

1 Chapter 1: An Introduction to statistics

AnIntroductionToStatistics.tex or .pdf
Edward Morey - August 30, 2010

1.1 The name of this course is "Mathematical Statistics for Economics" (Econ 7818)

So, this is a course in *statistics*.

What are statistics? *Statistics* is the plural of the word *statistic*.

Definition 1 *A statistic is a function of one or more random variables.*

To know exactly what this means, one must first define a *variable* and then a *random variable*.

Put simply a variable is something that varies. That is, a variable can take on different numerical values (a realized value of a variable is some number). Each number represents a distinct *state*. For example, the variable G might represent gender, where 1 corresponds to the state female and 0 corresponds to the state not-female. Or, for example, if X is a variable such that $0 \leq x \leq 123$, X can take any numerical value between zero and 123, inclusive, the variable X might represent age of a human. Here, I have assumed all humans are not younger than zero and not older than 123 years (everyone would not agree on this lower bound).¹

X is the name of the random variable (e.g. age, price, or amount of sexual activity), and x is a numerical value of X .²

¹From Wikipedia: The longest unambiguously documented lifespan is that of Jeanne Calment of France (1875–1997), who died at age 122 years and 164 days. She met Vincent van Gogh at age 14.[1] This led to her being noticed by the media in 1985, at age 110. Subsequent investigation found that her life was documented in the records of her native city of Arles beyond reasonable question.[2] More evidence for the Calment case has been produced than for any other supercentenarian case, which makes her case a standard among the oldest people recordholders.[citation needed]. http://en.wikipedia.org/wiki/Oldest_people

²The issue of how to distinguish between a random variable and a specific realization of that random variable can be confusing, and the literature is not consistent in how it notationally distinguishes between the two - I won't be either. I will try to use upper case to denote the name of a random variable, and lower-case to denote a specific value of that random variable. However, I, and others, might use x to refer to both and hope the reader can determine which is meant by the context.

To say that the value of a rv can be expressed with a number is not vary restrictive: if X has cardinal properties, e.g. if X represents age, the the numbers have cardinal meaning; if X only has meaning in terms of a ranking, e.g. class rank, then only the ordinal properties of the numbers are important; if X simply denotes categories, e.g. gender or race, then the numbers mean nothing other than different numbers represent different categories.³

Let's assume X is a **random** variable (I will define random variable in a second). In which case

$$y = f(x)$$

$$\beta = g(x)$$

and

$$m = 4 + 7x$$

are each statistics, all of the same random variable, X . The letters f and g are the names of particular functions.

Or more generally, imagine three random variables: X , Y and Z . In which case

$$\alpha = \alpha(x, y, z)$$

$$b = h(x, y, z)$$

$$\beta_1 = \beta_1(x, y, z)$$

and

$$\beta_2 = \beta_2(x, y, z)$$

are each statistics.

So $c = m(b)$ is a rv.

Alternatively, one could let x refer to the rv and let x_i refer to a value i of the rv. This approach will be pretty clear if there is only one rv being considered. But what if there are three rvs? Do I give them different names, like x , y and z ? If so z_i refers to a realized value of z . But, what if instead of x , y and z , I had denoted the three random variables in the text x_1 , x_2 and x_3 where the subscripts now refer to different rv's, not different observation on x . One must be viligant.

We need to be careful and figure out what is going on by the context. Being explicit about what we mean is also a good thing if we don't want the reader to get confused.

³You are probably ready to conclude that I like footnotes. I do; they allow me to digress—a former student accused me of being the "King of digression"—and tangents. I discuss the properties of numbers in Morey (confuser surplus). *Latent Gold*, a statistical program I use, allows one to change the specification of random variables between, cardinal, ordinal and nominal (categorical). The differences between many statistical and economic models are often only the numerical properties of the dependent and independent variables.

All statistics are random variables, but all random variables are not statistics, unless one defines $x = x$ as a function.

1.2 So, what is a random variable?

Definition 2 X is a random variable if it is a variable and if it has a distribution. Said another way, X is a random variable if $\forall a$ and b one can determine the probability that $a \leq x \leq b$ if one knows the distribution of X

Note that X takes specific values (e.g., if X is weight, each of us has a specific weight but weight, in the population, has some distribution.)

The above definition is not self-contained. It requires that we know what a distribution is, and we have yet to define that term, other than we have defined it as something that allows us to calculate $\Pr(a \leq x \leq b)$

Also note the definition requires that X has a distribution, but it does not require that we know what distribution.

