Basic Probability and Statistics for Economic Research

C. Luke Watson
FDIC

This version: March 2025

Abstract

These notes present necessary statistical tools for economic research. The material should be remedial for all those with an undergraduate degree in economics (and/or finance, political science). The aim is to provide vocabulary, notation, and intuition for these concepts for reading other more advanced material. Specifically, the first first of these notes are for a study group for reading Mostly Harmless Econometrics.

Contents

Introduction

This set of notes briefly *re*introduces statistical concepts necessary to understand and conduct economic research. These notes assume the reader has *at least* seen these ideas and answered a question or two on a test. These notes do not aim to be encyclopedic but provide a touch-point for vocabulary to seek additional information if required. Think of them as an extended glossary while reading more advanced material outside of these notes.

Most importantly to note: these notes **will** oversimplify important concepts and in many ways could be viewed as incorrect. However, I believe the notes give just enough of a description to be useful. The reader is encouraged to refine their understanding as they take more classes, read other textbook treatments, or expand their knowledge in other ways.

I am loosely following the outline in the Appendices of 'Baby Wooldridge,' I am borrowing some of the treatment from Herman Bennett's MIT Open Course notes, and I am using claude.ai for some coding and to keep the discussion intuitive.

Part I

Fundamentals of Probability

> Probability is the branch of mathematics concerning events and numerical descriptions of how likely they are to occur.
>
> — Wikipedia: Probability

1 Random Variables and Probability Distributions

1.1 Random Variables

A random variable is a variable whose value is uncertain and determined by chance ('stochastic'). Each possible value has a probability of occurring, and these probabilities form what we call a probability distribution.

An example of a random variable you are already familiar with is the value of flipping a coin. If you pick up a coin and wish to record its values (i.e., the sequence of heads or tails), then you do not know what that sequence is until you start flipping the coin. The values that you get from this sequence of flipping the coin are the realizations of the random variable and the values you get are determined by the probability distribution for that coin. A perfectly balanced and 'fair' coin would yield exactly a one-half chance or heads or tails; however, any given coin might technically have a bias.

A fundamental realization in studying the world is that we can treat observations in real life as if they were random variables. The coin flipping is one example; another is commuting to work. I take the WMATA Blue Line to work from my house. I know that it takes me five minutes to get to the platform and seven minutes to walk from Farragut West to the FDIC lobby (as long as the motorcade is not stopping traffic); however, the train ride can take anywhere from 23 minutes to 30 minutes depending on various factors. So when I start my commute, the total length it takes is unknown to me. The reasons why I am late may be known to someone, but they are certainly not always known to me while I am on the train.

It turns out that I do not really need to know what specific events caused my train to be fast or slow on a given day. However, suppose I have an important meeting early in the morning, and so I want to guess when I should leave for work that day. I could record my

commute times over the course of many days and get some understanding of what are the possible values and relatively likelihoods of different commute times. If the meeting is extremely important (e.g., a meeting with the President), then I might use the maximum length I have ever experienced as my estimate. If the meeting is with my section, then I might use the average length. If it is just a meeting with Troy...

This is the key to using statistics for empirical work. While there may be various reasons for different things, we can often model empirical outcomes as realizations of random variables.

Notation: use upper case letters (e.g., X, Y, Z) be random variables and let lower case (e.g., x, y, z) be particular values of the random variable. Consider again flipping a coin. Let X be the random variable from flipping a coin and so x would be the value of the flip. Suppose we flip the coin ten times, then we would have ten random variables: $(X_1, X_2, \ldots, X_{10})$ and ten realizations $(x_1, x_2, \ldots, x_{10})$. Depending on the context, we might refer to an arbitrary random variable (coin flip) as X_i and its value x_i.

1.2 Probability

An intuitive definition of probability is the likelihood or chance of a particular event occurring. It is a way to quantify uncertainty and express how likely something is to happen. To define a probability, first one must know all the things that can happen, second one must know the chance that one of those things happens. We then normalize the 'size of the chance' so that the sum of every chance is equal to one.

Consider (again) flipping a coin. The only possibly outcomes are heads or tails. For a fair coin, the chances are even for either outcome. We can write this as $\{(H, T); (\frac{1}{2}, \frac{1}{2})\}$. More generally, we could consider a set of N outcomes $\mathcal{A} = (a_1, a_2, a_3, \ldots, a_N)$ and a set of chances $\mathcal{P}_\mathcal{A} = (p_1, p_2, p_3, \ldots, p_N)$, so we would have $\{\mathcal{A}; \mathcal{P}_\mathcal{A}\}$.

There are three rules ('axioms') for probability:

1. $\Pr(a_i) = p_i \geq 0$ – probabilities must be non-negative,

2. $\Pr(\mathcal{A}) = 1$ – the probability that some outcome happens is assured,

3. If $a_i \neq a_j$, then $\Pr(a_i \vee a_j) = \Pr(a_i) + \Pr(a_j)$ – as long as two events are different ('mutually exclusive'), then the probability of either event is the sum of the probability of each event.

1.3 Discrete Random Variables

We will classify two types of random variables: discrete and continuous. It turns out that while the properties are almost the same, they require slightly different mathematics. We first describe the discrete random variables since they are more intuitive, and then we will go to continuous.

Discrete random variables can only take on a finite amount of values. The classic example is (again) heads or tails from flipping a coin. Because the outcomes are finite, then we can explicitly right out the probabilities of each event: $\Pr(X = \text{Heads}) = p_H$ and $\Pr(X = \text{Tails}) = p_T$. Note: by Rules 2 and 3, $p_H + p_T = 1$, so $p_T = 1 - p_H$.

1.3.1 Discrete PDF

A probability density function (PDF) for a discrete random variable is a function that gives the probability of each possible value that the random variable can take. Think of the function as a "lookup table" that tells you how likely each possible outcome is. If we are being very precise, then we denote the PDF for a random variable X as $f_X(x)$, where

$$f_X(x) = \Pr(X = x) \tag{1}$$

$$\forall x \in \mathcal{X}, f_X(x) \geq 0 \tag{2}$$

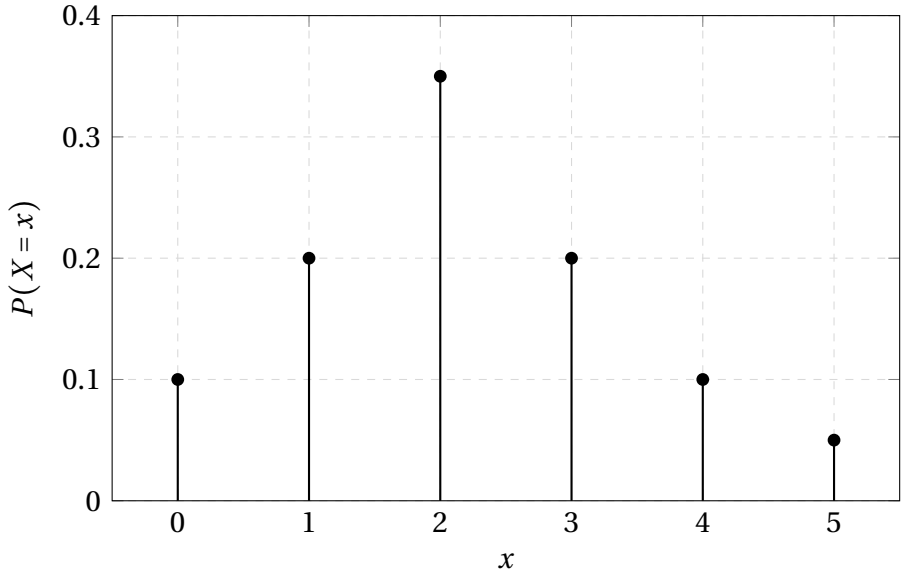
$$\sum_{x \in \mathcal{X}} f_X(x) = 1. \tag{3}$$

However, usually we will just write $f(x)$ rather than always specifying the dependence of the PDF on the random variable X.

Example:

[htbp]

Figure 1: Probability Density Function of a Discrete Random Variable



Note: This figure illustrates a hypothetical PDF for a discrete random variable with values ranging from 0 to 5.

1.3.2 Discrete CDF

A Cumulative Distribution Function (CDF) for a discrete random variable is a function that gives the probability that the random variable takes on a value less than or equal to a specific value. Think of the CDF as a "running total" (i.e., cumulative) that tells you the probability of getting a value up to and including each possible outcome. We specify this as $F_X(x)$, where

$$F_X(x) = \sum_{v \leq x} f_X(v) \tag{4}$$

$$\lim_{x \to \neg\infty} F(x) = 0 \tag{5}$$

$$\lim_{x \to \infty} F(x) = 1 \tag{6}$$

$$v \leq x \implies F_X(v) \leq F_X(x) \tag{7}$$

$$F_X(x) \text{ is right continuous.} \tag{8}$$

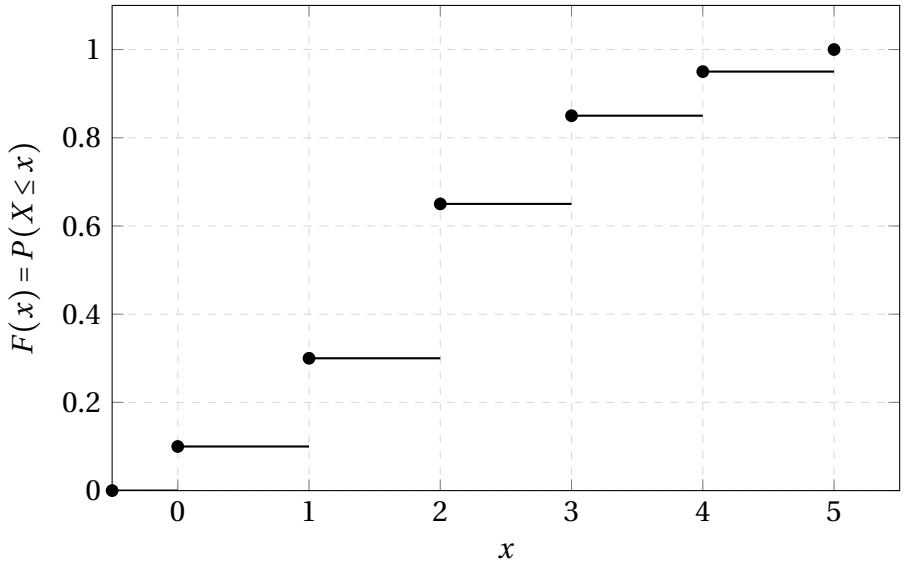
Again, we will usually just write $F(x)$. It is not a coincidence that the PDF is little $f(\cdot)$ and the CDF is big $F(\cdot)$ – more on this with continuous random variables. Two other features

of the CDF:

1. $\Pr(X > x) = 1 - \Pr(X \leq x) = 1 - F(x)$
2. $v \leq x \implies \Pr(v < X \leq x) = F(x) - F(v)$.

[htbp]

Figure 2: Cumulative Distribution Function of a Discrete Random Variable



Note: This figure illustrates the CDF corresponding to the PDF of the discrete random variable with values ranging from 0 to 5.

1.4 Continuous Random Variables

Building on our understanding of discrete random variables, let's now turn to continuous random variables. Unlike their discrete counterparts, continuous random variables can take on any value within a given range, including fractional or irrational numbers. That is, continuous random variables can take on an infinite amount of values. They are used to model quantities that can vary smoothly, such as time, temperature, or prices in financial markets. Going back to my commute example, it might seem like the number of minutes is discrete; however, if we measure the commute in minutes, seconds, or nanoseconds, then it looks like it is continuous.

With continuous random variables, we move from thinking about individual probabilities to considering probability densities over intervals. This shift requires us to use

slightly more complicated mathematical objects: integrals rather than sums when calculating probabilities.

1.4.1 Continuous PDF

A probability density function for a continuous random variable serves a similar role to the PDF for discrete variables, but with some key differences. Rather than giving the probability of exact values, a PDF describes the relative likelihood of the random variable falling within a particular range of values. Intuitively, if there are an infinite number of possible values, then probability of just one value must be zero. However, the probability of a range of values can have mass.

