the linear regression model. Although β_0 and β_1 are unknown, we can learn about them using a sample of data.

To illustrate this problem, we are going to analyze data on test scores and class sizes in 420 California school districts that serve kindergarten through eight grade. The test score is the districtwide average of reading and math scores for fifth graders. Class size is measured as the number of students in the district divided by the number of teachers, i.e., the districtwide student-teacher ratio.

Table 2 summarizes the distributions of test scores and class sizes for this sample. The average student-teacher ratio is 19.6 students per teacher, and the standard deviation is 1.9 students per teacher. Looking at the percentiles, only 10% of the districts have student-teacher ratios below 17.3, while the district at the 90th percentile

has a student-teacher ratio of 21.9.

Table 2: Summary of the Distribution of Student-Teacher Ratios and Test Scores

		Standard	Percentile				
	Average	Deviation	10%	25%	50%	75%	90%
Student-teacher ratio	19.6	1.9	17.3	18.6	19.7	20.9	21.9
Test Score	654.2	19.1	630.4	640	654.4	666.7	679.1

To get a sense of how the two variables relate in this sample, we can plot test scores against student-teacher ratios, as we do in Figure 14. The sample correlation coefficient is -0.23, indicating a weak negative relationship between the two variables. Although larger classes in this sample tend to have lower test scores, there are other determinants of test scores that keep the observations from falling perfectly along a straight line. Despite this low correlation, if one could draw a straight line through these data, then the slope would be an estimate of β_1. But, how can we choose among the many possible lines? By far the best way is to choose the line that produces the least squares fit to the data, that is, to use the OLS estimator.

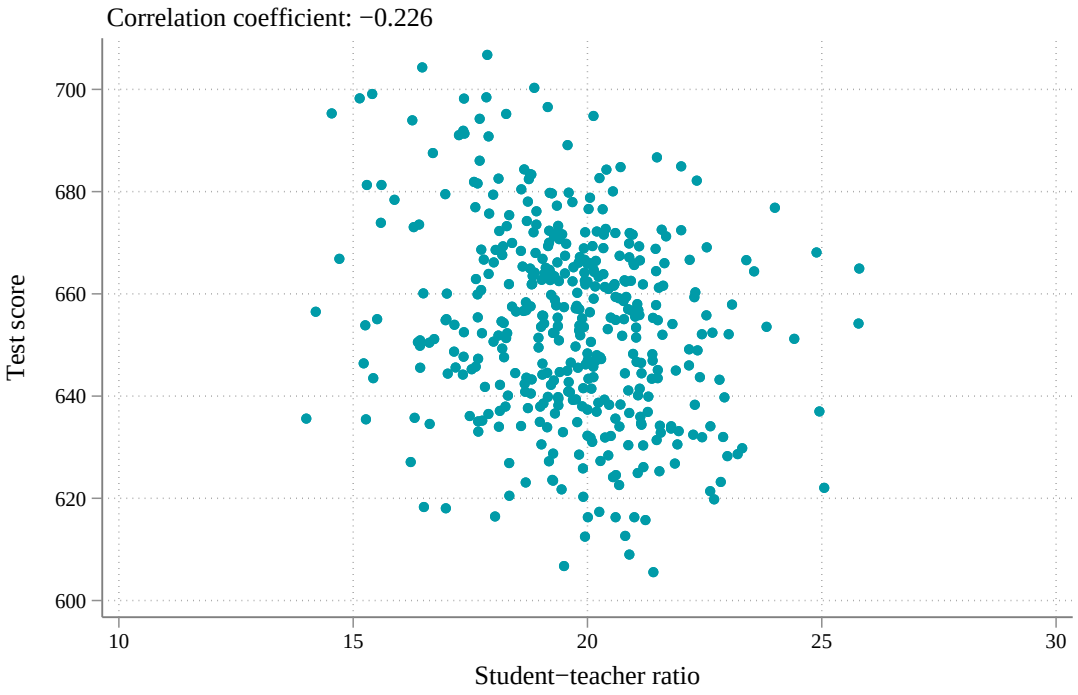
The **ordinary least squares (OLS)** estimator chooses the regression coefficients so that the estimated regression line is as close as possible to the observed data, where closeness is measured by the sum of the squared errors made in predicting Y given X. As we said in Section III.B, the sample average $\overline{Y}$ is the the least squares estimator of the population mean $\mathbb{E}[Y]$, because it minimizes the total squared estimation errors $\sum_{i=1}^{n}(Y_i - m)^2$ among all possible estimators m. The OLS estimator extends this idea to the linear regression model.