The book, *Introduction to the Theory of Statistics* (Mood, Graybill and Boes) defines a continuous random variable as follows:

The variable X is a one-dimensional, continuous random variable if there exists a function $f(x)$ such that $f(x) \geq 0 \forall x$ in the interval $-\infty \leq x \leq \infty$, and the probability that $(a \leq x \leq b)$ is⁴

$$\Pr(a \leq x \leq b) = \int_a^b f(x)dx$$

The function $f(X)$ is called a *density function* (or a *probability density function*). The function, $f(X)$, describes the distribution of X .

Any function, $f(X)$, can serve as a density function as long as

$$f(x) \geq 0, \quad -\infty \leq x \leq \infty$$

and⁵

$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

⁴Note the qualifying adjective *continuous*.

⁵Note that $f(x) \leq 1$ is not a requirement (necessary condition). It is required for certain types of density functions, but not all of them. What types? What is required is that $\int_a^b f(x)dx \leq 1$, which follow from the restriction that $\int_{-\infty}^{+\infty} f(x)dx = 1$

Why do we care about density functions? Economic models typically assume outcomes (how much you drink, whether the interest rate will rise) are the result of some process with a random component: the model contains a random variable. Or said differently, the behavior of a variable in the model is described by some density function.

Two more things:

1. A sample is the result of a random process, so a sample is a vector of random variables. Therefore, a function of a sample is a statistic. More on this in a bit.
2. The realized value of a random variable is not a random variable: it is a fixed number; it does not vary, so not a variable.

For example, consider the rv A , age at death. Assuming the determination of when you "kick the bucket" has a random component, A is a random variable, but once the dice is thrown and Tori "buys the ranch" a_{Tori} is determined, fixed, and not a random variable. Up until that moment, a_{Tori} was both a variable and random, but neither afterwards.⁶

Consider how many hours UnJung, a former student, sleeps a night; assume its determination has a random component. Let S denote the number of hours he sleeps in a night, so s_t is how many hours he sleeps on night t . Before night t , s_t is a rv, but once the night is over, s_t is a fixed number.

An interesting question is whether the world is inherently random—god rolls dice, as in quantum mechanics—or the world is deterministic and it just seems random from our perspective because we cannot observe or measure all of the things that determine things. While interesting, this distinction is not critical for studying statistics in 7818.

⁶"Kick the bucket" and "buy the ranch" are colloquial expressions for dying. There are hundred of colloquial expressions for dying.

1.2.1 Make up a density function for a continuous random variable

Prove that your function is a density function.

Graph your density function.

Some student examples

For a continuous variable x , we can make up a density function

$$f(x) = \begin{cases} e^x & -\infty < x \leq 0 \\ 0 & x > 0 \end{cases}.$$

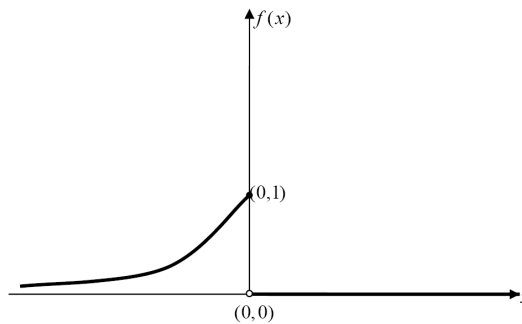
Proof:

1. It is obvious that $f(x) \geq 0$ for all x in the domain $(-\infty, +\infty)$.

2. $\int_{-\infty}^{+\infty} f(x)dx = \int_{-\infty}^0 e^x dx + \int_0^{+\infty} 0 dx = e^x \Big|_{-\infty}^0 + 0 = e^0 - e^{-\infty} = 1 - 0 = 1$.

So, the function $f(x) = \begin{cases} e^x & -\infty < x \leq 0 \\ 0 & x > 0 \end{cases}$ is a density function.

