Michael Boozer
IDE Econometrics
Economics 558

Review of Probability and Statistics: Definitions and Concepts

A probability distribution for a random variable X completely characterizes its behavior. We follow the usual convention of letting upper case letters like X represent random variables, and lower case versions of them like x represent particular realizations of the random variable. A familiar example of this notation is to let X represent the roll of a die (before it happens) and once the die is rolled, a specific number from the set $\{1, 2, 3, 4, 5, 6\}$ occurs, and this constitutes the realization x of the roll. The example of the roll of the die is also an example of a discrete random variable, where the number of possible realizations is 6. In general for discrete random variables, we let the parameter k denote the (finite and integer valued) number of possible realizations for a discrete random variable X, and the set of realizations be $\{x_1, x_2, ..., x_k\}$. Associated with each outcome of a discrete random variable we have the probability of its occurrence, and the notation for that is:

$$p_i \equiv \text{Prob}(X = x_i) \qquad \forall \;\; i = 1, ..., k \tag{1}$$

where the p_i are the shorthand notation for the probability of each realization x_i, and the p_i must all be non-negative (i.e. $p_i \geqslant 0 \;\; \forall \;\; i$) and must sum to 1 (i.e. $\sum_{i=1}^{k} p_i = 1$ which basically requires that the list of potential realizations be exhaustive).

Now with real data, due to the finite precision with which things like incomes, weights, distances, rainfall amounts, etc are measured, all *data* are ultimately discrete. Nonetheless, we treat such data as arising from a continuous random variable, where the number of realizations is no longer finite (and in fact uncountable). Continuous random variables are in a sense easier to work with, so going forward, we tend to assume we are simply working with continuous random variables, even if that may not always be the case. The basic difference is that we use integral expressions to define the relevant quantities for continuous random variables, and we use summation operators for discrete random variables. Also, owing to the purely continuous nature of a continuous random variable, the probability of a realization being any *particular* number is zero - the idea being that if say weight is the random variable in question, the probability that someone's weight is exactly 138.132 pounds is zero - there is always a precision to which anyone's weight will never be *exactly* this number (although again, as I mention above, with real *data*, this of course can happen - just not with the underlying continuous random variable which is generating the data). Thus, for a continuous random variable we define and work with a continuous probability density function which we frequently abbreviate as the pdf, and we use the notation of $f(x)$. $f(x)$ has the properties that: (i) $f(x) \geqslant 0 \;\; \forall \;\; x$ (ii) $\int_{-\infty}^{+\infty} f(x)\, dx = 1$ (iii) $\int_{a}^{b} f(x)\, dx = \text{Prob}[a < X < b]$.

From the last property, one can see the implication that Prob$[X = a] = 0$ for any particular number a when X has a continuous density, as I discussed before, and that as a related result Prob$[a < X < b]$ = Prob$[a \leqslant X \leqslant b]$ - it is immaterial if we use strict inequalities for continuous random variables or not.

Associated with the pdf is its integral, the cumulative distribution function or cdf defined by: $F(b) \equiv \int_{-\infty}^{b} f(x)\ dx$ which is the mass under a pdf $f(x)$ up to the specified point b. It then follows from the definitions and properties of the pdf just given that (i) $F(\infty) = 1$ (the pdf integrates to 1) and (ii) Prob$[a < X < b] = F(b) - F(a)$. Notice that for $b \geqslant a$, this is a well defined probability, and so it requires that the associated probability of the event $a < X < b$ be non-negative - i,e, that $F(b) \geqslant F(a)$, which is the requirement that $F(\bullet)$ is everywhere a non-decreasing function. Also notice since the pdf integrates to 1, the probability of the event $a < X < b$ is guaranteed to be $\leqslant 1$ satisfying the other criterion of a proper probability (that is be non-negative and less than or equal to 1). I leave it to you to be sure you can define the cdf for a discrete pdf - can you think of what the diagram of the cdf for a discrete random variable will look like? These are good thought exercises to make sure you understand the basic definitions and concepts given here.