Mathematically, the PDF is a non-negative function whose integral over any interval gives the probability of the random variable taking a value in that interval. The total area under the PDF curve must equal 1, reflecting that the probability of all possible outcomes sums to 100%. Formally, continuous RV PDFs have the following rules:

$$\int_{x_1}^{x_2} f_X(v)\mathrm{d}v = \Pr(X \in (x_1, x_2)) \tag{9}$$

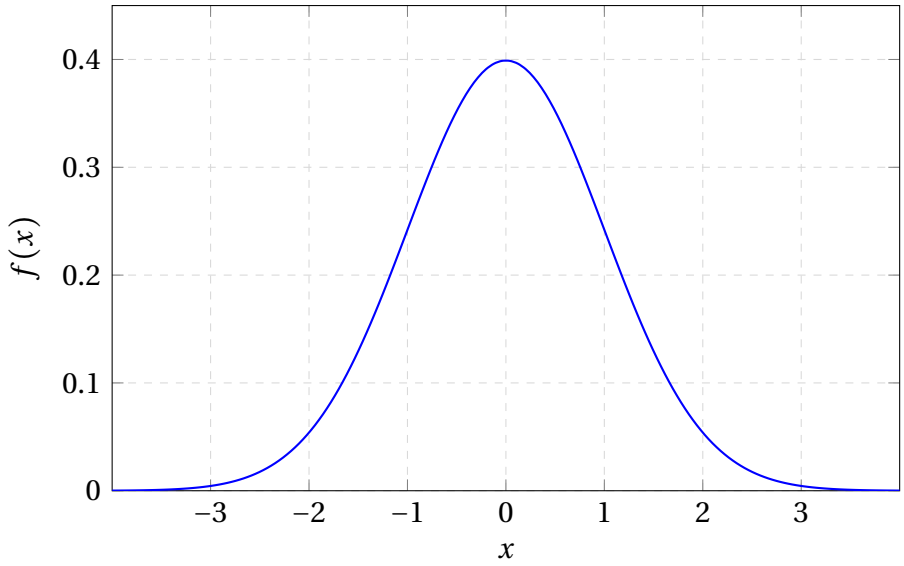
$$\forall x \in \mathcal{X}, f_X(x) \geq 0 \tag{10}$$

$$\int_{\underline{x}}^{\overline{x}} f_X(v)\mathrm{d}v = 1, \tag{11}$$

where $\underline{x} = \inf \mathcal{X}$ and $\overline{x} = \sup \mathcal{X}$.

[htbp]

Figure 3: Probability Density Function of a Continuous Random Variable



Note: This figure illustrates the PDF of a continuous random variable, truncated at $\{-3, 3\}$.

1.4.2 Continuous CDF

The Cumulative Distribution Function (CDF) for a continuous random variable, like its discrete counterpart, gives the probability that the random variable takes on a value less than or equal to a given value. However, for continuous variables, the CDF is a smooth, continuous function rather than a step function. It starts at 0 for the lowest possible value, $\underline{x}$, and increases monotonically to 1 for the highest possible value, $\overline{x}$. The CDF at any point represents the area under the probability density function (PDF) up to that point. Formally,

$$F_X(x) = \Pr(X \leq x) = \int_{\underline{x}}^{x} f_X(v)\mathrm{d}v \tag{12}$$

$$\lim_{x \to \neg\infty} F_X(x) = 0 \tag{13}$$

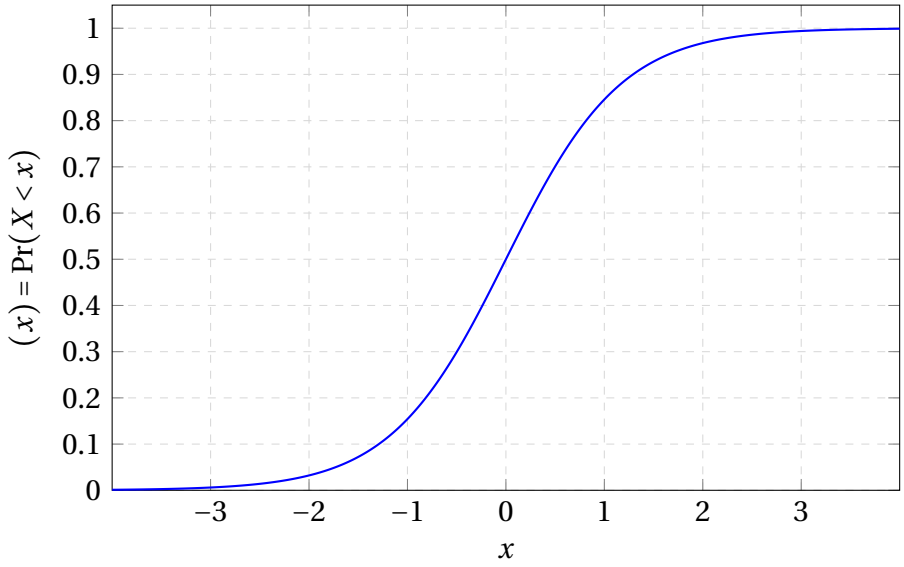
$$\lim_{x \to \infty} F_X(x) = 1. \tag{14}$$

As promised, I said that it was not a coincidence that the PDF is little $f(\cdot)$ and the CDF is big $F(\cdot)$. As can be seen in equation 12, integrating the PDF up to x yields the CDF, and so likewise differentiating the CDF at x yields the PDF: $\frac{\mathrm{d}}{\mathrm{d}x} F_X(x) = f_X(x)$. This provides an intuition for the PDF as the probability of X falling within the infinitesimal interval

$[x, x + \mathrm{d}x]$.

[htbp]

Figure 4: Cumulative Distribution Function of a Continuous Random Variable



Note: This figure illustrates the CDF of a continuous random variables, truncated at $\{-4, 4\}$.

1.5 Multiple Random Variables

We now think about multiple random variables (sometimes called random vectors). Multivariate random variables ("MRVs") extend the concept of single random variables to situations where we are interested in multiple related outcomes simultaneously. Instead of describing a single uncertain quantity, like a coin flip, multivariate random variables allow us to model and analyze multiple interconnected random quantities together. This approach is crucial in real-world scenarios where outcomes are often correlated or jointly influenced by underlying factors. By considering multiple variables at once, we can capture more complex relationships and dependencies in our data, providing a richer and more accurate representation of the systems we are studying.

For my commuting example, my commute depends on how long the train takes as well as how fast I can walk between the stations. If the presidential motorcade blocks Pennsylvania Ave, then I would have to wait. We could model my commute to work as depending on a continuous train ride duration and a binary variable for the motorcade blocking traffic.

1.5.1 Joint Distribution

Both the PDF and CDF apply to when there are multiple random variables. We should just say a PDF is $f_{X,Y}(x,y) = \Pr(X = x, Y = y)$ and CDF is $F_{X,Y}(x,y) = \Pr(X \leq x, Y \leq y)$. Most of the points about PDFs and CDFs go through, just with a little more math.

In the next to subsubsections, we will introduce the Marginal and the Conditional distribution, which are somewhat opposites. Suppose we have a joint distribution $D(X,Y)$. The X Marginal Distribution is the distribution that describes the distribution of X without reference to Y (but this does *not* mean Y is ignored!). The Conditional Distribution of X given Y is the distribution of X that is contingent on the values of Y. While marginal distributions give us a broad view of a single variable across all possible values of other variables, conditional distributions provide a focused snapshot of one variable when another is fixed. An example of the two ideas is: a marginal distribution is like looking at average temperatures for a city across the whole year; a conditional distribution is like looking at temperatures specifically in July.

1.5.2 Marginal Distribution

The X Marginal Distribution attempts to collapse the information on Y so that we focus on how X behaves. An intuition of this is to imagine collapsing a 3D mountain range that you look at into a 2D picture. A mathematical intuition is that we wish to 'integrate' or 'average' out the variation Y when considering X. Later, we will apply this concept for the Law of Iterated Expectations.

Annoyingly, the notation for a marginal distribution and a univariate RV is the same. If the variation in Y provides no information about X (which might be true in my commuting example), then the X marginal distribution *is* the same as the univariate; otherwise, they are not the same. We will use the notation that is used widely, but we do not need to be happy about it.

Say we have a discrete MRV's PDF $f_{X,Y}(x,y)$, then the X marginal distribution is defined by:

$$F_X(x) = F_{X,Y}(x, \overline{y}) \tag{15}$$

$$f_X(x) = \sum_{y \in \mathcal{Y}} f_{X,Y}(x,y), \tag{16}$$

where we see that $F_X(x)$ and $f_X(x)$ are the same thing we would write if talking about a

univariate RV.[1]

1.5.3 Conditional Distribution

A conditional distribution describes the behavior of one random variable given that another variable has taken on a specific value. It is like slicing through a joint distribution at a particular point to answer 'what does X look like when we fix Y.' For the commuting example, we would ask 'what is the probability my commute will be less than 30 minutes conditional on the fact that there will *not* be a motorcade.' This concept is powerful because it allows us to explore how variables influence each other, revealing relationships and dependencies that might not be apparent when looking at the overall joint distribution or individual marginal distributions.

For a joint distribution $D(x, y)$, we define the probability of Y conditional X as:

$$\Pr(Y = y \mid X = x) = \frac{\Pr(Y = y \wedge X = x)}{\Pr(X = x)}, \tag{17}$$

where we require that $\Pr(X = x) > 0$. Essentially, we are normalizing the probability of the two events by the probability on the conditioning event (in this case, $X = x$).[2]

The density function for Y conditional on X is:

$$f_{Y|X}(y; x) = \frac{f_{X,Y}(x, y)}{f_X(x)}, \tag{18}$$

where $f_{X,Y}(x, y)$ is the joint PDF and $f_X(x)$ is the X marginal density. Also note, the arguments in the conditional density for $Y \mid X$ are reversed (i.e., the y is in front of x) and we use a semi-colon to designate that x is a fixed argument of the function (note: this is not notation used by everyone).

1.5.4 Example using Bivariate Normal distribution

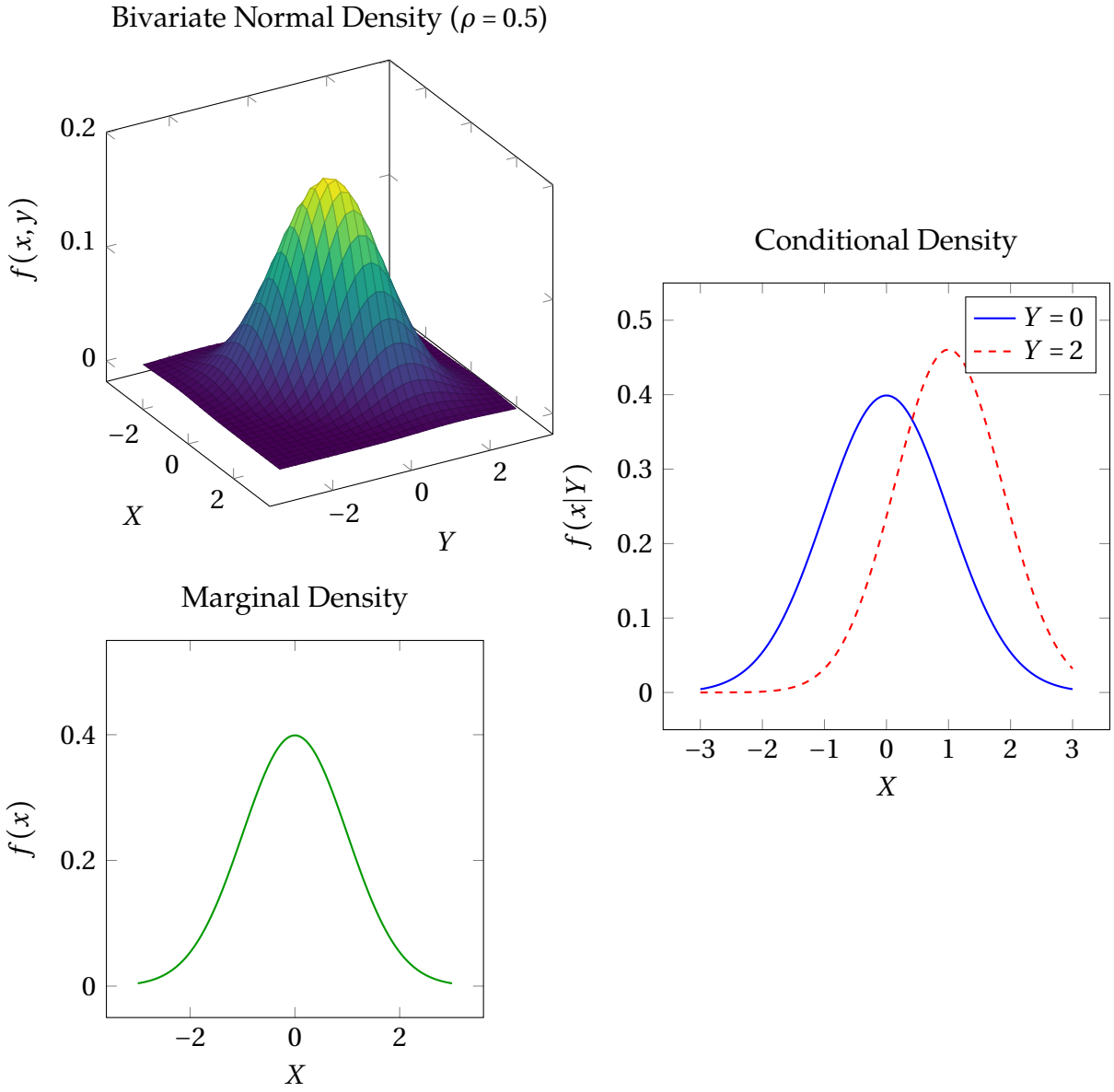
See Figure 5 for an example of a bivariate density (specifically a bivariate standard normal with correlation of 0.5) and a conditional density.

[1] Note: I assumed again that $D(x, y)$ is discrete.

[2] More generally, for any two events $\{E_1, E_2\}$, the probability of E_1 conditional on E_2 is $\Pr(E_1, E_2)/\Pr(E_2)$. In the above case, the events are $E_1 := Y = y$ and $E_2 := X = x$.