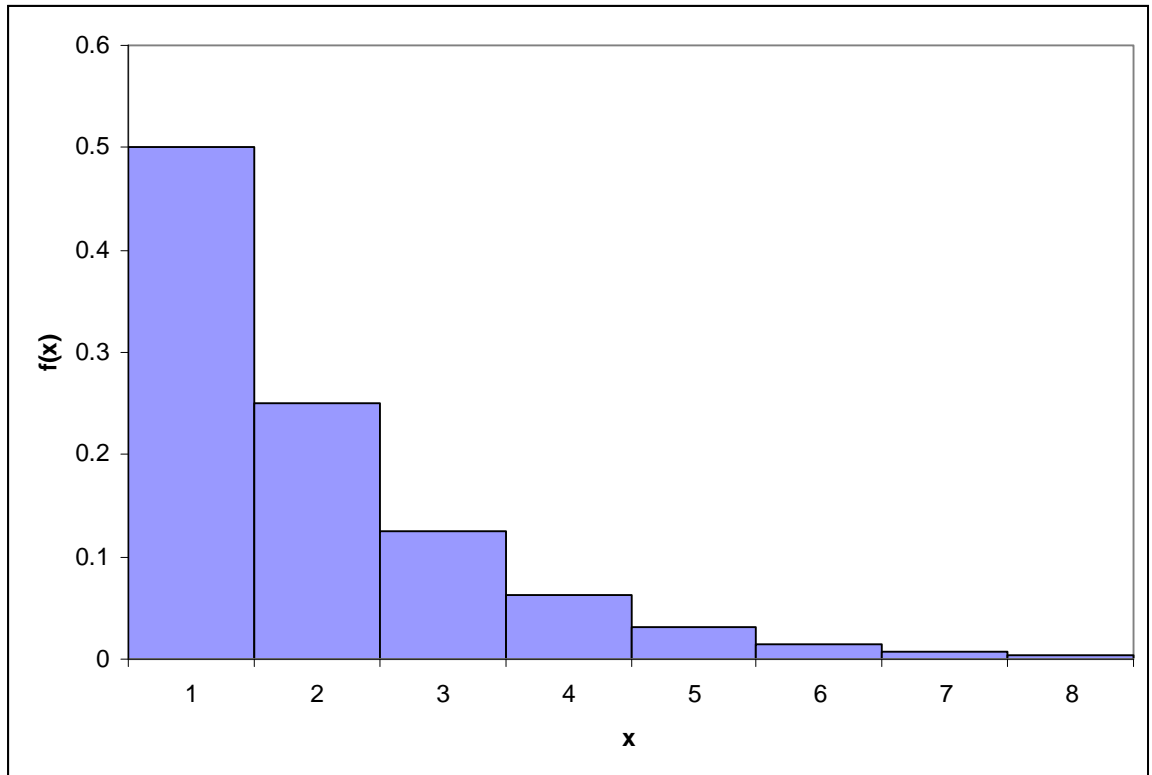
The graph of this density function is as the following:



Another example

$$f(x) = \begin{cases} 0 & \text{if } x < .5 \\ .5^{\text{round}(x)} & \text{if } x \geq .5 \end{cases} \quad \text{where } \text{round}(x) \text{ is defined as the integer}$$

closest to x , with $.5$ rounded up.

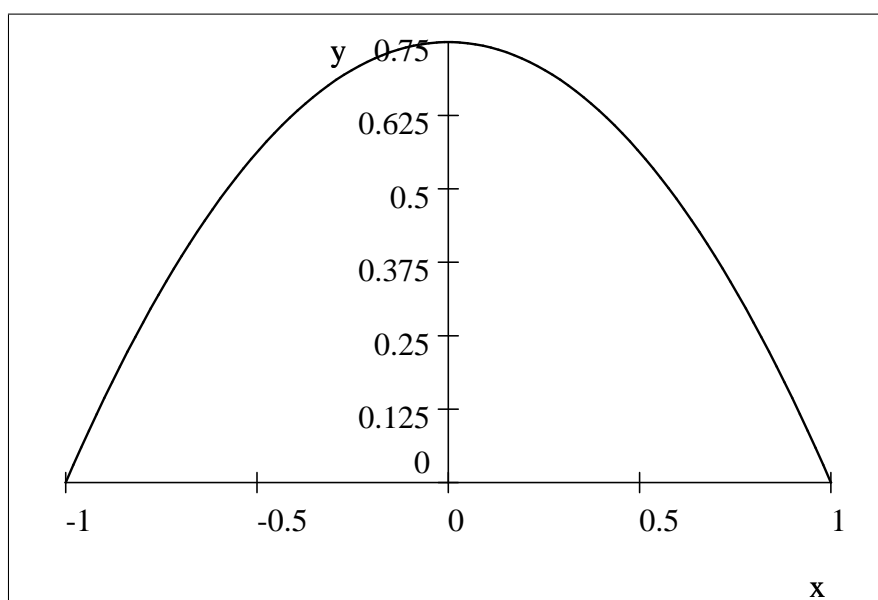


Can you prove this is a density function.

