

# 1. Review of Probability and Statistical Inference

Erick F Molina

Summer, 2024

## 1 Introduction

Throughout the applied microeconometrics course, we will use different concepts and tools that we have learned in previous courses. For this reason, this document will present a brief review of some topics in probability and statistics. First, we will analyze what lies behind random variables and then we will study what happens when we do not know the entire population and only have a sample from it.<sup>1</sup>

## 2 Random variables

**Definition 1.** A **random variable**  $X$  is a function that assigns a numerical value to the outcome of a random event. In mathematical terms, it is a function from a sample space  $S$  to the real numbers  $\mathbb{R}$ .

Generally, we talk about two types of random variables: discrete and continuous. Discrete variables take their values in a discrete set, which is usually a small number of distinct values. For example, the outcome of a die can only be an integer between 1 and 6. Continuous variables, on the other hand, take values in a continuous set. An example of a continuous random variable is the price of gasoline.

For each random variable, we can create a function that quantifies the uncertainty of the outcome of the experiment. We call this function the *probability distribution function* and it is defined on the interval between 0 and 1.<sup>2</sup> Thanks to the axioms of probability, we do not need to write the probability function in full; instead, we can use a *probability mass function* or a *probability density function*, depending on whether the variable is continuous or discrete.

---

<sup>1</sup>This document is based on the bibliography I describe in the course syllabus and on notes made by Xiaoxia Shi, Arturo Aguilar, Mauricio Romero, Laura Schechter, Shaun McRae, Kirill Borusyak, among others. Any remaining errors are mine.

<sup>2</sup>Formally, these functions satisfy the three axioms of probability: 1)  $P(A) \geq 0$ , 2)  $P(S) = 1$ , and 3) finite additivity.

**Definition 2.** The **probability mass function (PMF)** of  $X$ , a discrete variable, is a function that describes the probability that  $X$  is equal to a certain value  $x$ . That is,  $f_X = P(X = x)$ . If  $f_X$  is the PMF of  $X$ , we say that  $X$  follows the distribution  $f_X$ . The notation we use is  $X \sim f_X$ .

The PMF is not the only function with which we can describe the distribution of a random variable. We can also describe it using the *cumulative distribution function*.

**Definition 3.** The **cumulative distribution function (CDF)** of  $X$  is a function such that  $F_X(x) = P(X \leq x)$ . If the CDF of  $X$  is  $F_X(x)$ , we can also say that  $X \sim F_X$ .

There is a relationship between the PMF and the CDF, since  $F_X(x) = \sum_{x' \leq x} f_X(x')$ . In the same way, one can go from the CDF to the PMF. It is worth remembering that up to this point we have only analyzed discrete random variables.

The advantage of working with CDFs is that they can be used for continuous random variables. However, we cannot use the PMF for the latter. If a random variable  $X$  is continuous, then  $P(X = x) = 0$ . What we can do is compute the probability that  $X$  lies within some range of values. To do so, we use the *probability density function*.

**Definition 4.** The **probability density function (PDF)** for a continuous random variable  $X$  with CDF  $F_X$  is a non-negative function  $f$  that satisfies the following equation for all values of  $x \in \mathbb{R}$ :

$$F_X(x) = \int_{-\infty}^x f(t) dt \quad (1)$$

## 2.1 Joint distributions

In most cases, we do not analyze only one variable, but several. Therefore, we usually refer to a joint density function. Suppose we have two discrete random variables,  $X$  and  $Y$ . The joint density function is the probability that each one takes a certain value at the same time, that is,  $f_{X,Y}(x, y) = P(X = x, Y = y)$ . If we want to know how one of them is distributed, we can obtain the marginal density function  $f_X(x) = P(X = x)$ .

In addition, sometimes we are interested in how a variable  $Y$  behaves when a variable  $X$  is already given. For that, we use the conditional density function, which is given by  $f_{X|Y}(y|x) = P(Y = y | X = x) = \frac{P(Y=y, X=x)}{P(X=x)} = \frac{f_{X,Y}(x,y)}{f_X(x)}$ . It may also happen that knowing the value of one variable tells us nothing about the other. This situation is called *independence*. We say that a variable  $Y$  is independent of another variable  $X$  if

$$P(Y = y | X = x) = P(Y = y) \quad (2)$$

Similarly,  $X$  and  $Y$  are independent if

$$P(X = x, Y = y) = P(X = x)P(Y = y) \quad (3)$$

## 2.2 Expected value, variance, and other moments

Once we know that random variables follow a distribution, the next natural question is: what can we say about the distribution of a variable?