[htbp]

Figure 5: Bivariate, Marginal, and Conditional Densities



Note: This figure shows a bivariate density (top left), two conditional densities (right), and a marginal density (bottom left).

1.5.5 Independence and Dependence

Statistical independence and dependence describe the relationship between random variables. Two variables are considered statistically independent if the occurrence or value of one does not affect the probability of the other; i.e., knowing the outcome of one variable

provides no information about the other. On the other hand, statistical dependence implies that there is a relationship between the variables, where the value or occurrence of one variable influences the probability or distribution of the other.

In the commuting example, the duration of the train trip is almost certainly independent of whether the motorcade is going through. However, if longer train rides are correlated with more people traveling on the trains and the motorcade is active during morning traffic, then possibly they could be correlated.

For two variables with the joint distribution $D(x, y)$, we say X and Y are independent (denoted as $X \perp\!\!\!\perp Y$) if:

$$\Pr(X = x, Y = y) = \Pr(X = x) \cdot \Pr(Y = y), \tag{19}$$

$$F_{X,Y}(x, y) = F_X(x) \cdot F_Y(y), \text{ or} \tag{20}$$

$$f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y). \tag{21}$$

Independence is a useful property for simplifying things. For example, if $X \perp\!\!\!\perp Y$, then:

$$\Pr(Y = y \mid X = x) = \frac{\Pr(Y = y \wedge X = x)}{\Pr(X = x)} \text{ [by definition]} \tag{22}$$

$$= \frac{\Pr(Y = y) \cdot \Pr(X = x)}{\Pr(X = x)} \text{ [by independence]} \tag{23}$$

$$= \Pr(Y = y). \tag{24}$$

Another consequence is that in this case the conditional and marginal probabilities are the same.

2 Distributions to Know

In this section, we will go over some common univariate distributions that are useful to know. There are many distributions, so this section is only going to cover the ones that are essential / useful to know. There are three distributions that are essential to know: Bernoulli, Normal, and Uniform. One must intuitively understand these distributions. There are then three highly useful distributions to know: Log-Normal, Extreme Value, and Logistic. Anyone who seriously does economic research will use these distributions. Finally, there are three distributions that are useful to know: Poisson, Exponential, and Pareto. These distributions will pop-up in applications and it is good to be familiar.

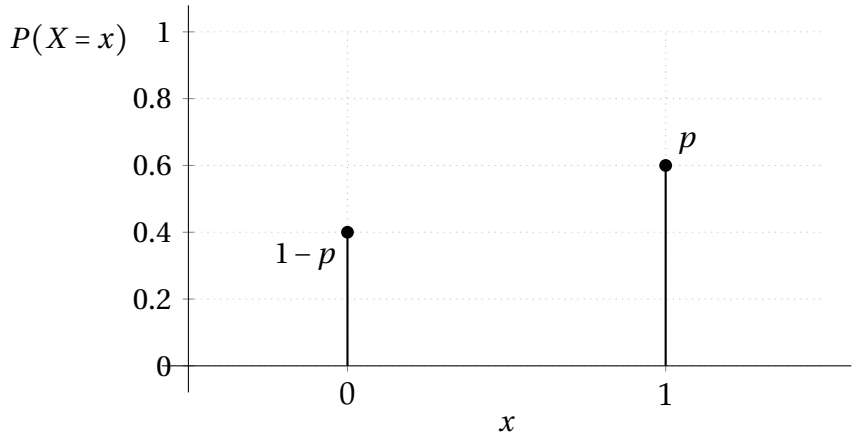
For these distributions (as well as others one may learn about later), one should know:

- Finite or Infinite Support?
- Discrete or Continuous?
- What is the shape of the PDF?
- What are the key parameters?
- What are the mean and variance?

2.1 Bernoulli – essential

The Bernoulli distribution is a discrete probability distribution for a random variable that has only two possible outcomes, typically labeled as success (1) or failure (0). It is named after Swiss mathematician Jacob Bernoulli and is a special case of the binomial distribution where only a single trial is conducted. This distribution is commonly used to model binary events such as coin flips, yes/no surveys, or the presence/absence of a particular characteristic in a population. The Bernoulli distribution is parameterized by a single parameter p, which represents the probability of success (1). If $X \sim \mathsf{Bernoilli}(p)$, then the mean is $\mathsf{E}[X] = p$ and $\mathsf{Var}(X) = p \cdot (1 - p)$.

Figure 6: Bernoulli Distribution PDF

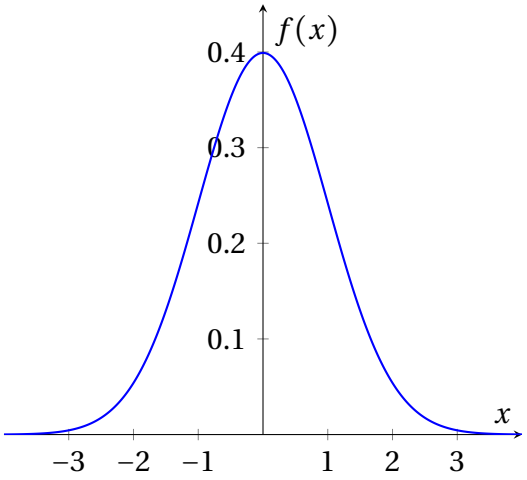


2.2 Normal – essential

The Normal distribution, also known as the Gaussian distribution, is a fundamental probability distribution in statistics and mathematics. It is characterized by its distinctive

bell-shaped curve, which is symmetric around the mean. The distribution is continuous and has infinite support. This distribution is ubiquitous in nature and plays a crucial role in many fields, including physics, biology, and finance. It serves as a good approximation for many real-world phenomena and is central to the concept of statistical inference. The Normal distribution is parameterized by a two parameters (μ, σ^2). If $X \sim \mathsf{N}(\mu, \sigma^2)$, then the mean is $\mathsf{E}[X] = \mu$ and $\mathsf{Var}(X) = \sigma^2$).

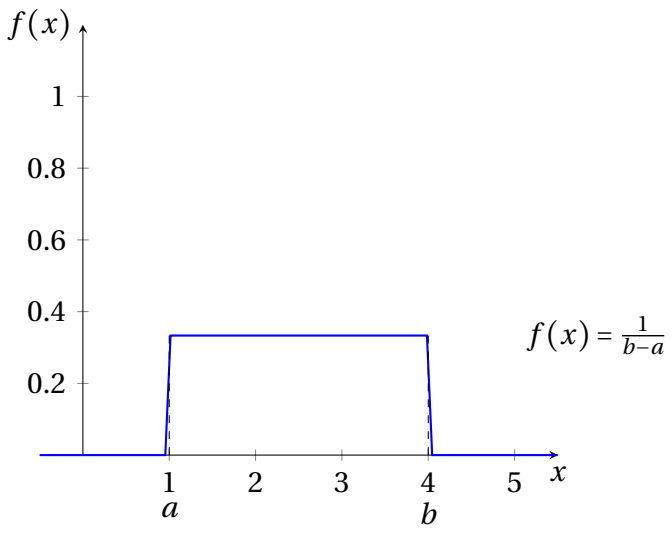
Figure 7: Normal Distribution PDF



2.3 Uniform – essential

The Uniform distribution is a simple yet important probability distribution in statistics. It describes a scenario where all outcomes within a specified range are equally likely to occur. This distribution is characterized by its constant probability density function over the defined interval, resulting in a rectangular shape when graphed. The Uniform distribution is often used in simulations, random number generation, and as a prior distribution in Bayesian inference when no prior information is available about a parameter except for its possible range of values. The Uniform distribution is parameterized by a two parameters (a, b), where a is the lower bound of the interval and b is the upper bound of the interval. If $X \sim \mathsf{U}(a, b)$, then the mean is $\mathsf{E}[X] = (a + b)/2$ and $\mathsf{Var}(X) = (b - a)^2/12$.

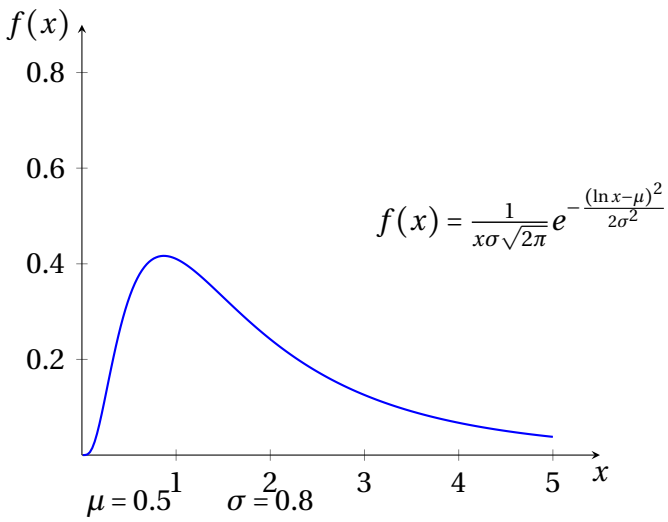
Figure 8: Uniform Distribution PDF



2.4 Log-Normal – highly useful

The log-normal distribution is a continuous probability distribution for a random variable whose logarithm follows a normal distribution. It is closely related to the normal distribution and arises naturally in many fields, particularly in economics, finance, and biology. This distribution is often used to model variables that are always positive and have a right-skewed shape, such as stock prices, income distributions, or the size of biological organisms. The log-normal distribution is parameterized by two parameters: μ and σ, which are the mean and standard deviation of the variable's natural logarithm, respectively. If $Y \sim \mathsf{LogNormal}(\mu, \sigma^2)$, then $X = \ln(Y) \sim \mathsf{Normal}(\mu, \sigma^2)$. The mean of Y is given by $\mathsf{E}[Y] = e^{\mu + \frac{\sigma^2}{2}}$ and its variance is $\mathsf{Var}(Y) = (e^{\sigma^2} - 1)e^{2\mu + \sigma^2}$.

Figure 9: Log-normal Distribution PDF



2.5 Logistic – highly useful

The Logistic distribution is a continuous probability distribution that plays a fundamental role in binary choice models and logistic regression analysis. When modeling binary outcomes, a key insight is that if the log-odds ratio is linear in covariates, then the conditional expectation function is the same as the conditional logistic distribution. This distribution is particularly valuable in modeling binary outcomes (e.g., growth curves, and survival analysis) as it captures the S-shaped relationship between predictor variables and probability outcomes, and provides the theoretical foundation for logistic regression in statistics and econometrics. The Logistic distribution is characterized by two parameters: μ (location parameter) and s (scale parameter), where $s > 0$. If $X \sim \text{Logistic}(\mu, s)$, then its probability density function is $f(x) = \frac{e^{-(x-\mu)/s}}{s(1+e^{-(x-\mu)/s})^2}$.[3] The mean is given by $\mathsf{E}[X] = \mu$ and variance by $\text{Var}(X) = \frac{\pi^2}{3}s^2$. The cumulative distribution function has the closed form $F(x) = \frac{1}{1+e^{-(x-\mu)/s}}$, which emerges naturally from the linear log-odds assumption.[4]

Example: Consider modeling the probability of a student passing an exam based on their study hours. If we let H be the hours studied and assume the log-odds of passing is

[3] The scale parameter s is related to the variance by $\text{Var}(X) = \frac{\pi^2}{3}s^2$, and the distribution is symmetric around μ.

[4] This function is sometimes called the Sigmoid function or the logit function.