Another example: choose $a > 0$ and define $f(x)$ as follows:

$$f(x) = \begin{cases} 0 & x < -a \\ -\frac{3}{4a^3}x^2 + \frac{3}{4a} & -a \leq x \leq a \\ 0 & x > a. \end{cases}$$

The graph of this function is an upside-down parabola centered around the y -axis, which touches the x -axis at points $-a$ and a , and crosses the y -axis at $\frac{3}{4a}$. Here is what it looks like if $a = 1$.



Note that it is much easier to make up a density function $f(X)$ when X has a limited domain, rather than an infinite domain. It is tough to find functions where the domain is infinite and the area under the function is one.

After one has chosen some population variable to study, how does one decide what to assume about its density function?

For example, what if the rv is number of times living individuals of have been married.

What restrictions might one realistically impose on this distribution.

1.2.2 Note that we used the notion of probability to define a random variable

A variable is a rv if there exists some **probability** that the variable lies in the interval a, b .

It is sometimes easy to forget that statistics are all about determining or estimating probabilities.

For example, in OLS regression analysis, something many of you know, the probabilities are not always explicit. Many, unfortunately, fixate on the OLS parameter estimates, but the probabilities are there. For example, given the OLS parameter estimates, what is the probability that the true value of the parameter lies between a and b ? We should be most interested in that question.

1.3 What do statisticians do?

Put simply, statisticians do statistics. Sam, the statistician, does statistics in the same sense that Tiger, the golfer, does golf, but Sam's does not do statistics as well as Tiger does golf.

Theoretical statisticians and theoretical econometricians propose statistics to better understand random processes—most random processes are determined, in part, by other random processes. They then try to determine the properties of those statistics.

For example, for some random process with unknown parameters, statisticians develop statistics based on samples of random variables from that process, statistics that either help to describe the process or to our estimates of the unknown parameters or both.

The goal is to find a statistic that describe the process and that has desirable properties for the question at hand. (To accomplish this goal, one has to obviously decide on what properties are, and are not, desirable - what is desirable also depends on the question at hand.)

A statistic is like a *significant other*, one wants one with desirable properties.

Applied statisticians use these desirable statistics, along with data, to estimate things about the world. Implicit in the approach is the idea that observed outcomes are the result of some random process. Remember that a statistic is a function of random variables, and the function has parameters, which the statistician wants to estimate—estimation means using data to estimate the parameters in statistics.



k0190546 www.fotosearch.com

If one reads statistics books, one quickly gets the idea that statisticians have a thing about urns and drawing balls from urns. When entering the kitchen to make the kids breakfast, the statistician takes the lid off the "breakfast urn" and draws a ping-pong ball. If the ball is red it is eggs, blue cereal, Maybe there is different urn for weekend breakfasts. Tonight will it be TV in bed or sex with the spouse, it all depends of the draw from the bedroom urn.

1.3.1 Why doing statistics is more difficult than watching TV

A statistic is a function of random variables, so a random variable. Random variables have density functions, so a statistic has a density function. It is often damn difficult to determine that density function.

For example, imagine that X , Y and Z are each a random variable, such that X has the famous Guber distribution, Y has the famous Gomer distribution, and Z has the not-so-famous Snerd distribution.

Now define the statistic

$$\begin{aligned} s &= s(x, y, z) \\ &= \exp(\tan(x^2y) + xy(\ln(.5z \exp(zy)))) \end{aligned}$$

and figure out the density function for S , $f(s)$.

1.3.2 Econometricians are, one would hope, a subspecies of statisticians

Some quotes from Peter Kennedy's *A Guide to Econometrics*

Econometrics is what econometricians do.

Econometrics is the study of the application of statistical methods to the analysis of economic phenomena.

What distinguishes an econometrician from a statistician is the former's preoccupation with problems caused by violations of statistician's standard assumptions; owing to the nature of economic relationships and the lack of controlled experimentation, these assumptions are seldom met.

Econometricians are often criticized, and often by other econometricians. They have a bad habit of ignoring the quality of their data. Again, some quotes from *A Guide to Econometrics*⁷

Econometricians are often accused of using sledgehammers to crack open peanuts while turning a blind eye to data deficiencies and the many questionable assumptions required for the successful application of these many techniques.