**Definition 5.** The **expected value** of a random variable  $X$  is the average value that  $X$  takes after many realizations. In other words, it is a weighted average of the possible values of  $X$ , where the weights are given by the PMF.

For discrete variables, the expected value is computed as follows:

$$\mathbb{E}[X] = \sum_x x f_X(x) \quad (4)$$

For continuous variables, the expected value is computed as follows:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx \quad (5)$$

Those of us who work constantly with expected values tend to call it by different names. Among them are: expectation, mean, first moment, or we use the Greek letter  $\mu$ .

Likewise, expected values have properties that will be useful throughout the course. For each constant  $a$  and  $b$  and random variables  $X$  and  $Y$ :

1.  $\mathbb{E}[a] = a$
2.  $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$
3.  $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$

In addition to expected value, there are other measures from which we can obtain information about distributions. There is a set of measures about the distribution of a random variable that are called moments.

**Definition 6.** The  $m$ -th **moment** of  $X$  is  $\mu_m = \mathbb{E}[X^m]$ , and the  $m$ -th central moment of  $X$  is  $\mathbb{E}[X - \mathbb{E}[X]]^m$ .

**Definition 7.** The **variance** is a measure that quantifies the dispersion of a variable around its mean. Clearly, variance is the second central moment:

$$\sigma^2(X) = Var(X) = \mathbb{E}[X - \mathbb{E}[X]]^2 = \mathbb{E}[X^2] - (\mathbb{E}[X])^2. \quad (6)$$

We usually use expected value and variance to summarize the central tendency and dispersion of a distribution. However, variance does not behave in the same way as expected value, since it is not a linear operator. The properties of variance are as follows:

For each constant  $a$  and  $b$  and random variables  $X$  and  $Y$ :

1.  $Var(a) = 0$
2.  $Var(aX + b) = a^2Var(X)$
3.  $Var(aX + bY) = a^2Var(X) + b^2Var(Y) + 2abCov(X, Y)$

Sometimes, variance results in very large numbers and is difficult to interpret. It is also a nonlinear measure. To have a linear measure, we commonly use the square root of the variance, which we call the *standard deviation*:  $\sigma_X = \sqrt{Var(X)}$ .

Returning to the analysis of two variables, the third property of variance implies the covariance between two variables. Covariance measures the relationship between two variables. If covariance is positive, then both variables tend to move in the same direction.  $\sigma_{XY} = Cov(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$ . When two variables  $X$  and  $Y$  are independent,  $Cov(X, Y) = 0$ .<sup>3</sup>

The units of covariance are the units of  $X$  multiplied by the units of  $Y$ . So we could be comparing apples and oranges. To better interpret the relationship between  $X$  and  $Y$ , we can use the correlation coefficient. The correlation of  $X$  and  $Y$  is covariance rescaled by the standard deviations of both variables:  $\rho_{XY} = Corr(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$ .

---

<sup>3</sup>Note: This only holds in one direction. If  $Cov(X, Y) = 0$ , we cannot ensure that two variables are independent.

### 3 Sampling

In the real world, which is what we do econometrics with, it is very difficult to obtain all the data from a *population*. Commonly, we only have a set of observations on different individuals and/or time periods. A set of observations is called a *sample*. The way in which the sample is collected affects the method for doing inference and the properties of estimators.

In general, talking about sampling is very complex. Since this course does not focus on sampling, we will assume that the samples we work with are representative of the population. This means that the sample behaves very similarly to how the population behaves.

When a sample is not representative, a sampling bias is incurred. This bias leads us to make incorrect inference about the data. To evaluate whether there is sampling bias, it is important to know the population and have a way to compare with other representative samples from the same population.

**Definition 8.** A set of observations is a **random sample**  $\{X_1, \dots, X_n\}$  from a population  $F$  if they are mutually independent with identical marginal distributions  $F$ , where  $n$  is the sample size. In this case we also say that  $\{X_1, \dots, X_n\}$  are i.i.d. (independent and identically distributed).