Formally, let b_0 and b_1 be some estimators of β_0 and β_1. The regression line based on these estimators is $b_0 + b_1 X$. Then, the error made in predicting the ith observation is $Y_i - (b_0 + b_1 X_i)$. The sum of these squared prediction errors over all n observations is

$$\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2, \tag{39}$$

which is just an extension of the sum of the squared errors for the problem of estimating the population mean (in fact, if there was no regressor, the two problems would be identical). Just as there is a unique estimator $\overline{Y}$ that minimizes the sum of squared errors when predicting the population mean, so is there a unique pair of estimators of β_0 and β_1 that minimize (39). We denote these estimators $\hat{\beta}_0$ and $\hat{\beta}_1$.

Figure 14: Scatterplot of Test Score vs. Student-Teacher Ratio (California School District Data)



They can be computed as

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} = \frac{s_{XY}}{s_X^2} \tag{40}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}. \tag{41}$$

Proof. The OLS estimator is the solution to

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2. \tag{42}$$

We can solve this problem by setting the partial derivatives equal to zero:

$$\frac{\partial}{\partial \hat{\beta}_0}\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2 = -2\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right) \tag{43}$$

$$\frac{\partial}{\partial \hat{\beta}_1}\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2 = -2\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right) X_i \tag{44}$$

By the fact that $\sum_{i=1}^{n} Y_i = n\overline{Y}$ and rearranging terms, we can write (43) as

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}. \tag{45}$$

Meanwhile, we can rearrange (44) to write

$$\sum_{i=1}^{n}\left(X_i Y_i - \hat{\beta}_0 X_i - \hat{\beta}_1 X_i^2\right) = 0. \tag{46}$$

Substituting our result for $\hat{\beta}_0$ gives us

$$\sum_{i=1}^{n}\left(X_i Y_i - \left(\overline{Y} - \hat{\beta}_1 \overline{X}\right) X_i - \hat{\beta}_1 X_i^2\right) = 0, \tag{47}$$

and distributing the sum to each term, we get

$$\sum_{i=1}^{n} X_i Y_i - \overline{Y}\sum_{i=1}^{n} X_i + \hat{\beta}_1 \overline{X}\sum_{i=1}^{n} X_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 = 0, \tag{48}$$

which can be rewritten as

$$\sum_{i=1}^{n} X_i Y_i - n\bar{Y}\bar{X} + \hat{\beta}_1 n\overline{X}^2 - \hat{\beta}_1 \sum_{i=1}^{N} X_i^2 = 0. \tag{49}$$

Solving for $\hat{\beta}_1$,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{Y}\bar{X}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}. \tag{50}$$

We can note that the numerator is equal to $\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)$, and the denominator is equal to $\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$. Substituting these above gives the result we were looking for, and dividing the numerator and the denominator by n gives $\hat{\beta}_1 = s_{XY}/s_X^2$. □

Therefore, the ***OLS regression line***, also called the *sample regression line* or *sample regression function*, is the straight line constructed using the OLS estimator: $\hat{\beta}_0 + \hat{\beta}_1 X$. The ***predicted value*** of Y_i based on the OLS regression line is

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \tag{51}$$

and the ***residual*** for the ith observation is the difference between Y_i and its predicted value:

$$\hat{u}_i = Y_i - \hat{Y}_i. \tag{52}$$

We also say that $\hat{\beta}_0$ and $\hat{\beta}_1$ are the sample counterparts of the population coefficients β_0 and β_1. Similarly, the OLS regression line $\hat{\beta}_0 + \hat{\beta}_1 X$ is the sample counterpart of the population regression line $\beta_0 + \beta_1 X$, and the OLS residuals $\hat{u}_i$ are the sample counterparts of the population errors u_i.