Econometric theory is like an exquisitely balanced French recipe, spelling out precisely with how many turns to mix the sauce, how many carats of spice to add, and for how many milliseconds to bake the mixture at exactly 474 degrees of temperature. But when the statistical cook turns to raw materials, he finds that hearts of cactus fruit are unavailable, so he substitutes chunks of cantaloupe; where the recipe calls for vermicelli he used shredded wheat; and he substitutes green garment die for curry, ping-pong balls for turtle's eggs, and for Chalifougnac vintage 1883, a can of turpentine. (Valavanis)

⁷Google Books at <http://books.google.com/books?id=B8I5SP69e4kC&dq=Kennedy+a+guide+to+econometrics&printsec=from>

It is the preparation skill of the econometric chef that catches the professional eye, not the quality of the raw materials in the meal, or the effort that went into procuring them. (Griliches⁸)

The art of the econometrician consists in finding the set of assumptions which are both sufficiently specific and sufficiently realistic to allow him to take the best possible advantage of the data available to him (Malinvaud)⁹

The applied econometrician: The applied econometrician, unlike the theoretical econometrician, needs to worry as much about her data as about the theory. The forecasts and predictions generated by the econometric model are only as good as the data that produced them.

A well-known econometrician recently mentioned to me that he was hired by a group of wealthy gamblers to use his choice-modeling skills to predict the outcome of horse races. It might be important that he get it right.¹⁰

1.4 Other perspective on statistics and statisticians

The following quote is from the front of *The Advanced Theory of Statistics*, Vol. 2, by M.G. Kendall and A. Stuart. They attributed it to the fictitious K.A.C. Manderville, *The Undoing of Lamia Gurdleneck*.)

"You haven't told me yet," said Lady Nuttal, "what it is your fiancé does for a living."

"He's an statistician." replied Lamia, with an annoying sense of being on the defensive.

Lady Nuttal was obviously taken aback. It had not occurred to her that statisticians entered into normal social relationships. The species, she would have surmised, was perpetuated in some collateral manner, like mules.

"But Aunt Sara, it's a very interesting profession," said Lamia warmly.

⁸More about Zvi Griliches at Wiki: http://en.wikipedia.org/wiki/Zvi_Griliches

⁹Edmond Malinvaud published in 1964, *Statistical Methods in Econometrics*. More details about Edmond can be found at Wikihttp://en.wikipedia.org/wiki/Edmond_Malinvaud

¹⁰It should not surprise that many statisticians and econometricians gamble; probability theory developed to improve one's odds in games of chance.

"I don't doubt it," said her aunt, who obviously doubted it very much. "To express anything important in mere figures is so plainly impossible that there must be endless scope for well-paid advice on how to do it. But don't you think that life with an statistician would be rather, shall we say, humdrum?"

Lamia was silent. She felt reluctant to discuss the surprising depth of emotional possibility which she had discovered below Edgar's numerical veneer.

"It's not the figures themselves," she said finally, "it's what you do with them that matter."

Some additional quotes:

To understand God's thoughts we must study statistics, for these are the measure of His purpose. (Florence Nightingale)

Statistics are like a bikini. What they reveal is suggestive, but what they conceal is vital. (Aaron Levenstein)

The first lesson that you must learn is, when I call for statistics about the rate of infant mortality, what I want is proof that fewer babies died when I was Prime Minister than when anyone else was Prime Minister. That is a political statistic. (Winston Churchill)

There are three kinds of lies: lies, damned lies, and statistics. (Benjamin Disraeli, but sometimes attributed to Mark Twain)

To bad we can't email Florence and ask here what the hell she meant. Maybe she mean that "casualty statistics", estimates of maimed and dead soldiers, are a "measure of [God's] purpose": part of God's big plan.



1.4.1 How Takeshi Amemiya defines *statistics*

Statistics is the science of assigning a probability to an event on the basis of experiments (Amemiya, Introduction to statistics and econometrics, p. 2)¹¹

Statistics is the science of observing data and making inferences about the characteristics of the random mechanism that has generated the data. (p. 3)

¹¹Takeshi Amemiya is a professor of economics and classics at Stanford. His home page is <http://economics.stanford.edu/faculty/amemiya>

Both of statements suggest that one needs data to do statistics. I would disagree; one can develop statistics and investigate their properties without ever seeing any data. .

So, how does my definition differ from those of Amemiya. Mine is

more technical math-driven, Amemiya's more a *what-is-it-for* definition. In his first definition, Amemiya gives the purpose: determining or estimating probabilities (consistent with my definition), and in his second definition, while not mentioning probability, he does mention *random*. One might consider appending my definition on the front of his definitions.

Implicit in his definitions, and in much of econometrics, is the assumption that what we observe in the world is the result of draws from populations where different outcomes have different probabilities of occurring. Think in terms of drawing one or more balls from an urn, where the urn holds different colored balls in different proportions.