Once we have a random sample, we can use it to answer some questions. In particular, we can use *statistics*, which are functions of the sample, to learn something about a *parameter* of the population. Analogously, a parameter is a function of the population. Since we do not have the complete population, we need to use statistics to construct estimators of parameters.

**Definition 9.** A (point) **estimator**  $\hat{\theta}$  of a parameter  $\theta$  is a statistic that serves as an approximation to  $\theta$ . An estimator is a function that uses the elements of the sample as inputs, not a specific value. When we evaluate an estimator with given data, we obtain an **estimated value**.

For example, it is common to use  $\bar{X}_n$  as an estimator of  $\mu$ . This is the most natural analog estimator, since  $\bar{X}_n$  is the mean of the sample, while  $\mu$  is the mean of the population.

There is no such thing as a unique estimator to approximate a parameter. For this reason, it is necessary to have tools to evaluate which estimator works better than others. The two most popular tools we use are unbiasedness and efficiency, which are properties of the estimators themselves.

**Definition 10.** We say that an estimator is **unbiased** if its bias is equal to zero. The **bias** of an estimator is defined as

$$Bias(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta \quad (7)$$

**Definition 11.** We say that an estimator  $\hat{\theta}_1$  is more **efficient** than another  $\hat{\theta}_2$  if:  
 $Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$ .

Since an estimator is a statistic, which in turn is a function of random variables, an estimator is also a random variable. Being a random variable, we can try to understand its distribution. There are different ways to approach this question, for example: computing its bias, deriving its exact distribution under normality, finding its asymptotic distribution, running a bootstrap simulation, etc. In this document, we will focus on asymptotic theory, since it will be useful for doing econometrics.

### 3.1 Asymptotic theory

An estimator can have asymptotic properties. These properties only apply when estimators come from samples with very large  $n$ . Even so, there is no convention on the size of  $n$  for a sample to be considered large. For most of the analysis we will do in this course, we will say that samples are large enough to apply the asymptotic properties of estimators.

Before defining asymptotic properties of estimators, it is necessary to recall the following concepts:

**Definition 12.** A sequence of random variables  $\{Z_n\}$  **converges in probability** to  $Z$  as  $n \rightarrow \infty$  if for every  $\varepsilon > 0$ , we have:

$$\lim_{n \rightarrow \infty} P(|Z_n - Z| \geq \varepsilon) = 0 \quad (8)$$

To denote convergence in probability, we write  $Z_n \rightarrow_p Z$ .

**Definition 13.** If an estimator  $\hat{\theta}$  of  $\theta$  satisfies  $\hat{\theta} \rightarrow_p \theta$  as  $n \rightarrow \infty$ , then  $\hat{\theta}$  is a **consistent** estimator of  $\theta$ . In other words, if  $n$  is large enough, then  $\hat{\theta}$  should be very close to  $\theta$  with high probability, so it can be interpreted as  $\theta$ .

**Definition 14.** A sequence of random variables  $Z_n$  **converges in distribution** to  $Z$  as  $n \rightarrow \infty$  if

$$F_n(x) = P(Z_n \leq x) \rightarrow F(x) = P(Z \leq x) \quad (9)$$

at all points of continuity of  $F$ . To denote convergence in distribution, we write  $Z_n \rightarrow_d Z$ .

Now that we know convergence in probability, consistency, and convergence in distribution, we can define some properties of estimators.

**Definition 15. Weak Law of Large Numbers (WLLN).** If  $\{X_1, X_2, \dots, X_n, \dots\}$  are an i.i.d. sequence from  $F$  and  $\mathbb{E}[X_i] < \infty$ , then  $\bar{X}_n \rightarrow_p \mu$  as  $n \rightarrow \infty$ . In other words, if we want to approximate the population mean, we can use the sample mean when we have sufficiently large samples. <sup>4</sup>

The WLLN implies that the sample mean  $\bar{X}_n$  is a consistent estimator of  $\mu$ . From this, we know it is a property of a point estimator, but it does not tell us things about the distribution of the estimator. To talk about the distribution we use the central limit theorem.