With the pdf and the associated cdf in hand, we can now define the moments of a random variable. While a pdf describes an entire distribution, often we wish to work with particular features of the distribution, such as a measure of 'central tendency' like the expected value also known as the population mean. (When we come to talk about data and estimators using the data, we will define the sample mean which is a function of the data and is almost surely something you know from your early schooling - the population mean does not pertain to data, but instead is an algebraic function of the pdf, defined next.) The notation and definition of the expected value for a discrete random variable is:

$$E(X) \equiv \sum_{i=1}^{k} p_i x_i \tag{2}$$

and for the continuous case, the definition is:

$$E(X) \equiv \int_{-\infty}^{+\infty} x f(x) dx \tag{3}$$

and in either case, we also use the shorthand notation of $\mu_x \equiv E(X)$ in many instances. Both expressions lead to the intuitive notion of the expected value or population mean as being the probability-weighted average of the possible realizations, hence the linkage to the concept of "central tendency". The population mean or expected value is the so-called "first moment" and it is in essence a measure of the 'location' of the pdf for the random variable X, as it provides a measure of where on the real line the density of the random variable X tends to lie. However, as the population mean is just one number, obviously it cannot by itself capture all of the features of the entire distribution (at least in general), and so for example, the expected value tells us nothing about the spread of the distribution of X across the real line. The formal terminology for the usual measure of spread is the (population) variance of a random variable X, which is defined as:

$$Var(X) \equiv E\left\{[X - E(X)]^2\right\} = E(X^2) - [E(X)]^2 \tag{4}$$

Notice that since $Var(X)$ is defined as an expectation of something squared, there is no way it can ever go negative, and thus by definition $Var(X) \geqslant 0$, which implies that $E(X^2) \geqslant [E(X)]^2$ and notice that except in exceptional circumstances, $E(X^2) \neq [E(X)]^2$ – please do not make that mistake. By its definition, the expectations operator is a *linear operator* which means it "passes through" linear functions of random variables, as in $E[aX+b] = aE(X)+b$, where a and b are non-random constants or parameters. But while $E[aX^2+b] = aE(X^2)+b$ it does **not** equal $a[E(X)]^2+b$ simply because the function $g(X) = X^2$ is not a *linear* function of X.

Finally, back on the topic of the definition of the variance, it is the so-called (centered) second moment of the random variable X, and often we will use the short-hand notation of $Var(X) \equiv \sigma_x^2$ and since this is in squared units of the original random variable X, we also define the square root of the variance as the *standard deviation* as

$$\sigma_x \equiv \sqrt{Var(X)} \tag{5}$$

as a measure of spread, but in the original units of the random variable. Some questions for thought here: in terms of the integral or summation sign operators, what are the definitions of the variance in the continuous and discrete cases? If we change the mean of a random variable X how does that change the resulting variance? If the original random variable X was measured in US dollars, but now we switch to measuring it in British pounds, how will that change the mean, the variance, and the standard deviation?

As I mentioned in the previous paragraph, the expectations operator is a *linear operator* and so we can use it as an 'operator onto itself' without having to always go back to its original definition in terms of integral signs if we remember to respect this property. So as another example, $E(\frac{1}{X}) \neq \frac{1}{E(X)}$, yet this is also a common error made when people first start working with expectations operators. What about the same properties of the Variance operator? Well the variance operator isn't ever a linear operator, but a very useful property of how it interacts with linear functions is that $Var[aX + b] = a^2 Var(X)$, which can be shown by just using the original definition of the variance (you should confirm this result to yourself). Intuitively, the parameter b is just a location shifter, and so not surprisingly, it has no influence on the variance of the linear function of X we are considering here, which is $aX + b$. The fact that the parameter a on the other hand *multiplies* the random variable X means that it effectively changes it units, and since a variance can never be negative, yet both a and/or b can be negative numbers (we never restricted them in any way), it makes intuitive sense that the way in which a influences the expression for the variance must be in a way that does it non-negatively - i.e. it squares it. Again, this is the intuition, but you should do the underlying algebra to prove this to yourself (and definitely see me if you cannot).