Consider an urn that contains all dead smokers, pickled in brine, and one wants to determine the probability that a smoker will get lung cancer - one draws a sample of dead smokers from the urn.¹² One biopsies the lungs of each

to determine whether the rv Cancer, C , takes a value of 0 or 1 for that individual. The result is a vector of random variables, c_1, c_2, \dots, c_s , where c_3 indicates the cancer status of the third guy drawn. One plugs these observations, the sample, into a statistic to estimate the probability that a smoker will get lung cancer.

With his example in mind, consider Amemiya's definition of a random variable: "A random mechanism whose outcomes are real numbers is called a *random variable*." (p. 4).¹³

He goes on to say, "The characteristics of a continuous random variable are captured by a *density function*." (p. 4 - later in the book he provides amore technical definitions of a rv). (Note that whether one has lung cancer is not a continuous rv, at least not given the way we define cancer)

On to his third definition:

¹²Alternatively, imagine a cemetery where all and only smokers are buried. One digs up a bunch of the decomposing and takes from each a snip of lung tissue to see whether the smoker had lung cancer. Here Reference the study that dug up frozen guys from WWI to see what kind of flu they had.

¹³I would modify this to "whose outcomes can be expressed with real numbers." For example one would still have a random variable if the variable was hair color and one used letters of the alphabet, rather than numbers, to denote the different colors hair can take.

"Statistics is the science of estimating the probability distribution of a random variable on the basis of repeated observations drawn from the same random variable."¹⁴ (p. 4)

¹⁴The term density function is typically only used to describe the distribution of the rv if the rv varies continuously over some range (it is a *continuous rv*). If a rv can take only a countable number of values (it is a *discrete rv*), each with some probability, we don't call its distribution a *density function*. Rather we call it a *discrete distribution*. The term *probability distribution* refers to either a density function, a discrete distribution, or some combination of the two.

1.5 Many statistics of interest are call *estimators*

An *estimator* is a type of statistic. Specifically, it, with data, generates an estimate of a parameter, or parameter range, in the data-generating process.

We assume members of our population-of-interest are generated by a process, a process with a random component - a *data-generating process*.

And assume members of the population can be characterized in terms of some small number of random variables. A member is simply a realization of those random variables.

Write down a simple data-generating process:

For example, assume that the rv of interest is y and y_i is a realization of y .¹⁵

I am going to assume that $y_i = ax^{\varepsilon_i}$ where $\varepsilon \sim N(0, \sigma)$

or another data-generating process:

Assume the rv of interest is W , glasses of wine a day, and $f(w) = \frac{\mu^w e^{-\mu}}{w!}$ where $\mu > 0$.

If the population is all humans now alive we might be interested in how this population varies in terms of age, gender, height and weight. We might be willing to assume these four variables are random variables - their realized values are generated by a random process.

Each of you can be described as realized value of that random process: you have some age, gender, height and weight.



One's interest in this population might be in finding someone to date. You are writing out your application for the online dating service and have gotten to the question about what kind of person you want to date. You want a female between 25 and 30, over six feet tall and less than 150 pounds, but are not sure you should be this restrictive, maybe they only occur rarely in the population

So, you ask yourself, "what is the probability someone has signed up who is female, between 25 and 30 years old, over 6 ft. tall, and weighs less than 150 pounds." To answer this questions, you need to learn about/estimate the joint density function for humans that have signed up for this dating service, and then use it to determine the probability that your dream date exists—keep in mind that you might not be her dream date.

¹⁵Note that I have broken my "rule" about uppercase for the name of the rv.

To say that that members of the population are generated by a process with a random component is equivalent to saying that each member is a draw from some density function. That density function has some functional form and we want to estimate its form and its parameters.

Said another way, populations have properties and we want to come up with estimates of those properties.

Said another way, what we observe is the outcome of a process that is driven by parameters, and we want to estimate those population parameters.

1.5.1 Consider something some people do: smoke cigarettes.

One could define a random variable c as the number of cigarettes smoked per day by an individual, where c_i denotes the number of cigarettes smoked by individual i : c is a random variable and c_i is a realized value of this random variable.¹⁶

We might want to learn about the distribution of this random variable in our population of interest: determine its density function. The data generating process is draws from that density function.

Note the term *population of interest*. For example, the density function for cigarettes smoked by residents of Italy is very different from the density function for cigarettes smoked by residents of the U.S. And the density function for cigarettes smoked by foreign, male graduates students in Boulder is different from the density function for all Boulder residents.