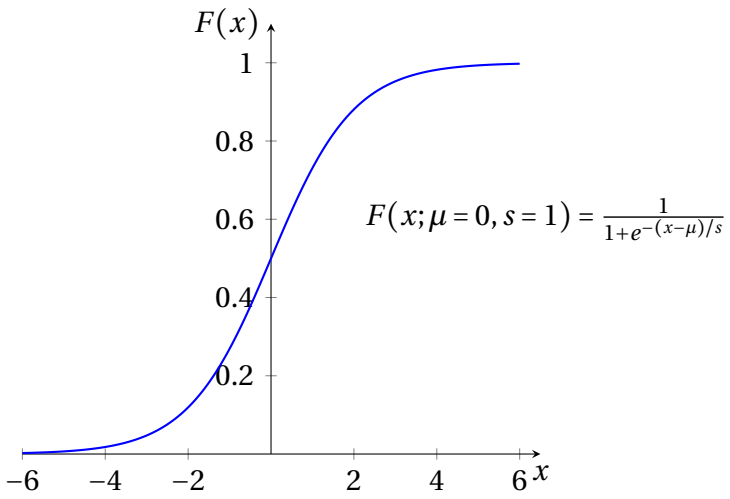
linear in H, we have:

$$\ln\left(\frac{\Pr(\text{pass}|H)}{1-\Pr(\text{pass}|H)}\right) = \beta_0 + \beta_1 H. \tag{25}$$

This linear log-odds assumption (i.e., Eq 25) implies that the conditional probability of passing must take the form of a logistic CDF with location parameter $\mu = -\beta_0/\beta_1$ and scale parameter $s = 1/\beta_1$:

$$\Pr(\text{pass}|H) = \frac{1}{1+e^{-(\beta_0+\beta_1 H)}}. \tag{26}$$

Figure 10: Logistic Distribution CDF (Sigmoid Function)



Digression: Typically in economics, we take a different approach to get to the logistic conditional expectation function. Using the example, suppose that we say the binary observed outcome 'passing given H hours of study,' denoted P, is determined by a latent stochastic variable $\text{P}^\star = \beta_0 + \beta_1 H + \epsilon$, where $\epsilon \sim \text{Logistic}(0,1)$, is above some threshold, κ. Without loss of generality, we can set $\kappa = 0$ and it is just absorbed into β_0. Then, we can model the conditional probability of passing as:

$$\begin{aligned}
\Pr(\text{P} == 1 \mid H) &= \Pr(\text{P}^\star > \kappa \mid H) && (27)\\
&= \Pr(\beta_0 + \beta_1 H + \epsilon > 0 \mid H) && (28)\\
&= \Pr(\epsilon > -(\beta_0 + \beta_1 H) \mid H) && (29)\\
&= \text{Logit}^{\neg 1}(\beta_0 + \beta_1 H) && (30)\\
&= \frac{1}{1+e^{-(\beta_0+\beta_1 H)}}. && (31)
\end{aligned}$$

2.6 Extreme Value – highly useful

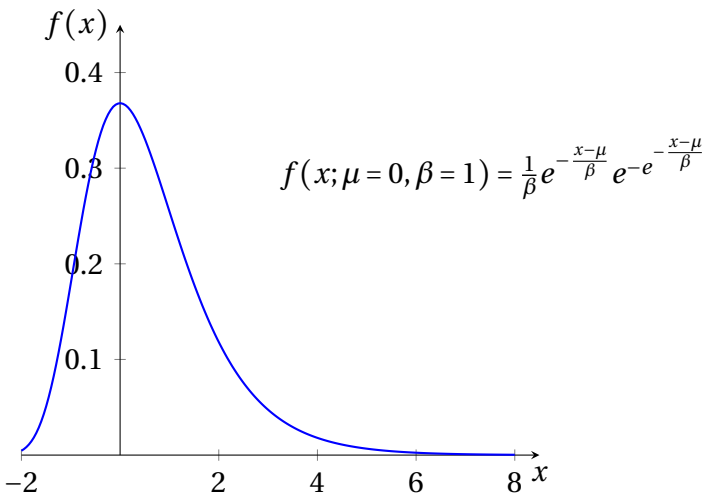
The Type 1 Extreme Value (T1EV) distribution (also known as the T1EV distribution) is a continuous probability distribution that plays a fundamental role in discrete choice theory and economic modeling. It naturally arises in random utility models where individuals choose among discrete alternatives, forming the theoretical foundation for logit models in econometrics. This distribution is particularly valuable in modeling consumer choice behavior, market share analysis, and transportation mode selection, as it captures the random component of utility in decision-making processes. The T1EV distribution is characterized by two parameters: μ (location parameter) and β (scale parameter), where $\beta > 0$. However, in discrete choice models, these parameters are typically normalized with $\mu = 0$ and $\beta = 1$ since only differences in utility matter for choice probabilities, and one parameter must be normalized for identification.[5] If $X \sim \mathsf{T1EV}(\mu = 0, \beta = 1)$, then its probability density function simplifies to $f(x) = e^{-x}e^{-e^{-x}}$. Generally, the mean is given by $\mathsf{E}[X] = \mu + \beta\gamma$ (where $\gamma \approx 0.5772$ is the Euler-Mascheroni constant) and variance by $\mathsf{Var}(X) = \frac{\pi^2}{6}\beta^2$. In multinomial logit models, the difference between two T1EV-distributed random variables follows a logistic distribution, leading to the familiar logistic choice probabilities.

Example: Consider a commuter choosing among three transportation modes: car (1), bus (2), and train (3). The utility of each mode can be written as $U_i = V_i + \epsilon_i$, where V_i is the deterministic component and ϵ_i follows a T1EV distribution. If $V_1 = 2$ (car), $V_2 = 1$ (bus), and $V_3 = 1.5$ (train), the probability of choosing each mode is given by:

$$P_i = \frac{e^{V_i}}{\sum_{j=1}^{3} e^{V_j}} = \begin{cases} P_1 = \frac{e^2}{e^2+e^1+e^{1.5}} \approx 0.48 \text{ (car)} \\ P_2 = \frac{e^1}{e^2+e^1+e^{1.5}} \approx 0.18 \text{ (bus)} \\ P_3 = \frac{e^{1.5}}{e^2+e^1+e^{1.5}} \approx 0.34 \text{ (train)} \end{cases} \tag{32}$$

[5] This normalization is necessary because choice probabilities depend only on utility differences. The scale parameter β is a non-trival normalization as it is inversely related to the variance of the error terms and thus determines the model's sensitivity to differences in deterministic utility.

Figure 11: Type 1 Extreme Value Distribution PDF



2.7 Poisson – useful

The Poisson distribution is a discrete probability distribution that models the number of events occurring within a fixed interval of time or space. The Poisson distribution naturally arises in processes where events occur continuously and independently at a constant arrival rate. It is particularly useful in modeling rare events such as customer arrivals, equipment failures, or accidents at a given location. This distribution is characterized by a single parameter λ (lambda), also known as the arrival rate and is always positive, that represents the average number of events in the given interval. The arrival rate is also equal to both the mean and variance of the distribution.[6] The probability mass function for a discrete random variable X following a Poisson distribution is given by $P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$ for $k = 0, 1, 2, \ldots$ If $X \sim \text{Poisson}(\lambda)$, then its expected value is $\mathsf{E}[X] = \lambda$ and variance is $\mathsf{Var}(X) = \lambda$. The Poisson distribution is often used as an approximation to the binomial distribution when n is large and p is small, with $\lambda = np$.

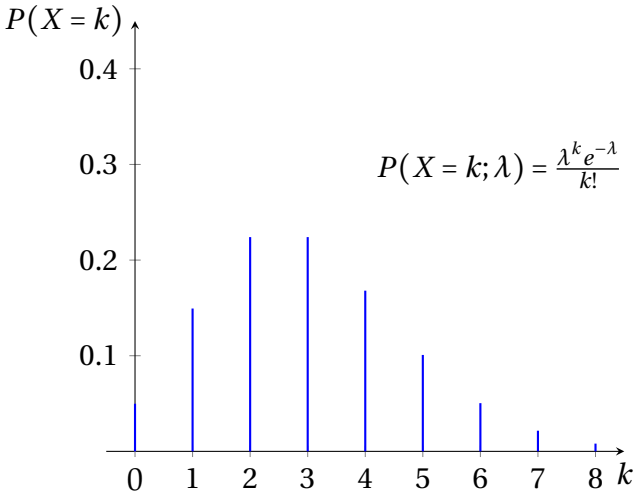
Example: Consider a fast-food restaurant that receives an average of $\lambda = 3$ customers per 10-minute interval during lunch hour. The probability of exactly k customers arriving in

[6] This equality of mean and variance is a distinctive property of the Poisson distribution, known as equidispersion.

any 10-minute period is given by:

$$P(X=k)=\frac{3^k e^{-3}}{k!}=\begin{cases} P(X=0)=e^{-3}\approx 0.050 \text{ (no customers)} \\ P(X=1)=3e^{-3}\approx 0.149 \text{ (one customer)} \\ P(X=2)=\frac{9e^{-3}}{2}\approx 0.224 \text{ (two customers)} \\ P(X=3)=\frac{27e^{-3}}{6}\approx 0.224 \text{ (three customers)} \\ \quad \ldots\text{and so on}\ldots \end{cases} \tag{33}$$

Figure 12: Poisson Distribution PMF



2.8 Exponential – useful

The Exponential distribution is a continuous probability distribution that describes the time between events in a Poisson point process, making it fundamental in reliability theory, queueing theory, and survival analysis. When modeling time-to-event data, a key feature of the exponential distribution is its "memoryless" property, meaning the probability of waiting an additional time t is independent of how long we have already waited. This distribution is particularly valuable in modeling waiting times, lifetimes of electronic components, and interarrival times in queuing systems, as it provides a simple yet powerful way to model processes where events occur continuously and independently at a constant average rate. The Exponential distribution is characterized by a single parameter: λ (rate parameter), where $\lambda > 0$. If $X \sim \text{Exponential}(\lambda)$, then its probability

density function is $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$. The mean is given by $\mathsf{E}[X] = \frac{1}{\lambda}$ and variance by $\mathsf{Var}(X) = \frac{1}{\lambda^2}$. The distribution has the unique property of constant hazard rate $h(x) = \lambda$. The cumulative distribution function has the closed form $F(x) = 1 - e^{-\lambda x}$ for $x \geq 0$.[7]

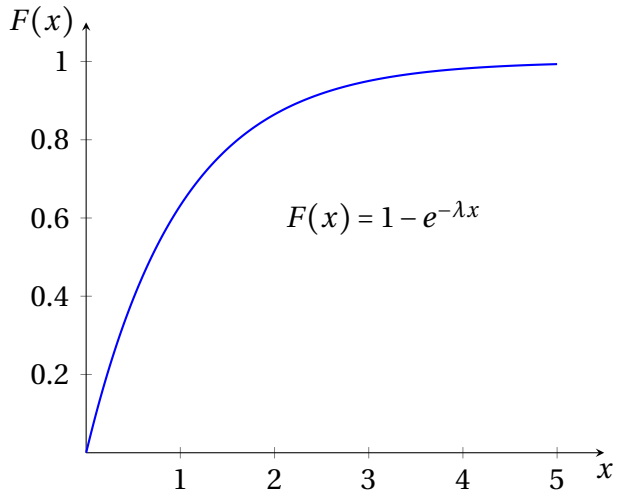
Example: Consider modeling the lifetime of electronic components in a manufacturing process. If we let T be the time until failure and assume a constant hazard rate λ, we have:

$$h(t) = \frac{f(t)}{1 - F(t)} = \lambda \tag{34}$$

This constant hazard rate assumption (i.e., Eq 34) implies that the probability of survival beyond time t must take the form of an exponential survival function:

$$\Pr(T > t) = e^{-\lambda t}. \tag{35}$$

Figure 13: Exponential Distribution CDF



2.9 Pareto – useful

Digression:

> *The 80-20 Rule:* also known as the Pareto principle, the 80-20 rule states that roughly 80% of effects come from 20% of causes in many real-world situations. For example:

[7] This function represents the probability that the waiting time is less than or equal to x.

- 80% of wealth is owned by 20% of the population
- 80% of sales come from 20% of customers
- 80% of complaints come from 20% of clients
- 80% of software bugs are caused by 20% of the code

The exact percentages need not be precisely 80% and 20%, but rather represent a general principle of imbalance where a small proportion of inputs tends to drive a large proportion of outputs.

Power Laws: A relationship between two quantities follows a power law if one quantity varies as a power of another. Mathematically, if y and x are related by a power law, then:

$$y = cx^{-\alpha} \tag{36}$$

where c and α are constants. Power laws have three distinctive properties:

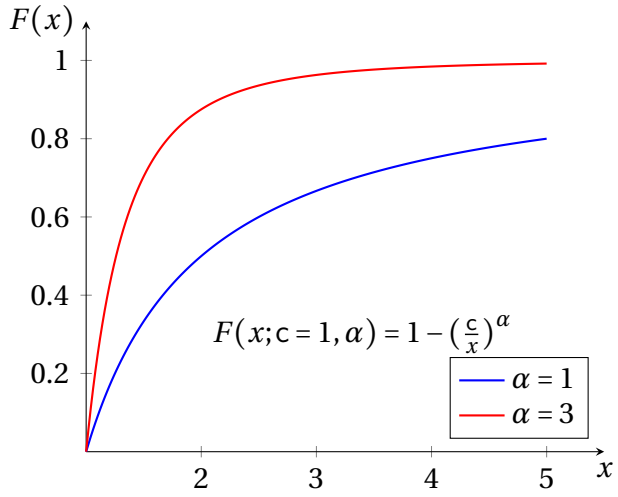
1. Scale invariance: Scaling the input x by a constant factor results in scaling the output y by a corresponding power of that factor
2. Linear relationship in log-log plots: Taking logarithms of both sides yields $\ln(y) = \ln(c) - \alpha \ln(x)$
3. Heavy tails: When $\alpha > 0$, power laws decay more slowly than exponential functions, leading to higher probabilities of extreme events

End Digression.

The Pareto distribution is a continuous probability distribution that exemplifies the "80-20 rule" and power law behavior found in many natural and social phenomena. When modeling phenomena where a small proportion of causes produce a large proportion of effects, the Pareto distribution emerges naturally. Its power law tail behavior makes it particularly suitable for modeling size distributions where extreme events are more common than would be predicted by light-tailed distributions. This distribution is invaluable in modeling wealth inequality, city populations, file sizes in computer networks, and natural disaster magnitudes. It was originally developed by Vilfredo Pareto to describe the distribution of wealth among individuals, where he observed that approximately 80% of the wealth was held by 20% of the population.