0.1 Joint Distributions

Thus far we have discussed distributions and moments for *univariate* distributions - that is, distributions over a single random variable. While these are incredibly important for describing random variables and so-called 'descriptive statistics' of various variables in datasets, ultimately econometrics is about *joint* or *multivariate* distributions which pertain to the joint behavior of many random variables. To start, we do like much of economics and econometrics and consider the simplest possible case of two random variables, denote them as X and Y. Let the joint pdf for (X, Y) be denoted as $f(x, y)$, and as in the univariate case this joint density must be everywhere non-negative $f(x, y) \geqslant 0$ and must integrate over the domains of X and Y to 1 - i.e. $\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y)\, dx\, dy = 1$. With a joint distribution, we can now consider two further associated pdf's that we could not consider in the univariate case. The first is the *marginal distribution* which integrates over the domain of one of the random variables, say y, as in $\int_{-\infty}^{+\infty} f(x, y)\, dy \equiv f(x)$. Intuitively, the marginal distribution is the 'average' of the joint distribution over the y dimension - which is why it 'averages out' and so the resulting marginal distribution is not functionally dependent on y. Once we have defined the marginal distribution, then we can also go and define the *conditional distribution* of say y conditional on x as:

$$f(y|x) \equiv \frac{f(x, y)}{f(x)} \tag{6}$$

which describes the behavior of the joint distribution for a *particular* realization or value of $X = x$. Thus, to think of a concrete example, let Y=Income of a person and X=Education of a person. The joint distribution $f(x, y)$ describes the relationship of Income and Education in the population. The marginal distribution $f(x)$ describes the univariate distribution of educational levels in the population (and of course there is also the univariate distribution of incomes $f(y)$ as well) and so $f(y|x)$ describes the univariate distribution of income for a *given* level of education, say $x = 16$ which would be for college graduates. Presumably as education increases, incomes increase as well, but the conditional distributions across various educational levels also reveal if incomes become more or less disperse as educational levels rise, and if the changes in dispersion are mainly in the 'left tail' of the distribution, which is the relatively poor people, or in the 'right tail' which are the relatively rich people. Obviously the marginal and conditional densities are of great value as statistical tools in their own right, even before we use them to define further statistical quantities.

Notice also the marginal and conditional densities are themselves 'proper' densities in that they satisfy the non-negativity requirement, from the definition of the marginal distribution $f(x)$ is is obvious that it is an average of the joint distribution $f(x, y)$, and so $f(x)$ will always be $\geqslant 0$ as long as $f(x, y) \geqslant 0 \; \forall \; x, y$. It is also easy to see the marginal density itself integrates to 1, since by definition, $\int_{-\infty}^{+\infty} f(x)dx \equiv \int_{-\infty}^{+\infty} \left[\int_{-\infty}^{+\infty} f(x, y)dy \right] dx = 1$. Similar proofs demonstrate the same attributes for the conditional density, and so I leave them to

IMPORTANT

from Morris DeGroot "Probability and Statistics"

For each fixed value of y, the function g_1 will be a p.d.f. for X over the real line, since $g_1(x|y) \geq 0$ and

$$\int_{-\infty}^{\infty} g_1(x|y)\,dx = 1.$$

It should be noted that Eq. (2) and Eq. (4) are identical. However, Eq. (2) was *derived* as the conditional probability that $X = x$ given that $Y = y$, whereas Eq. (4) was *defined* to be the value of the conditional p.d.f. of X given that $Y = y$.