What properties, if any, must these density function have?

Can the number of cigarettes smoked take any value or must it be an integer? Can it be a negative number? Can it be zero? Can it be 1000 a day? Someone want to check the world record for number of cigarettes smoked in 24 hours?

Go to http://www.jimmouth.com/tv04_body.html to see some idiot smoke 159 cigarettes at once

¹⁶Not that for many i , $c_i = 0$, particularly residents of Boulder. The only people in Boulder who smoke appear to be foreign graduate students.

More specifically, we might want to estimate what the distribution of c would be when cigarettes cost a dollar each, and compare this to what the distribution would be if cigarettes cost 10 cents each: this is a problem in demand estimation.

Make sure you understand why this is a problem in **demand estimation**.

Econometricians want to estimate the properties of populations (humans, smokers, interest rates, prices). We do this by taking a sample(s) from the population - we sample random variables that describe the population.

We then propose statistics of the sampled values of those random variables, statistics that will hopefully be good estimators of population properties. That is, we want to estimate population parameters. **The statistics that we will use to estimate population parameters are called estimators.** We want our estimators to do a good job of estimating the population parameter.

An estimator is a function of random variables. If one plugs particular values of the random variables into the function one gets an *estimate*. Note the difference between *estimators* and *estimates* - estimators are functions, estimates are realized values of an estimator - estimates are outcomes/numbers.

1.5.2 One population parameter that is often of interest is its mean

Consider some rv H . If the population is small we can sometimes observe the whole population. If so, we can calculate (not estimate) the population mean.

But, most of the time we do not observe the whole population, so are limited to estimating the mean of the rv H . The function that we use to estimate the mean is an estimator. The inputs into this function are one or more rv's. Plugging in a vector of realized values of the rv's, out comes an estimate of mean H - remember that the mean of H is not a rv, but our estimate of it is a rv.

Different realizations of the random variables will generate different estimates of the population mean of H - the estimated mean will vary from sample to sample, have *sampling variation*.

For example, assume the goal is estimating the mean weight in the U.S. population, ω .¹⁷

¹⁷Note that we have assumed that there is a mean weight. Not all random processes have finite means. The mean weight in the U.S. population continuously increases. Maybe it will eventually reach infinity.

According to WolframAlpha the mean is 180 and the median is 173. How did they get their estimate of the mean?

We observe four weights in the population: denote the weight of the first person observed, w_1 , second person w_2 , third w_3 and fourth w_4 .¹⁸

Every time we sample, we get four different observations: a different sample. In the U.S. population there is a very large number of different possible samples (different sets of 4 people). Let $\mathbf{w}^s \equiv (w_1^s, w_2^s, w_3^s, w_4^s)$ denote sample s . \mathbf{w}^s is a vector of four random variables, so any function of \mathbf{w}^s is a statistic.

Consider the following three statistics

$$\tilde{w} = f(w_1, w_2, w_3, w_4) = w_1 + 3w_2 + (w_3w_4)^2$$

$$\hat{w} = g(w_1, w_4) = \frac{w_1 + w_4}{2}$$

$$\bar{w} = h(w_1, w_2, w_3, w_4) = (.25 \ln w_1 + .25e^{w_2})^{w_3} + w_4$$

Where did these three statistics come from? I made up three functions of the four random variables.

I now declare each of these an estimator of ω ; anything can be an estimator of anything, so what I declare is not untrue, each is an estimator of ω . That said, they may be lousy estimators of ω .

Every time we plug in values from a different sample, we will get new estimates. For example $\bar{w}^s = h(w_1^s, w_2^s, w_3^s, w_4^s) = (.25 \ln w_1^s + .25e^{w_2^s})^{w_3^s} + w_4^s$ is the estimate of \bar{w} for sample s .

If God said that $\hat{w} = g(w_1, w_4) = \frac{w_1 + w_4}{2}$ was the best estimator of ω , the applied statistician/econometrician would always use this estimator to estimate ω , no matter what sample they had collected.¹⁹

God is either unavaible or unwilling; so, we need to decide which of all feasible estimators is the preferred estimator (which has the most desirable properties). To determine which is the preferred estimator, from those available, we ask the theorists what properties we would like an estimator to have (not all theorists agree), and which estimator has the most of those properties.

¹⁸Note that here the subscript refers to different observations on w , not to four different rv's. But, that said, one could also think of them as four different rv's. For example in every sample there will be a first observation, w_1 , and this will vary from sample to sample.

¹⁹This would be an interesting God. If she wanted to be helpful, why didn't she just tell us ω ?