The Pareto distribution is characterized by two parameters: c (scale parameter or cutoff value or minimum value) and α (shape parameter or Pareto index), where both $\mathsf{c}, \alpha > 0$. If $X \sim \text{Pareto}(\mathsf{c}, \alpha)$, then its probability density function is $f(x) = \frac{\alpha \mathsf{c}^\alpha}{x^{\alpha+1}}$ for $x \geq \mathsf{c}$.[8] The mean is given by $\mathsf{E}[X] = \frac{\alpha \mathsf{c}}{\alpha - 1}$ for $\alpha > 1$ and variance by $\text{Var}(X) = \frac{\mathsf{c}^2 \alpha}{(\alpha-1)^2(\alpha-2)}$ for $\alpha > 2$.[9] The cumulative distribution function has the closed form $F(x) = 1 - (\frac{\mathsf{c}}{x})^\alpha$ for $x \geq \mathsf{c}$.[10]

Figure 14: Pareto Distribution CDF for Different Shape Parameters



Example 1: Consider modeling the distribution of annual incomes above a minimum threshold. If we let W be the annual income and set c as the minimum income threshold, the probability that an individual's income exceeds some value w follows a power law:

$$\Pr(W > w) = \left(\frac{\mathsf{c}}{w}\right)^\alpha \tag{37}$$

This power law relationship (i.e., Eq 37) implies that the log survival function is linear in log income:

$$\ln(\Pr(W > w)) = \alpha \ln(\mathsf{c}) - \alpha \ln(w). \tag{38}$$

Example 2: The distribution of city sizes provides a classic example of how the Pareto distribution manifests as Zipf's law when data is rank-ordered. While the Pareto distri-

[8] The shape parameter α determines the thickness of the tail, with smaller values corresponding to heavier tails. The kth moment exists only when $k < \alpha$.

[9] Note that the mean exists only for $\alpha > 1$ and the variance only for $\alpha > 2$, highlighting the heavy-tailed nature of the distribution.

[10] The survival function $S(x) = \left(\frac{\mathsf{c}}{x}\right)^\alpha$ follows a power law, a defining characteristic of the distribution.

bution describes the probability of observing a city of size X or larger, Zipf's law emerges when we rank these cities by population. Let X be the population of a city, and assume it follows a Pareto distribution with parameters c (minimum city size) and α (Pareto shape parameter). For a country with N cities, the expected rank r of a city with population x is:

$$r = N \Pr(X > x) = N \left(\frac{\mathsf{c}}{x} \right)^{\alpha} \tag{39}$$

Solving for x gives:

$$x = \mathsf{c} \left(\frac{N}{r} \right)^{1/\alpha} \tag{40}$$

When $\alpha \approx 1$, this becomes Zipf's law:

$$x \approx \frac{\mathsf{c} \cdot N}{r}. \tag{41}$$

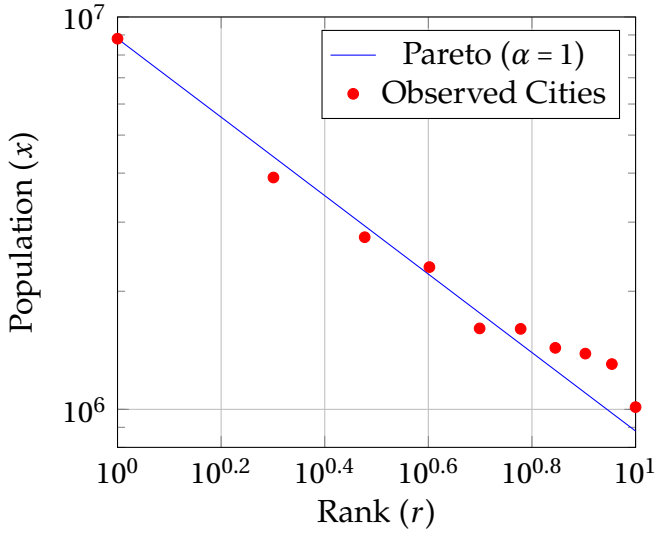
Apply this to US City Populations: Consider the populations of major US cities based on the 2020 census data. If we fit a Pareto distribution to these populations, we expect:

- The survival function to follow a power law
- The rank-size relationship to approximate Zipf's law
- $\alpha \approx 1$ in the Pareto distribution

Table 1: Top 10 US Cities: Pareto-Zipf Analysis (2020)

Rank (r)	City	Population (x)	$\ln(r)$	$\ln(x)$	$\ln(N/r)$
1	New York	8,804,190	0.000	15.991	7.090
2	Los Angeles	3,898,747	0.693	15.176	6.397
3	Chicago	2,746,388	1.099	14.825	5.991
4	Houston	2,304,580	1.386	14.651	5.704
5	Phoenix	1,608,139	1.609	14.290	5.481
6	Philadelphia	1,603,797	1.792	14.288	5.298
7	San Antonio	1,434,625	1.946	14.176	5.144
8	San Diego	1,386,932	2.079	14.142	5.011
9	Dallas	1,304,379	2.197	14.081	4.893
10	San Jose	1,013,240	2.303	13.829	4.787

Figure 15: Pareto-Zipf Analysis of US Cities



2.10 Bonus: Uniform Value Trick to Drawing RVs

Basic Principle Suppose you want to draw random variable X with CDF $F_X(x)$. Let $U \sim \text{Uniform}(0,1)$. Then: $X = F_X^{-1}(U)$ has the same distribution as X.[11] This method works because $P(F_X^{-1}(U) \leq x) = P(U \leq F_X(x)) = F_X(x)$.

Normal Distribution For the standard normal distribution, we use: $X = \Phi^{-1}(U)$, where Φ^{-1} is the inverse of the standard normal CDF. For a general normal distribution with mean μ and standard deviation σ: $X = \mu + \sigma\Phi^{-1}(U)$

Pareto Distribution For a Pareto distribution with scale parameter c and shape parameter α, the CDF is $F_X(x) = 1 - (\frac{c}{x})^\alpha$ and so the inverse CDF is: $F_X^{-1}(u) = \frac{c}{(1-u)^{1/\alpha}}$. Therefore, to generate Pareto random variables: $X = \frac{c}{(1-U)^{1/\alpha}}$.

Algorithm To generate random samples:

1. Generate $U \sim \text{Uniform}(0,1)$
2. Calculate $X = F_X^{-1}(U)$

[11] The Inverse CDF is called the quantile function.

Part II

Fundamentals of Statistics

> Statistics (from German: *Statistik*, orig. "description of a state, a country") is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of data.
>
> — Wikipedia: Statistics

3 Basic Concepts

Broadly speaking, statistics is concerned with using tools of probability and observed data to 'understand the world.' Another way of think about statistics is that statistics is methodology describing how to use data to undercover the probabilities governing the world.

In this section, we discuss basic concepts of statistics. Second, we discuss about ways of describing data. Third, we discuss properties of estimators. Fourth, we discuss other statistical concepts that use the ideas we will have introduced. Finally, we touch upon statistical inference.

Econometrics: The difference between statistics and econometrics is that econometrics incorporates standard assumptions from economics. While statistics provides the mathematical framework and tools for analyzing relationships in data, econometrics specifically builds upon economic theory and its standard assumptions.

For example, when analyzing how price affects quantity, econometricians frame their analysis within established economic frameworks. The econometrician will realize the answer depends on which quantity one is interested in: quantity supplied or quantity demanded. Supply and price are positively related while demand and price are negatively related, so to properly measure how equilibrium quantity is affected by price, one must think about the underlying economic model of demand and supply.

3.1 Population vs Sample

3.1.1 Population

A statistical population refers to the entire group of individuals, items, or events that are the subject of a study or analysis. A population can be finite or infinite. In the context of economic research, this could be all households in a country, all firms in an industry, or all transactions in a market. The population encompasses every possible observation that fits the defined criteria of interest.

3.1.2 Sample

A statistical sample is a subset of a statistical population. For most economic applications, populations are large, so it is impractical or impossible to study every member of a population. This is why researchers typically work with samples, which are subsets of the population, to make inferences about the population as a whole. Understanding the nature and characteristics of the population is crucial for designing studies, selecting appropriate sampling methods, and drawing valid conclusions from statistical analyses.

3.1.3 Random Sampling

Random sampling is an approach to collecting a sample randomly as opposed to selecting it based on some deterministic rule. A strict mathematical definition is that any sample of size n, $\mathbf{Y}_n$, has equal probability of being selected from the population.

In practice, there are several ways of generating a random sample, and detailing them is far beyond the scope of this note. We usually refer to a random sample implicitly in terms of its ideal version and as the opposite of a *biased* sample. Here is an intuitive idea of the difference between a random sample and a biased sample:

- Random:
    1. Drawing ten jellybeans from a jar without looking
    2. Drawing ten student names from a hat
- Biased:
    1. Picking out ten blue jellybeans
    2. Picking out ten students based on who raises their hand the most

3.1.4 Independent and Identically Distributed

If n observations are drawn from the same distribution and are independent, then we say that they are independent and identically distributed – shortened to iid. Baby Wooldridge provides the following definition that is implicitly for an iid random sample.

Definition 3.1. If a sample $\mathbf{Y}_n = \{Y_1, Y_2, \dots Y_n\}$ is independent random variables with a common probability density function $f(y;\theta)$, then $\mathbf{Y}_n$ is a random sample from the population represented by $f(y;\theta)$.

Almost all econometric theory assume we have an iid random sample. In practice, we as researchers need to ensure that we are actually using data that approximates an iid random sample.

3.2 Estimators vs Estimates

The following is a set of circular definitions. An estimator is a function that takes a sample as its argument to produce an estimate, and estimates are the output value of estimators given a sample.

Suppose we want to know the mean age of all Americans (a), and that we have a random sample of ages of Americans, $\mathbf{a}_N = \{a_i\}_{i \in N}$. The sample average is the estimator of the mean, and the estimate is the sample average of the sample that we have. The sample average is a generic function: $\text{Avg} : \mathbf{S}_n \to \mathbb{R}$. This function's argument is a sample of a given size, and its output is a real number. This is just like saying $f(x) = x^2$, except the 'x' is a sample rather than a single number. The estimate is when we put a specific sample into the function. Again, this is similar to having $f(x = 2) = (2)^2 = 4$.

Estimators can be quite complex or simple. We could say that our estimator for the mean age of Americans will be the first age that we observe, the median age, or the average after removing the ten greatest and ten least values that we observe. We could use all of these methods as an estimator to estimate the mean. Statistics is the study of which of all the possible estimators is 'the best' given some definition of what 'the best' means.

Econometrics: The mathematics of regressions tell how to calculate the estimator for regression coefficients. The estimator is $B_{\text{OLS}}(X, y) = \hat{\beta} = (X'X)^{-1}X'y$, and we get estimates when we supply data on covariates, X, and a dependent variable, y, to the function.

3.2.1 Models vs Estimators

One of Wooldridge's biggest pet peeves is when people confuse estimators and models. Here's how I would describe the difference: the model is the (economic) relationship you think exists that relates one or more variables to each other via some parameters; the estimator is what ever statistical method you use to estimate the parameters of that relationship. Models describe the population; estimators are tools to learn about the population relationship given data.

For example, suppose you say quantity demanded is a function of price: $D(p) = \alpha \cdot p^{\varepsilon} \cdot u$. This is a model. You could estimate this model several different ways. One way is to estimate a linear regression of the logged values of the variables because the model then implies a linear relationship: $\ln[D] = a + \varepsilon \ln[p] + \ln[u]$. Thus, the OLS coefficient for log price is the OLS estimator of ε, and given data we have the specific OLS estimate.[12]

3.3 Estimators as Functions of Random Variables

The next concept to internalize is that, since our data is random from a population, our sample is just a complicated random variable. Before we draw our data, we could draw any sample of data as governed by the underlying distribution of the data and the random sampling procedure. Once we actually draw it, then we have an observed sample – just like how before we flip a coin, there is a probability distribution over which side will land up.

Further, if the sample is a complicated random variable, then that means an estimator is a function of a random variable, which we can also call a random variable (or random function). This means that an estimator is governed by the probability distribution of a sample may be drawn from and the aspects of the function. For example, suppose age is distributed according to a normal distribution: $a_i \sim \mathcal{N}(\mu, \sigma)$. Now, for whatever reason, suppose we have a function $Y(a) = a^2$. Since a is a random variable, then $Y(a)$ is now a random function, and we can characterize how this function behaves by knowing how the square of a normally distributed random behaves.[13] Note, as I said we can consider the random function a random variable, we can just name this random variable $y = Y(a)$.

Understanding that estimators have probability distributions is crucial for several reasons. First, it tells us how precise our estimates are likely to be – some estimators might

[12] In this example, with additional assumptions about u, we could estimate a using the OLS estimator for the regression intercept.

[13] In fact, we do know how this behaves: this follows (roughly) a Chi-squared (χ^2) distribution.

consistently give values close to the true parameter, while others might be more variable. Second, it allows us to construct confidence intervals, telling us things like "we're 95% confident the true population mean lies between $(\underline{x}, \overline{x})$." Third, it enables hypothesis testing, where we can formally assess claims like 'the true parameter is zero' or 'treatment A is better than treatment B.' Without knowing the probability distribution of our estimator, we couldn't make any of these statistical inferences about the population from our sample.

Econometrics: Since the regression coefficient formula shows that regressions are functions of data, then the OLS estimator is a random function, and thus the coefficients are random variables that are subject to the probability distribution of the data and the rules of the function. This is where the statistical inference on regression coefficients comes from. The regression coefficients have a probability distribution, and we can use the properties of that distribution (along with specific hypothesis tests we want to do) to tell use where a given estimate is for that distribution.