The definition given in Eq. (4) has an interpretation that can be understood by considering Fig. 3.15. The joint p.d.f. f defines a surface over the xy-plane for which the height $f(x, y)$ at any point (x, y) represents the relative likelihood of that point. For instance, if it is known that $Y = y_0$, then the point (x, y) must lie on the line $y = y_0$ in the xy-plane, and the relative likelihood of any point (x, y_0) on this line is $f(x, y_0)$. Hence, the conditional p.d.f. $g_1(x|y_0)$ of X should be proportional to $f(x, y_0)$. In other words, $g_1(x|y_0)$ is essentially the same as $f(x, y_0)$, but it includes a constant factor $1/[f_2(y_0)]$ which is required to make the conditional p.d.f. integrate to unity over all values of X.

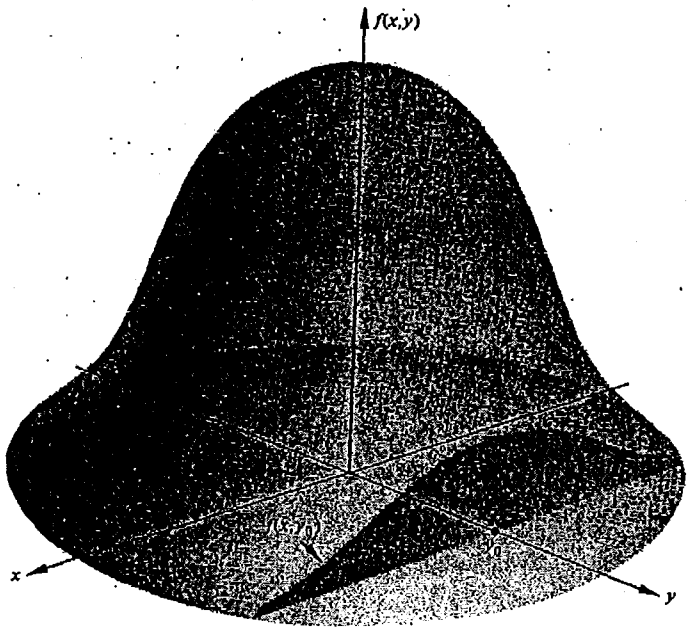

Figure 3.15 The conditional p.d.f. $g_1(x|y_0)$ is proportional to $f(x, y_0)$.

$$g_1(x|y) = \frac{f(x,y)}{f_2(y)} \qquad f(x,y_0) = f_2(y_0)\cdot g_1(x|y_0)$$

Similarly, for any value
given that $X = x$ is defined a

$$g_2(y|x) = \frac{f(x, y)}{f_1(x)}$$

This equation is identical to

Example 2: Calculating a C
joint p.d.f. of X and Y is a
determine the conditional p.
probabilities for Y given the

The set S for which $f(x$
marginal p.d.f. f_1 was derive
It can be seen from Fig. 3.12
Therefore, for any given val
conditional p.d.f. $g_2(y|x)$ of

$$g_2(y|x) = \begin{cases} \frac{2y}{1-x^4} \\ 0 \end{cases}$$

In particular, if it is know

$$\Pr\left(Y \geq \frac{3}{4} \,\middle|\, X = \frac{1}{2}\right) = \int_3$$

Construction of the Joint D

Basic Relations. It follows
$f_2(y) > 0$ and for any value c

$$f(x, y) = g_1(x|y)f_2(y).$$

Furthermore, if $f_2(y_0) = 0$ fo
of generality that $f(x, y_0) =$
(6) will be 0 and the fact that
Eq. (6) will be satisfied for *al*

Similarly, it follows fron
represented as follows for all

$$f(x, y) = f_1(x)g_2(y|x)$$

Example 3: Choosing Points
is chosen from a uniform di

you. There is one issue for the conditional density that didn't arise for the marginal density, and that is for the definition of the conditional density $f(y|x)$ we divide by the marginal density $f(x)$, and in order to avoid dividing by 0, we need to define the conditional density to be 0 over the parts of the domain of X where $f(x) = 0$.