Since we can never know the true population mean of H , ω , we cannot judge an estimate of mean H by how close it is to the true value. (If we knew the true value, we would not need to do estimation.)

We judge estimators, not estimates, this point is lost on many souls—those souls should be damned to Purgatory, maybe the third level of Purgatory.

Words that come to mind when we think about the properties of an estimator include *simple*, *linear*, *unbiased* (vs. *biased*), *efficient*, *asymptotically unbiased*, *consistent*, and *easy to estimate*.

1.5.3 So, how does my description of estimators relate to good-old *ordinary least squares* (OLS)?

You have to wonder.

Consider a one-parameter version of the classic linear-regression model

$$c_i = c(p_i, \varepsilon_i) = 25 + \beta p_i + \varepsilon_i \quad (1)$$

where c_i is the number of cigarettes consumed by individual i , and p_i is the price of cigarettes for individual i (assumed a variable but not a random variable). ε is assumed a random variable (rv) and ε_i is a random draw from ε . Assume that the density function of ε is normal with mean 0 and variance σ^2 . β is a parameter, not a variable, it has some fixed value in the population of interest. An estimation problem only exists because we do not know β or σ

First note that ε is a rv, so the c is a rv, and $c(p, \varepsilon) = 25 + \beta p + \varepsilon$ is a statistic (a function of a rv).

Equation 1 describes the process by which cigarette consumption is determined. Note that a statistical model/process has been assumed: the process/model has a random component and we have assumed the density function for this rv belongs to the family of normal distributions.²⁰

We have assumed most of the estimation problem away. All that is unknown about the population is the values of parameters β and σ^2 .²¹ These we want to estimate.

We want an estimator for β .

We will use that estimator, along with realized values of the rv's that are the variables in the estimator, to get an estimate of β .

In OLS, we make our estimator of β a function of a sample drawn from the assumed population. In this case, one observation in the sample is the i th pair drawn, (c_i, p_i) - we don't observe the ε_i . A sample consists of N drawn pairs: $(c_1, p_1), (c_2, p_2), \dots, (c_N, p_N)$.

The OLS **estimator**/statistic of β is the b that minimizes

$$\sum_{i=1}^N (c_i - bp_i)^2$$

²⁰Note that this is a pretty stupid (unrealistic) model because most people smoke no cigarettes, in addition no one smokes a negative number of cigarettes, so consumption cannot be normally distributed.

²¹Econometricians like to assume away most of the estimation problem. We impose a lot of assumptions on our models, often independently of anything the data might suggest.

Denote this estimator b_{OLS} . Every sample taken will generate a different b_{OLS} estimate of β . Note that b_{OLS} is a rv - β is not a rv; it is a constant. Let b_{OLS}^s denote the OLS estimate generated by sample s .

As applied econometricians, we often mistakenly concentrate on the obtained estimate rather than keeping in mind that our b_{OLS} , b_{OLS}^1 , is just one draw from a distribution of b_{OLS} .²²

That is, b_{OLS} is a random variable with some density function $f(b_{OLS})$.

Much of the work of the classical linear regression model has to do with deriving that density function.

Once we have it, we can answer questions such as "Given β , what is the probability that an estimate, b_{OLS} , will be between $(\beta - \alpha)$ and $(\beta + \alpha)$?" Or, of more relevance, "What is the probability that β is between $(b_{OLS} - \alpha)$ and $(b_{OLS} + \alpha)$?"

So, put simply, the OLS estimator is a special type of statistic, an estimator. And, OLS estimates are rv's with some distribution.

We like OLS estimates - when we assume the classical linear-regression model - because we can show that the OLS estimator has nice properties: it is, if one buys the assumptions, BLUE (a Best Linear Unbiased Estimator).

Note that what has nice properties is the estimator, **not** any particular estimate generated by the estimator. Our actual OLS estimate of β often sucks.

²² b_{OLS}^1 is the estimate from the first sample. I assume that most of the time we only collect one sample. When we do simulations, we will collect many samples.

Note that in all of my years as an applied econometrician I have published many models, but I have never published a paper that report the results of a linear regression.

There is much more to econometrics than linear regressions. Consider again urns. Urns might sound far afield from what econometricians do, but they're not.

Drawing a sample is akin to drawing balls from urns. Consider a sample of c, p pairs. One could view the world as consisting of a number of urns, each corresponding to a different price of cigarettes, for example, urn sixteen might include cigarette consumption by everyone who faces a price of \$3.00 a pack for cigarettes.