4 Distributional Characteristics

4.1 Moments

In mathematics, the concept of **moments** are used to describe the features and shape of a function. It turns out that moments are very useful in describing random variables' properties. We do not need to learn too much about moments in general; we will only focus on the specific moments that are used frequently in statistics and estimation.

To calculate the n^{th}-moment:

$$\mathsf{m}(n) = \begin{cases} \sum_i \left(x_i^n \cdot f(x_i) \right), & \text{discrete distribution} \\ \int \left(x^n \cdot f(x) \right) \mathrm{d}x, & \text{continuous distribution} \end{cases} \tag{42}$$

If $F(x)$ is a cumulative density function, then the n^{th}-moment of a random variable X is the **expectation** of X raised to the n:

$$\mathsf{m}(n) = \mathsf{E}[X^n]. \tag{43}$$

We can also calculate *central* moments:

$$\mathsf{m}^{\mathsf{c}}(n) = \begin{cases} \sum_i \left(\left(x_i^n - \mathsf{m}(1) \right) \cdot f(x_i) \right), & \text{discrete distribution} \\ \int \left(\left(x_i^n - \mathsf{m}(1) \right) \cdot f(x) \right) \mathsf{d}x, & \text{continuous distribution} \end{cases} \tag{44}$$

where we calculate the central moment about the first 'raw' moment.[14]

Proposition 1. *If all moments from two distributions are always equal, then the two distributions are actually the same.*

Note: the important thing is not the above definitions; rather, it is to understand the moments we describe below come from a set of mathematical tools about describing distributional characteristics of random variables.

4.2 First Moment: Mean

The first moment of a distribution is the **mean**, also called the expectation:

$$\mu_X = \mathsf{E}[X]. \tag{45}$$

We often denote the mean with the Greek letter μ.

The mean is a measure of central tendency for a distribution, representing the 'center' of the distribution. While the mean summarizes the distribution, the mean does not necessarily need to be a value in the domain of the distribution. For example, consider a Bernoulli with probability $p \in (0,1)$; the mean *is* p, but a Bernoulli random variable only takes values $\{0,1\}$.

4.2.1 Sample Arithmetic Average

We use the sample arithmetic average as the estimate of the mean. We calculate this as:

$$\hat{\mu}_{X,n} = \frac{1}{n} \sum_{i \in n} x_i. \tag{46}$$

The mean considers all possible values of the variable weighted by their probability of being observed. The sample average considers only the values that we happened to draw.

[14] Thus, $\mathsf{m}^{\mathsf{c}}(n) = \mathsf{E}\left[(X - \mathsf{m}(1))^n \right]$.

We will cover properties of estimators in the next section, but let's think about why the sample average is a good estimate of the mean.

Sample Average Estimator of the Mean: Let's consider on discrete random variables that take on finitely many values, denoted as G.[15] We can label the distinct values $\{x^1, x^2, \ldots, x^G\}$. We observe a n observation size sample $\mathbf{X}_n = \{x_i\}_{i \in n}$. Given the above assumptions, we can group our sampled observations according to which distinct value they have:

$$\mathbf{X}_n = \{\{x_i^1\}_{i \in n_1}, \{x_i^2\}_{i \in n_2}, \ldots, \{x_i^G\}_{i \in n_G}\}. \tag{47}$$

Let's now consider a binary random variable that takes the value of 1 if a value in the sample equals a particular distinct value, x^g, from the distribution; else is zero: $\iota_i(x^g) = 1[x_i = x^g]$. With this new random variable, we can rewrite each observation in the sample as: $x_i^{g(i)} = \sum_{f \in F} x^g \cdot \iota_i(x^g)$, where $g(i)$ is the value for the given observation.

Finally, let's write out the sample average using these assumptions:

$$\begin{aligned}
\hat{\mu}_{X,n} &= \frac{1}{n} \sum_{i \in n} x_i && (48)\\
&= \frac{1}{n} \sum_{i \in n} \left(\sum_{g \in G} x^g \cdot \iota_i(x^g) \right) && (49)\\
&= \sum_{g \in G} \left(x^g \cdot \frac{1}{n} \sum_{i \in n} (\iota_i(x^g)) \right) && (50)\\
&= \sum_{g \in G} \left(x^g \cdot \frac{n_g}{n} \right) && (51)\\
&\approx \sum_{g \in G} (x^g \cdot \Pr(x = x^g)) = \mu_X && (52)
\end{aligned}$$

If $\frac{n_g}{n} \approx \Pr(x = x^g)$, then the sample average approximates the mean. Therefore, what we need for the sample average to approximate the mean is that the observations are random draws from the distribution of interest. If the sample is biased relative to the distribution of interest, then the sample average will not yield the mean of interest. It is the connection between the likelihood that we see an observation in our sample and the true probability that an observation is observed in the population that a random sample provides that assures us that the sample average is a good estimator for the mean.

[15] For example, flipping a coin has $G = 2$ or rolling a die has $G = 6$.

4.3 Second Central Moment: Variance

The second central moment of a distribution is the **variance**:

$$\sigma_X^2 = \mathsf{Var}(X) = \mathsf{E}\left[(X - \mu_X)^2\right]. \tag{53}$$

We often denote the variance with the Greek letter σ^2.

Variance measures the dispersion of the distribution about its mean. Greater variance implies greater dispersion. A variance of zero would mean all observations are equal to the mean. This fact implies the following property: if a is a constant, then $\mathsf{Var}(a) = 0$.

Note the following:

$$\mathsf{E}\left[(X - \mu_X)^2\right] = \mathsf{E}[X^2] - \mathsf{E}[X]^2, \tag{54}$$

where we could have written the second term as $(\mu_X)^2$. This formulation helps to clarify the following property: $\mathsf{Var}(X) \geq 0$.

We can also work out the following properties:

- $\mathsf{Var}(X + a) = \mathsf{Var}(X)$
- $\mathsf{Var}(a \cdot X) = a^2 \cdot \mathsf{Var}(X)$
- $\mathsf{Var}(a \cdot X + b \cdot Y) = a^2 \cdot \mathsf{Var}(X) + b^2 \cdot \mathsf{Var}(Y) - 2 \cdot a \cdot b \cdot \mathsf{Cov}(X, Y)$,

where $\mathsf{Cov}(X, Y)$ is the *co*variance of X and Y that will be covered later.

4.3.1 Standard Deviation

We often also talk about the standard deviation: $\mathsf{sd}(X) = \sqrt{\mathsf{Var}(X)}$. We often denote the standard deviation with the Greek letter σ, as in the square root of σ^2 that denotes the variance.

The standard deviation also measures dispersion (since the square-root is a monotonic function); however, the standard deviation is in the same units as the variable itself. The variance is usually hard to compare to the mean or median. Thus, the phrase "a one standard deviation increase from the mean" which is relatively easy to interpret.

4.4 Higher Order Moment

Higher order moments are not often used except that sometimes one can use their properties for a proof. Worth mentioning though are the next two moments:

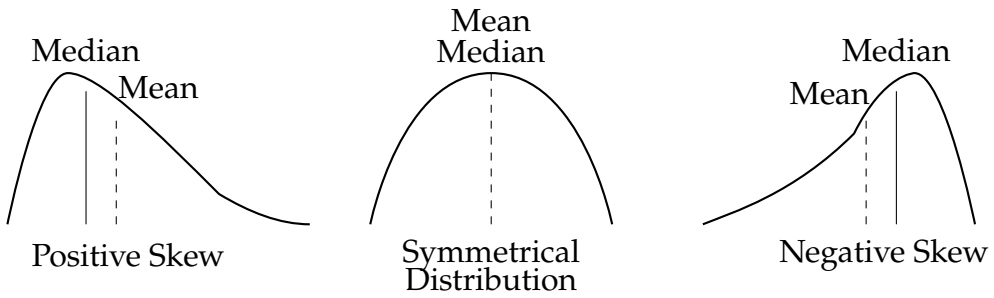
4.4.1 Third Central Moment: Skewness

Skewness measures the amount of asymmetry in a distribution about its mean.

- If the mean is greater than the median, then positive skew.
- If mean is less than the median, then negative skew.
- If the mean and median are the same, then no skew.

Figure 16 shows this visually.

Figure 16: Skewness



4.4.2 Fourth Central Moment: Kurtosis

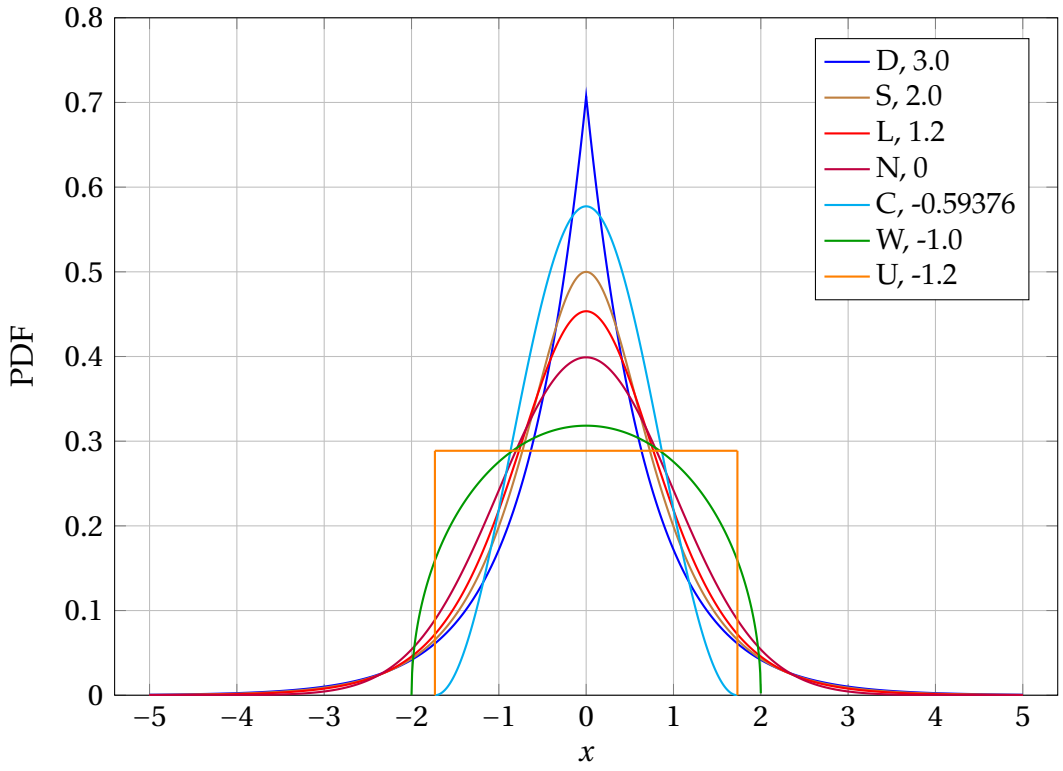
Kurtosis is a measure of the likelihood that a random draw from a distribution will be an 'outlier.' One could consider this a measure of the 'fatness' of the distribution tails.

Kurtosis is often interpreted as *excess kurtosis*: kurtosis minus three. The normal distribution has an excess kurtosis of zero, and this is used as a reference point. If the excess kurtosis is positive, then the likelihood of an outlier is greater than the normal distribution; if negative, then it is less likelihood than the normal distribution. Wikipedia has a cool figure that compares various distributions:

- D: Laplace distribution, also known as the double exponential distribution, red curve (two straight
- lines in the log-scale plot), excess kurtosis = 3
- S: hyperbolic secant distribution, orange curve, excess kurtosis = 2
- L: logistic distribution, green curve, excess kurtosis = 1.2

- N: normal distribution, black curve (inverted parabola in the log-scale plot), excess kurtosis = 0
- C: raised cosine distribution, cyan curve, excess kurtosis = 0.593762…
- W: Wigner semicircle distribution, blue curve, excess kurtosis = 1
- U: uniform distribution, magenta curve (shown for clarity as a rectangle in both images), excess kurtosis = 1.2

Figure 17: Kurtosis



4.5 Relationship Measures

4.5.1 Covariance

Covariance measures the joint variability of two random variables, defined as:

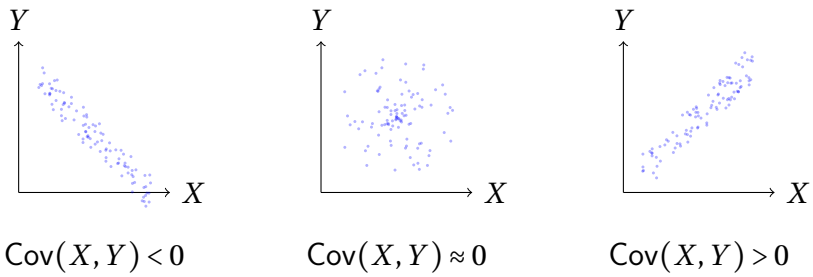
$$\mathrm{Cov}(X,Y) = \mathrm{E}[(X-\mu_X)\cdot(Y-\mu_Y)] \tag{55}$$

$$= \mathrm{E}[X\cdot Y] - \mathrm{E}[X]\cdot\mathrm{E}[Y] \tag{56}$$

Covariance is not usually denoted with a Greek letter, but sometimes one sees $\sigma^2_{X,Y}$.