Finally, given the definition and the concept of the conditional density, we can now define the very important concept of statistical independence: The random variables Y and X are said to be independent if the conditional density $f(y|x)$ does not depend on x. Intuitively what this means given our earlier example of Y = Income and X = Education is that if we compare the conditional income distribution at say $X = 12$ (high school graduates) versus $X = 16$ (college graduates) the distributions are *exactly* the same. Notice this is much stronger than just saying the means, variances, higher order moments, etc of the distributions are the same, it is saying *all* moments and features of the distribution are the same. Another way to think of the concept of independence is that giving a researcher knowledge of the value of the conditioning variable x does not help them one bit in saying something about the distribution of the other variable y. Now without independence we can use the definition of the conditional density to always write $f(x, y) = f(y|x)f(x)$, which is to say the joint density is always the product of the conditional density times the marginal distribution of the conditioning variable.[1] In the case of independence, we have for the conditional density that $f(y|x) = f(y)$, and so plugging this back in to the equation before, we have that in the case of independence, $f(x, y) = f(y)f(x)$, which is to say in the case of independence the joint density *factors* into two pieces, one dependent only on y the other piece only dependent on x. Finally, sometimes you are asked to show that two random variables are independent. Given knowledge of their joint distribution $f(x, y)$, the easiest way to do this is to do the calculus to compute the conditional density $f(y|x)$ (or $f(x|y)$ - clearly it doesn't matter which we work with if the variables are independent) and show that it is solely a function of y and doesn't depend on the conditioning variable x at all.

Now that we have described joint (which is to say bivariate, since we are dealing with the simplest case of just two random variables) distributions and the associated marginal and conditional distributions we can define some population moments analogous to the univariate case that we discussed above. Obtaining the population means in the joint distribution case is completely analogous to the univariate case, and so I leave this to you as a good exercise to work through to make sure you see how everything works - the same comment holds for

[1] In fact, this relationship of the joint density to the conditional and marginal densities has empirical implications in and of itself. DiNardo, Fortin, and Lemieux (1996) in a paper in *Econometrica* use this relationship to ask how much of the changes in the distribution of wages in the US over the 1980s was due to worker and labor market characteristics. To do this they exploit the fact that the joint density of y = wages, and x = characteristics in some initial time period, call it 0 for convenience, can be written as $f(y_0, x_0) = f(y_0|x_0)f(x_0)$. Now if we make the *assumption* that the functional form of the *conditional* density does not change from time period 0 to time period 1, then since the marginal distribution of charateristics is empirically observable in the later time period 1 $f(x_1)$, we can construct a *counterfactual* joint density of what the wage distribution in period 0 would have looked like *if* the workers had the characteristics in time period 1 by using: $f(y_0, x_1) = f(y_0|x_1)f(x_1)$. By integrating over this new joint distribution, we can recover the counterfactual marginal distribution of wages in the earlier period under the assumption the workers have the characteristics of workers in the later period. Mor recently, Altonji, Bharadwaj, and Lange (2008) exploit this 'prediction' idea to give an idea of what wages will look like for the cohort of workers soon to enter the labor market. This footnote is just the intuitive idea - see the papers to get all of the details.

the variances of Y and X, where the derivation ends up being much the same as in the univariate case. The moment that is new and that we didn't have in the univariate case but which is important for bivariate relationships is the covariance which intuitively describes how movements in one random variable, say X relate to movements in the other random variable, Y - which is to say how X and Y co-vary with each other. The covariance is no longer purely a measure of spread, as was the case with the variance, but instead measures how deviations in X from its mean relate to deviations in Y from its mean. This idea is reflected directly in the definition of the covariance:

$$Cov(X,Y) \equiv \sigma_{xy} \equiv E\left\{[X - E(X)]\,[Y - E(Y)]\right\} = E(XY) - E(X)E(Y) \tag{7}$$