**Definition 16. Central Limit Theorem (CLT).** If  $X_i : i = 1, 2, \dots$  are i.i.d. with  $\mathbb{E}[X_i] = \mu$  and  $0 < \text{Var}(X_i) = \sigma^2 < \infty$ , then

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \rightarrow_d Z \sim N(0, 1) \quad (10)$$

In other words, regardless of the population distribution of  $X$ , the standardized version of  $\bar{X}_n$  approaches a standard normal distribution as the sample size increases.

---

<sup>4</sup>To prove the WLLN, we can use Chebyshev's inequality, which we reviewed in the Probability course.

### 3.2 Hypothesis testing

Once we know estimators have properties, we can also formulate hypotheses about them. A hypothesis is a statement that has not been verified and can be tested. To test a hypothesis about population parameters, we can use a sample and construct estimators. Usually, the hypotheses we test are whether a parameter equals a certain value. This type of hypothesis is called the null hypothesis.

In statistics, we talk about *rejecting* or *failing to reject* the null hypothesis. Rejecting the null hypothesis does not mean that the alternative hypothesis is true. Similarly, failing to reject the null hypothesis does not mean that the null hypothesis is true. The only thing we can say is whether the estimated values from the sample are consistent with the hypothesis about the population.

Suppose we have a random sample from a population. Given the sample mean  $\bar{X}_n$ , we want to make inference about the population mean  $\mu$ . For that, we have the **null hypothesis** that the true population mean  $\mu$  is equal to some value  $\mu_0$ , while the **alternative hypothesis** is that  $\mu$  is different from  $\mu_0$ . This is expressed as follows:

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

It is important to note that the goal of hypothesis testing is not to know whether it is true, but to know whether our sample is consistent with it. This is because we do not have observations for the entire population, only a sample. In general,  $\bar{X}_n$  will rarely be equal to the  $\mu_0$  in our hypothesis for two reasons. First,  $\mu$  may truly be different from  $\mu_0$ . Second,  $\mu$  may in fact be equal to  $\mu_0$ , but  $\bar{X}_n$  differs because it comes from a random sample.

We can incur two types of error when trying to test our hypothesis. **Type I error** consists of rejecting the null hypothesis when it is actually true. **Type II error** consists of failing to reject the null hypothesis when it is actually false. In all hypothesis tests there is a *tradeoff* between the probabilities of these two types of error. We call **significance** the probability that a Type I error occurs.<sup>5</sup> Significance levels of 0.1, 0.05, and 0.01 are commonly used in hypothesis testing.

Once we know all this, we can carry out a hypothesis test. Following our assumptions about the sample, we need to build a test statistic that depends on how our estimators behave. Here, we assume that our sample mean  $\bar{X}_n$  has asymptotic properties and is normally distributed. Applying the CLT, we can construct a  $t$  test statistic:

---

<sup>5</sup>The probability of a Type II error is linked to statistical power, which we will study later in the course.

$$t = \frac{(\bar{X}_n - \mu_0)}{\sqrt{\frac{s^2}{n}}} \rightarrow N(0, 1) \quad (11)$$

Substituting the estimated values, we can obtain an estimated  $t$  statistic, which we call  $\hat{t}$ .<sup>6</sup> We compare  $\hat{t}$  to a reference value from the standard normal distribution using a significance level of our choice. If our  $\hat{t}$  is greater than the value given by the standard normal distribution at the chosen significance level, then we say we reject the null hypothesis at that significance level in favor of the alternative hypothesis.

Finally, in other classes we have heard about *p-values*. P-values are used together with significance levels, so it is useful to know what they are.

**Definition 17.** The **p-value** tells us up to what significance level we could reject a null hypothesis. If the significance level is greater than the p-value, the null hypothesis will be rejected.

To conclude, it is necessary to clarify some things about interpreting the results of hypothesis tests. In particular, rejecting the null hypothesis is a strong statement about the inconsistency between the sample data and the hypothesis. However, we cannot say anything definitive if we fail to reject the null hypothesis. For this reason, it is common to choose as the null hypothesis something we expect to be false.

---

<sup>6</sup>It is important to mention that there may be other test statistics that follow distributions different from the normal.