Figure 18 shows the how covariance might show itself if there was linear dependence.

Figure 18: Covariance



4.5.2 Correlation

Covariance, like variance, has the issue that it is in different units than the underlying variables, which can make it difficult to interpret. The Correlation between to variables provides a more interpretable parameter.

$$\mathsf{Corr}(X,Y) = \frac{\mathsf{Cov}(X,Y)}{\mathsf{sd}(X)\cdot\mathsf{sd}(Y)} \tag{57}$$

$$= \mathsf{E}\left[\left(\frac{(X-\mu_X)^2}{\sigma_X}\right)\cdot\left(\frac{(Y-\mu_Y)^2}{\sigma_Y}\right)\right]. \tag{58}$$

We often denote the correlation parameter with the Greek letter ρ.

The correlation has the nice property that it is bounded between $[-1,1]$, where $\rho = -1$ is perfectly negative correlation and $\rho = 1$ is perfectly positive correlation.

4.6 Other Measures of Central Tendency

4.6.1 Median

The median is a measure of central tendency that aims to find the value from the population that is 'in the middle,' such that half of the population is below and half is above. To emphasize, the median is about all elements of the population.

The median of a population given a CDF is a value m that satisfies:

$$\lim_{x\to m^-} F(x) \le \frac{1}{2} \le F(m). \tag{59}$$

An interesting property is that the distance between the median (m) and the mean (μ) is bounded by the standard deviation (σ):

$$|\mu - m| = |\mathsf{E}[X - m]| \leq \mathsf{E}[|X - m|] \leq \mathsf{E}[|X - \mu|] \leq \sqrt{\mathsf{E}[(X-\mu)^2]} = \sigma. \tag{60}$$

The first and last equal signs are definitional. The first and third inequalities are based on Jensen's Inequality.[16] The second inequality is based on the property that the median minimizes absolute deviations, so the deviation from the mean must be greater.

4.6.2 Mode

The mode is the value that appears the most often in a population. For a discrete random variable, the mode is calculated as:

$$\text{mode} = \arg\max_{x \in \mathcal{X}} \{\Pr(x)\}; \tag{61}$$

however, it is too tedious to bother with writing the continuous distribution version.

Most distributions you will encounter are unimodal (i.e., one mode), but there are multimodal distributions. For empirical work, it is useful to plot your data to get a sense of the modality of the data.

5 Properties of Estimators

To understand the properties of estimators, we need to consider the case where there is some parameter we are interested in and we have some estimator for it.

For notation purposes, let θ be the true parameter and let $\hat{\theta}_X$ be our estimator for it given a population. As alluded to above, the distribution of an estimator is a function of the underlying population that is used to construct it (and/or the sample from the population). We use this to inform these properties.

[16] Calude.ai: "Jensen's inequality is a fundamental mathematical principle that tells us something deep about averages and curved lines. For a convex function, the average value of the function evaluated at different points will always be greater than or equal to the function evaluated at the average of those points."

5.1 Bias

Bias is the difference between the true parameter and the expectation of the estimator:

$$\text{Bias}(\hat{\theta}_X) = \text{E}[\hat{\theta}_X] - \theta. \tag{62}$$

Note, the expectation is calculated based on the underlying random variable X and the estimator's functional form.

If $\text{Bias}(\hat{\theta}_X) = 0$, then the estimator is *unbiased*. Bias is a finite sample property of an estimator. If an estimator is biased, then no matter the sample size it will always be biased. Note, some estimators may not have an expectation (i.e., an undefined expectation), and so cannot have a bias measure – in this case we would use an alternative measure.[17]

A fun exercise is showing that the simple sample variance is biased, but that there is a simple correction to calculate an unbiased sample variance estimator. I leave this up to you to work out.

5.1.1 Mean Squared Error

The mean squared error can be decomposed into two terms, bias (squared) and variance:

$$\text{E}[(\hat{\theta}_X - \theta)^2] = \text{Bias}(\hat{\theta}_X)^2 + \text{Var}(\hat{\theta}_X). \tag{63}$$

This measure is good for displaying that any estimate is trading off bias verses precision. For example, an unbiased estimator with a large variance will likely not be very precise; however, the converse of bias but low variance may be preferable if the bias is not too large.

5.2 Consistency

5.2.1 Convergence in Probability

Consistency uses the concept of 'convergence in probability.' A sample converges in probability to a population random variable if the probability that the elements of the

[17] The most common example of such is the just identified instrumental variable estimator (i.e., 2SLS with one instrument and one endogenous variable) has no expectation.

sample sequence are larger than ϵ goes to zero as the sample size goes to infinity:

$$\lim_{n\to\infty} \Pr(|X_n - X| > \epsilon) = 0. \tag{64}$$

We denote a sample converges in probability using the *probability limit* operator plim, and we denote it as:

$$X_n \xrightarrow{p} X \qquad \Longleftrightarrow \qquad \text{plim}_{n\to\infty} X_n = X. \tag{65}$$

The most important property for basic statistics knowledge is that:

$$\text{For continuous function } g(\cdot), X_n \xrightarrow{p} X \Longrightarrow g(X_n) \xrightarrow{p} g(X); \tag{66}$$

this is an application of the *continuous mapping theorem*. Since estimators are functions of samples, this property says that as long as we have a sample that converges in probability to the population, then the function of that sample will also converge in probability to the population version.

5.2.2 Consistency of Estimator

Consistency of an estimator is that the estimator converges in probability to the population parameter: $\hat{\theta}_X \xrightarrow{p} \theta$. Consistency is an asymptotic property in that it relies on increasing the sample size. It is possible to have biased-but-consistent estimators and unbiased-but-inconsistent estimators. All else equal, consistency is more important for empirical work than unbiasedness (although, both are best).

Economics and modern empirical work mostly use large samples and as such we often rely on asymptotic justifications for our estimators. This is good in that as a rule asymptotic properties are easier to satisfy and do not require parametric assumptions. One thing you will notice is that it will make it harder to communicate with other academic disciplines, which tend to rely more on parametric assumptions. One example of this is that colleagues outside of economics may talk about testing whether the sample follows a normal distribution... in economics, we assume that we have a sample size large enough that we rely on asymptotics to kick in.

However, a reasonably fruitful line of research is finding cases where it turns out that the sample size needs to be **huge** (rather than just large), in which case the sample sizes we use in empirical work may not be big enough. If you find that this is true of a commonly used empirical technique, then you will almost certainly get a good academic journal

publication out of it.

5.3 Efficiency

Efficiency is a property that we ask for after we have unbiasedness or consistency (at least to a degree we are comfortable with). That is, we think about efficiency after we have ruled out estimators that unsuitable. Essentially, if $\text{Var}(\hat{\theta}_X) < \text{Var}(\tilde{\theta}_X)$, then $\hat{\theta}_X$ is more efficient than $\tilde{\theta}_X$, and $\hat{\theta}_X$ is the efficient estimator if $\text{Var}(\hat{\theta}_X) < \text{Var}(\tilde{\theta}_X)$ for *all* other potential estimators.

Efficiency allows one to compare different estimators. There could be many different estimators for a given population parameter, but the one we should use will be unbiased / consistent and as efficient as possible.

5.4 Standard Error

A standard error is the sampling standard deviation of an estimator. The standard error measures the following thought experiment: suppose we randomly sampled n values from the distribution $F(x)$ k times so we have $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_k\}$ samples, we calculate our estimator for each sample to get $\{\hat{\theta}_{\mathbf{X}_1}, \hat{\theta}_{\mathbf{X}_2}, \ldots, \hat{\theta}_{\mathbf{X}_k}\}$, and then calculate the standard deviation of these k different estimates to see what is the underlying variability of the estimator.

Typically in empirical projects we only ever get one sample, so how do we measure the underlying variability? We use the properties of $x_i \sim F(x)$. For many estimators that we use, we can use the underlying distribution of $F(x)$ and the functional form of the estimator *or* we can rely on asymptotics and the functional form. For example, if σ is the standard deviation of a random variable X, then the standard error of the sample average is $\sigma/\sqrt{n}$. Similarly, if $\hat{\sigma}$ is the *sample* standard deviation of a sample $\mathbf{X}$, then the *estimator* of the standard error of the sample average is $\hat{\sigma}/\sqrt{n}$.

5.4.1 Confidence Intervals

So we have a standard error... why do we care? Standard errors are used in hypothesis testing, which will be covered later. One thing that we use standard errors for is confidence interval estimation. A confidence interval (CI) is another way of describing the variability of the estimate that we calculate. A CI tells us how precise our estimate is, which can be vital if we are trying to estimate a parameter to use in policy analysis.

Typically, we use 95% CIs. Suppose we calculate that our 95% CI for a parameter estimate $\hat{\theta}$ to be $(c^{95}_{\text{low}}, c^{95}_{\text{high}})$, then we are saying if we could resample our data k times and recalculate the 95% CI each time, 95% of these CIs would contain the true (unknown) parameter value, θ. This is not an intuitive statistical object; however, it is very easy to understand what it says about the estimate. If the CI is very tight, then we have high precision; if loose, then we have low precision.

We calculate a 95% CI as :

$$c^{95}_{\text{low}} = \hat{\theta} - \mathsf{se}(\hat{\theta}) \cdot 1.96 \tag{67}$$

$$c^{95}_{\text{high}} = \hat{\theta} + \mathsf{se}(\hat{\theta}) \cdot 1.96, \tag{68}$$

where 1.96 is the critical value for the 97.5% point.[18] Thus, the standard error directly affects the size of the confidence intervals.

According to Wikipedia, it is important to note what the the CI *is not*:

- "A 95% confidence level does not mean that for a given realized interval there is a 95% probability that the population parameter lies within the interval (i.e., a 95% probability that the interval covers the population parameter)."
- "A 95% confidence level does not mean that there is a 95% probability of the parameter estimate from a repeat of the experiment falling within the confidence interval computed from a given experiment."
- "A 95% confidence level does not mean that 95% of the sample data lie within the confidence interval."

6 Foundational Theorems

This section discusses four foundational and fundamental theorems of statistics. The Law of Large Numbers states that the sample average converges to the true expected value as the sample size increases. The Central Limit Theorem tells us that the distribution of sample means approaches a normal distribution regardless of the underlying distribution (given certain conditions). The Law of Iterated Expectations says that the expected value of a conditional expectation equals the unconditional expectation. The Law of Total Variance

[18] If $Z \sim \mathcal{N}(0,1)$, then $\Pr(Z > 1.96) = 2.5\%$ and so $\Pr(-1.96 < Z < 1.96) \approx 95\%$. We use the standard normal as an application of the Central Limit Theorem, which is discussed below.

decomposes the total variance into the expected conditional variance plus the variance of conditional expectations.

These four theorems work together: the LLN and CLT help us understand sampling and asymptotic behavior, while the laws of iterated expectations and total variance help us understand the relationship between marginal and conditional distributions. Together, they form a powerful theoretical framework for statistical inference.

6.1 Law of Large Numbers

The Law of Large Numbers states that if (i) μ_X exists (is finite) and (ii) $\mathbf{X}_n$ is an iid sample, then $\bar{x}_n \xrightarrow{p} \mu_X$. That is, if there is a random variable with a mean that exists and we collect a random sample, then the sample average is consistent for the mean.

From Wikipedia:

> The LLN is important because it guarantees stable long-term results for the averages of some random events. For example, while a casino may lose money in a single spin of the roulette wheel, its earnings will tend towards a predictable percentage over a large number of spins. Any winning streak by a player will eventually be overcome by the parameters of the game.

6.2 Central Limit Theorem

The Central Limit Theorem says *something very close to* for large enough n, the distribution of $\bar{x}_n$ converges to $\mathcal{N}(\mu, \frac{\sigma^2}{n})$. It turns out there is no *single* version of the CLT – there are many with slightly different assumptions and results.

The classical CLT states: if (1) $\mathbf{X}_n$ is a random sample, (2) $\mathsf{E}[X_i] = \mu_X$ and $\mathsf{Var}(X_i) = \sigma^2 <$ inf, then $\sqrt{n} \cdot (\bar{x}_n - \mu_X) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$. There are other versions.

The CLT is useful widely in statistics, but primarily for econometrics, the CLT allows us to get standard errors. Essentially, it allows us to say $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2_{\hat{\beta}})$, so the empirical standard error of $\hat{\beta}$ is will be our estimate of the square root of $\sigma^2_{\hat{\beta}}$.

Note that the CLT does *not* say that the true population distribution of the random variable X is or must be normal. The CLT says that no matter the population distribution of the random variable X is, under certain conditions the distribution of $\bar{x}_n$ *will be* normal.

6.3 Law of Iterated Expectations

The Law of Iterated Expectations[19] states that $\mathsf{E}[X] = \mathsf{E}[\mathsf{E}[X \mid Y]]$. As simple as this may look, it turns out it is very powerful. Wooldridge considers this one of the most useful facts in doing econometric proofs.