Notice that the definition of the covariance generalizes (and thus includes) the definition of the variance given earlier, in that if we co-vary X with itself, we get: $Cov(X,X) \equiv \sigma_{xx} \equiv E\left\{[X - E(X)]\,[X - E(X)]\right\} = Var(x) = \sigma_x^2$. Notice also that if we expand the product in the definition of the covariance:

$$Cov(X,Y) \equiv E\left\{X\,[Y - E(Y)]\right\} - E\left\{E(X)\,[Y - E(Y)]\right\} = E\left\{X\,[Y - E(Y)]\right\} \tag{8}$$

simply because $E\left\{E(X)\,[Y - E(Y)]\right\} = 0$ (and of course the same holds true for the 'symmetric' version of this $E\left\{E(Y)\,[X - E(X)]\right\} = 0$), because $E(X)$ is non-stochastic and thus pulls out of the expectations operator, and $E\,[Y - E(Y)] = 0$ since the average deviation of a random variable from its population mean is 0 - by definition of the population mean. Thus, in working with simplifying expressions involving the covariance, this is a helpful 'trick' to remember that we only need to deviate one of the random variable from its mean, not both, and this helps in cutting down on the number of 'cross terms': $Cov(X,Y) = E\left\{X\,[Y - E(Y)]\right\} = E\left\{Y\,[X - E(X)]\right\}$.

Now just like the univariate variance measure was sensitive to the units of measurement of the random variable, the same is true of the covariance, and in some sense this is a more serious problem for the covariance if we want to use it as a 'measure of association' as is frequently the case in applied work. The issue is that we can get as large or as small a covariance that we want simply by changing the units of measurement on Y or X. A similar measure of association that avoids this units-of-measurement problem is the correlation coefficient that simply normalizes or standardizes the covariance by the standard deviations of the two variables - and thus the units of measurement divide out making the correlation coefficient a unit free measure of association! The definition of the correlation coefficient is:

$$Corr(X,Y) \equiv \rho_{xy} \equiv \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{9}$$

Given the definition of the correlation coefficient we can further appeal to a result in statistics known as the Cauchy-Schwarz Inequality which some of you have seen before (in brief, here is the C-S Inequality - for more, see a Probability and Statistics book like that by Morris DeGroot, p. 214: Take two random variables U and V , the Cauchy-Schwarz inequality is that $E(UV)^2 \leqslant E(U^2)E(V^2)$. If we let $U \equiv X - \mu_x$ and $V \equiv Y - \mu_y$, then the

C-S Inequality tells us $\sigma_{xy}^2 \leqslant \sigma_x^2 \sigma_y^2$ or that $\rho_{xy}^2 \leqslant 1$ or finally $-1 \leqslant \rho_{xy} \leqslant 1$). The bounds on the correlation coefficient that $-1 \leqslant \rho_{xy} \leqslant 1$ imply that at the upper bound, $\rho_{xy} = 1$, the random variables X and Y are perfectly dependent, which since the correlation is a *linear* measure of association, that X and Y are just positive linear functions of each other. The same holds for the extreme lower bound $\rho_{xy} = -1$, except that now X and Y are negative linear functions of each other. The case right in the middle of $\rho_{xy} = 0$ means that X and Y are, in a linear sense, unrelated to each other, but it is important to understand that a 0 correlation or covariance does NOT imply no relationship between X and Y in general. For example, we could let $Y = X^2$ and clearly Y is related to X since it is a *function* of X, but if X has a symmetric distribution (meaning its third and all higher-order odd moments are zero) around zero, then the $Corr(X, Y) = 0$ even though X and Y are *functionally* related to each other. Furthermore, while we can show that the concept of independence discussed above *does* imply that independent random variables have zero covariance with each other, the converse does not hold (except in the special case of normal random variables - discussed in future lectures) - a zero correlation or covariance does NOT imply independence - a counter-example is the one I just gave with $Y = X^2$ where conditioning on the value of x clearly does tell us something about the value of y even though the covariance between them is zero.