When OLS is used to estimate a line relating the student-teacher ratio to test scores using the 420 observations in Figure 14, the estimated slope is -2.28 and the estimated intercept is 698.9. Accordingly, the corresponding OLS regression line is

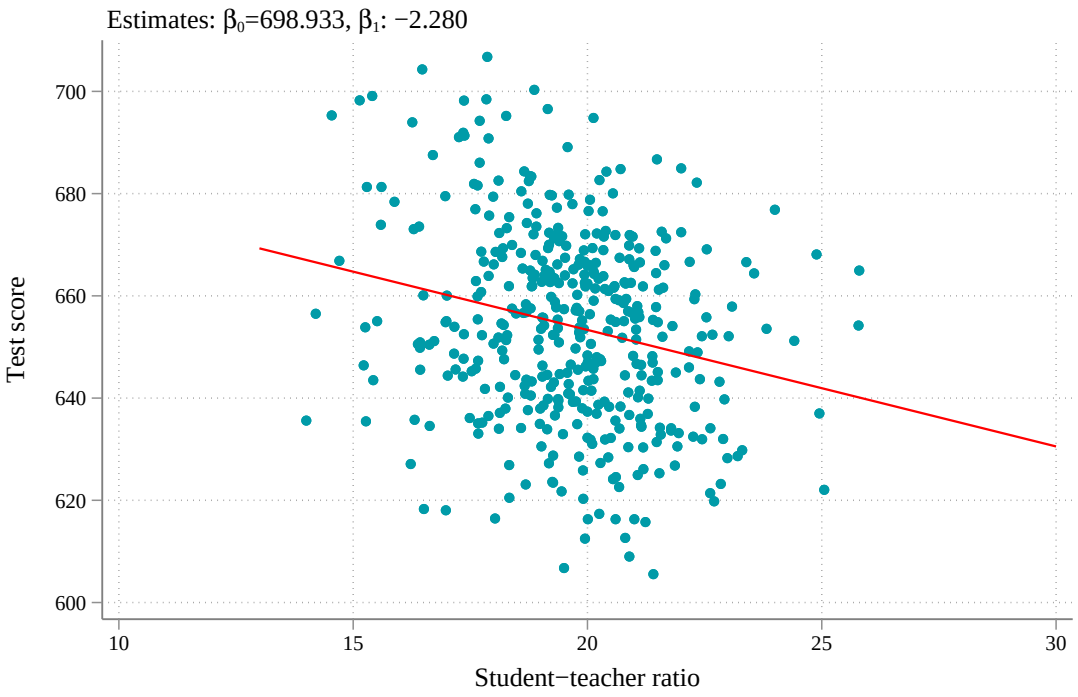
$$\hat{Y}_i = 698.9 - 2.28 X_i, \tag{53}$$

where $\hat{Y}_i$ is the average test score in the ith district and X_i is the student-teacher ratio. In Figure 15 we plot this OLS regression line together with the scatterplot previously shown.

The slope of -2.28 means that an increase in the student-teacher ratio by one student per class is, on average, associated with a decline in districtwide test scores by 2.28 points on the test. If the student teacher-ratio decreases by two students per class, there would be an associated increase in test scores of $(-2) \times (-2.28) = 4.56$, on average. Thus, the negative slope indicates that more students per teacher (larger classes) results in poorer performance on the test.

Using these regression results, we can also predict the districtwide test score given a value of the student-teacher ratio. For example, for a district with 20 students per

FIGURE 15: ESTIMATED REGRESSION LINE FOR CALIFORNIA SCHOOL DISTRICT DATA



teacher, the predicted test score is $\hat{Y}_i = 698.9 - 2.28 \times 20 = 653.3$. Of course, this prediction will not be exactly right because of the other factors that determine a district's performance, but it is the best prediction we can do given the information available.

Now, is this estimate of the slope large or small? Looking at Table 2, if we take a district with a student-teacher ratio and average test score equal to the median (19.7 and 654.4), decreasing the class size by 2 would move the district from the 50th percentile to very near the 10th percentile. As we have just seen, that would result in test scores increasing by approximately 4.6 points, that is, the district would move from the 50th to just short of the 60th percentile. Therefore, cutting the student-teacher ratio by two students per teacher would help and might be worth doing, but it would not be a panacea.

What if we consider a far more radical change, such as reducing the student-teacher ratio from 20 to 5? Unfortunately, our estimates would not be very useful. As we can see in Figure 15, the smallest student-teacher ratio in our dataset is 14. We do not have information on how districts with extremely small classes perform, so our data alone are not a reliable basis for predicting the effect of a radical move to such an extremely low student-teacher ratio.