It is easier to see this assuming Y is a discrete random variable. Suppose that $Y \in \{y^1, y^2, y^3\}$, then see that:

$$\mathsf{E}[X] = \Pr(Y = y^1) \cdot \mathsf{E}[X \mid Y = y^1] + \Pr(Y = y^2) \cdot \mathsf{E}[X \mid Y = y^2] + \Pr(Y = y^3) \cdot \mathsf{E}[X \mid Y = y^3]. \quad (69)$$

For example, suppose we are considering whether students will pass a test or not ($X \in \{0, 1\}$) given whether they have studied or not ($Y \in \{0, 1\}$). If a student has studied, then the student has a $\mathsf{E}[X \mid Y = 1] = \pi_{1,1}$ probability of passing; if not, then a $\mathsf{E}[X \mid Y = 0] = \pi_{1,0}$ of passing. Say that $\mathsf{E}[Y] = \lambda$. The expected pass rate of the test equals the portion of students who studied times the probability of passing conditional on studying plus the portion of students that did not study times their probability of passing:

$$\begin{aligned} \mathsf{E}[X] &= \lambda \cdot \pi_{1,1} + (1 - \lambda) \cdot \pi_{1,0} && (70) \\ &= \mathsf{E}[Y = 1] \cdot \mathsf{E}[X \mid Y = 1] + \mathsf{E}[Y = 0] \cdot \mathsf{E}[X \mid Y = 0]. && (71) \end{aligned}$$

6.4 Law of Total Variance

The Law of Total Variance[20] states that $\mathsf{Var}(Y) = \mathsf{E}[\mathsf{Var}(Y \mid X)] + \mathsf{Var}(\mathsf{E}[Y \mid X])$.

The first terms is called the unexplained variance while the second is called the explained variance. The first term is 'unexplained' because it is the average variance in Y that remains even after conditioning on the information we have from X. If X completely explained Y, then we would have 'explained all of Y'. The second term accounts for the variation in Y that comes from the fact that there is variation in our information X, which is 'explained.' In econometrics, we sometimes call the 'unexplained' as the within-group variation and the 'explained' is the between-group variation.

For example, suppose we wanted to explain the variation in grades. The total variance in grades world depend on (1) the variance in grades because some students studied and others did not and (2) the variance in grades within those who studied and within those who did not study. This would be useful to know since (1) tells you how the class's study

[19] It is also known as the Law of Total Expectations, but it is most commonly called LIE.
[20] It is also sometimes called a variance decomposition.

habits determine their grades while (2) tells you how much idiosyncratic factors explain their grades.

Consider a similar example but considering the distribution of wages for different education groups. What would this tell us?

7 Statistical Inference

The final section of this guide is about statistical inference. If the first step is to get an answer to an empirical questions, then the second step is to interpret that answer. All of the knowledge above culminates into statistical inference that then allows us to interpret the empirical work that we do.

7.1 Economic vs. statistical significance

Imagine that you have estimated some regression of an outcome Y on some policy X and gotten some coefficient estimate $\hat{\beta}$, which measures the effect of the policy on the outcome. How do you evaluate your estimate? There are three things to consider:

1. If the estimate was correct, then does it make sense and/or does it empirically matter for explaining outcomes?

2. Is the estimate statistically different from zero ('statistically significant')?

3. How wide is the confidence interval (i.e., how precise) is the estimate?

All three questions are paramount for the empirical analysis. Suppose we estimate the effect of studying on passing a test. We then calculate our estimate and its 95% confidence interval.

First, suppose that coefficient is only 0.1 percentage points; i.e., $\mathsf{E}[X \mid Y = 1] - \mathsf{E}[X \mid Y = 1] = 0.001$, and that its confidence interval is $(0.0005, 0.0015)$. In this case, while studying is statistically significant and precise, it is economically insignificant. Studying just does not matter for the passing the test.

Next, suppose the effect size is 10 percentage points but confidence interval is $(-0.2, 0.4)$. In this case, while the effect size is economically significant, the estimate fails on questions 2 and 3. We may be tempted to say studying does not affect test passing because the estimate is not statistically significant. However, the confidence interval tells us the true

effect could either very negative or very positive. We actually cannot strongly say much other than the estimate is imprecise.

Finally, suppose the effect is 0.1 percentage point and the confidence internal is $(-0.0005, 0.0025)$. In this case, the effect is neither economically nor statistically significant; however, it is actually very precise. In the second example, the estimate was too imprecise to confidently say studying either has no effect (statistically insignificant) or a positive effect (since the estimate was 10 percentage points). Here, we cannot reject that the effect is zero but we can reject that the effect is large. Being able to say when effects are not large is sometimes just as important as being able to provide an exact answer.

7.2 Hypothesis Test Steps

Here, I am going to describe a hypothesis test first without explanation so that you can see how everything fits together, and afterward will explain the different parts.

Suppose we run a regression of Y on X, and so we get $\hat{\beta}$ and $\hat{\sigma} = \mathsf{se}(\hat{\beta})$. We want to test that the coefficient is statistically different than zero. We call this our null hypothesis: $\mathsf{H}_0 : \beta = 0$; our alternative hypothesis is the opposite: $\mathsf{H}_A : \beta \neq 0$. We test this hypothesis by evaluating our test statistic and seeing what it says. Our test statistic is $\mathsf{t}(\hat{\beta}) = \hat{\beta}/\hat{\sigma}$, which follows a t-distribution with degrees of freedom $n - k$. Using this distribution, calculate $\Pr\big((|\,t\,| > |\,\mathsf{t}(\hat{\beta})\,|)\big) = \mathsf{p}$, the p-value. If $\mathsf{p} < 0.01$, then we say we reject the null hypothesis at the 1% level; if if $\mathsf{p} < 0.05$, then we say we reject the null hypothesis at the 5% level; if $\mathsf{p} < 0.10$, then we say we reject the null hypothesis at the 10% level; else, we say that we fail to reject the null hypothesis.

Let's make a list of things that are happening: (1) null and alternative hypothesis; (2) test statistic and test statistic distribution; (3) probability cut-offs.

7.2.1 Null and Alt. Hypotheses

That the hypotheses are about the true parameter! We are always trying to learn about the true parameters. The alternative hypothesis is always the opposite of the null. Usually, the null hypothesis is a 'restriction' and the alternative is the restriction being not true; e.g., $\beta = 0$ is a restriction that the effect is zero.

Note, the null hypothesis can be more complicated that described above. For example, it could be that $\beta > 0$ or it could be that $\beta^2 > \sqrt{(\beta)}$ or could involve multiple parameters $\beta_1 + \beta_2 > \beta_3$.

7.2.2 Test Statistic and Distribution

Generally, the test statistic is a measure that distinguishes the null from alternative hypothesis using the methodology and data we observe. Our test statistics are based on the data we have, and we use them to learn about the true parameter. Usually, because the null hypothesis implies a restricted and simplified model, it can be easier to describe the expected sampling distribution of the test statistic. Once we have this distribution, we can see if the probability that our calculated test statistic is sufficiently 'small' as to cause us to doubt the null hypothesis is true.

The test statistic described above is called the t-statistic and is the estimated coefficient divided by its standard error (also estimated). We call it the t-statistic because using the CLT, $\hat{\beta}$ *and* $\hat{\sigma}$ are both normally distributed and the ratio of normal variables is distributed according to a t-distribution with $n-k$ degrees of freedom ('DOF'), where n is the sample size and k is the number of estimated parameters. However, if we are relying on the CLT, which is based on $n \to \infty$ (so the DOF is also ∞), then we can also use the fact that the t-distribution turns into the standard normal distribution, $\mathcal{N}(0,1)$, as the DOF goes to ∞.

We use the t-statistic for hypothesis tests related to a single equation test (i.e., only one equal sign or one inequality). For hypothesis tests that use multiple equations (i.e., more than one equal or inequality sign), we use the F-statistic. An example of such is: $\mathsf{H}_0 : \beta_1 = 0 \ \& \ \beta_2 = 0$. The F-statistic follows... you guessed it... the F-distribution. It is kinda complicated to explain how we get to this distribution, but essentially it follows from the CLT and the continuous mapping theorem.

7.2.3 Probability Cut-Off

There were three potential probability cut-offs we considered: $0.01, 0.05, 0.10$. We often call these statistical significance levels. We compare the p-value to these levels, and if the p-value is less than level, we say the we can reject the null at the given level.

From Wikipedia, the p-value is "the probability of obtaining test results at least as extreme as the result actually observed, under the assumption that the null hypothesis is correct." Because the test statistic is designed to distinguish the null from the alternative hypothesis, a small p-value tells us the probability that the null is true yet we would get such an extreme value of the test statistic. A standard threshold in economics is 0.05, such that if that probability is less than 5%, then we reject the null.

7.3 False Positive / False Negative

Statistical inference thinks about two ways of being wrong: False Positives and False Negatives. False positives are when we wrong because something we think is true is actually false. False negatives are when we wrong because something we think is false is actually true. Finding an innocent person guilty is a false positive, and finding a guilty person not guilty is a false negative. For some sick reason, someone created another name for these: type 1 and type 2 errors. This is terrible because it does not tell you what the error is! False positives are type 1; false negatives are type 2.

The false positive rate is the probability of getting a positive result when the truth was negative; e.g., the likelihood of finding someone guilty who was actually innocent. The false positive rate for a statistical test is called the *significance* and is denoted by α. We tend to want tests that have low alpha.

The false negative rate is the probability of getting a negative result when the truth is positive; e.g., the likelihood of finding the guilty not guilty. The false negative rate is denoted with a β, and we call $(1-\beta)$ the *statistical power* of a test. We tend to want tests that high power.

7.4 Power

In short, power is the ability of a test to detect that an effect is present in the data. This is usually relevant to help distinguish *small* true effects from zero. If (1) there is a lot of true volatility in the data and (2) the true effects of a policy intervention is small, then you may estimate the true parameter but have too wide standard errors to be sure the effect is real. The easiest way to increase the power of a test is to have more data (large n).

However, suppose you are interested in designing an experiment, and you are wondering how many participants you need to have the ability to find an effect of size $\tilde{\beta}$. A rule of thumb to find the right n for 80% power and 5% significance is:

$$n^{\star} = 16 \cdot \frac{\sigma^2}{\tilde{\beta}}, \tag{72}$$

where σ^2 is the population variance in the outcome variable and the 16 comes from using the 80% power and 5% significance.[21]

[21] Specifically, $(1.96+0.84)^2 \approx (2.8)^2 \approx 7.84\ddot{O}2 \approx 16$; where 1.96 is the critical value of for the significance, 0.84 is the critical value for the power, and 2 is since it is a two-sided test.

If one conducts many hypothesis tests but they have low power, then a larger proportion of tests will effectively become false positives.

7.5 Looking at a Regression Table

Despite all the above, it is actually very easy to quickly to a basic hypothesis test for regression tables. In Figure 19, I show results from a paper.

First, notice that we did not add 'stars' or label whether we put t-statistics, SEs, or p-values. We also only include the most relevant variables for the analysis... we do not overload the table with information that is not relevant to the empirical story we are telling.

But, how can you tell what we did? Usually, this would be explicitly stated in the table footnotes, but there is often an easier way. SEs are always positive, t-statistics always have the same sign as the coefficient, and p-values are always between 0 and 1. This can help quickly figure out what you are looking at. In Table 4, one can use these facts to quickly see that we are using SEs.

Second, using the CLT, we can assume that our t-statistic follows a normal distribution, so we can find out what is the 5% value for a two-sided test... it is 1.96. So, if $\mathsf{t}(\hat{\beta}) > 1.96$, then the coefficient is significant at the 5% level. However, there is an even easier way. If $\hat{\beta} \pm \mathsf{se}(\hat{\beta})$ "crosses zero," then it is *not* significant at the 5% level. For example, in Table 2 column (1), $-0.012 < 0 < -0.012 + 0.032$ (crosses zero), so that coefficient is *not* statistically significant; however, in column (2) $0.161 > 0.161 - 0.080 > 0$ (does not cross zero), so it *is* significant.

Figure 19: Two Tables from Watson & Ziv, 2021

Table 2: The Relationship Between Ownership Concentration and Rent

	(1)	(2)	(3)	(4)	(5)	(6)
	$\ln[\text{Average r}_{j,g,t}]$					
	Panel (A): Manhattan					
$\ln[\text{HHI}_{f(j),g,t}]$	-0.012	0.161	0.075	0.009	0.162	0.075
	(0.032)	(0.080)	(0.076)	(0.038)	(0.076)	(0.076)
$\ln[s_{g,t}^{f(j)}]$				-0.028	0.002	-0.013
				(0.026)	(0.025)	(0.027)
Year FEs	Y	Y	Y	Y	Y	Y
Tract FEs	N	Y	N	N	Y	N
Building FEs	N	N	Y	N	N	Y
Observations	2,519	2,504	2,393	2,519	2,504	2,393
R^2	0.29	0.63	0.75	0.29	0.63	0.75

(a) Table 2

Table 4: Model Parameter Estimates for Four NYC Boroughs

	RC Logit	RC Nested Logit
α	-27.80	-23.74
	(13.97)	(4.23)
ρ		0.069
		(0.043)
BLP F Stat	88.0	32.4
Linear F Stat	111.6	121.9

(b) Table 4