0.2 Estimation

Thus far we have talked purely about random variables and their population distributions and moments. We have said nothing about data or samples or how to use data to say something about distributions and moments using the data to form estimators of the moments called, appropriately enough, sample moments. Sample moments or estimators are algebraic combinations of data, but it is important to distinguish between the properties of such estimators *before* a sample is drawn, in which case an estimator is just a combination of random variables, and therefore is itself a random variable with a distribution and population moments, etc, and the behavior of an estimator *after* a sample is drawn, in which case the estimator is just a combination of realizations, and so is itself just a realization or number. In short, because estimators are themselves random variables, we never truly estimate the true, say, population mean or population variance, but instead one which on average equals the true parameter. This is the fundamental or inherent uncertainty of statistical work which necessitates no matter how well you do your job as an analyst, there is always uncertainty associated with any estimator, and thus uncertainty regarding any conclusions drawn from our estimates and statistical work. If you can't deal with that, then applied statistics and econometrics isn't really for you, as it is a basic feature of the field. The flip side is to understand we never really get to see 'the truth', and so you need to always bear in mind no matter how good you think your analysis is, the end results and estimates all have uncertainty (or 'sampling error') associated with them, arising from the basic random variables generating the data. One last general comment, it is fairly common at this stage for students to confuse sample and population moments, so work hard at the outset to make sure you don't confuse them, and that you understand how sample moments are designed to say something useful about their population counterparts, in ways we make precise below.

We begin our discussion of estimation as we did in the case of discussing random variables, with the univariate case, where we wish to use a sample of data generated by the random variable X to say something about the population moments of X. The easiest form of a sample to work with is one which we say is *independently and identically distributed* or iid for short. What this means is that each of the sample realizations x_i for $i = 1, ..., N$ (N is the total sample size - i.e. the number of data points) is drawn from a common or identical distribution - that it, it does not change from observation to observation. Furthermore, the *independent* part of iid tells us that if we consider the joint distribution for the *entire* sample $f(x_1, x_2, ..., x_N)$ that it factors into the product of the N marginal densities $f(x_1)g(x_2) \cdots h(x_N)$ because telling us the realization say x_3 tell us nothing about the conditional density for x_4 and so on. Thus, the independence part of iid rules out things like serial correlation, where say knowing GNP per capita in a country in say 2005 is going to be *very* informative about its GNP per capita in 2006, and the identically distributed part rules out things like heteroskedasticity, where say the variance in R&D investment across industries is very likely to be different, and thus the distributions of R&D are likely *not* to be identical. Thus, iid is a simplifying assumption, and often not correct for the setting we are considering, and so while it is useful to start with, eventually we will have to consider what happens when this convenient assumption is simply flat out false.

For now, however, we make the iid assumption, as it makes evaluating the properties of estimators vastly easier. To begin, rather than deriving estimators, we simply consider familiar estimators, and ask or derive the properties of those estimators. For example, the *sample mean* (or *sample average*, as it is known more generally) is something you have known since your early secondary school days:

$$\overline{X} \equiv \frac{1}{N} \sum_{i=1}^{N} x_i \tag{10}$$

where x_i for $i = 1, ..., N$ is simply the N data points on X that we have in our sample. Under the iid assumption we have that the x_i are realizations from a common (or stable) distribution with common population mean denoted by μ_x and common population variance denoted by σ_x^2. Now the most basic feature of the random variable we can ask about is to ask what is $E(\overline{X})$, which in words is a bit of a mouth-full: what is the population mean of the sample mean? Now I am not one to give 'cookbook' advice, as in, "you should always do such and such", but in evaluating properties of estimators, there is in fact a two-step method you should always follow. First, write down the definition of the estimator in terms of the data - we already have that above. Second, plug in from the 'data generating process' or 'true model' the underlying parameters. In this case, the only data generating process is that the data points x_i come from a common distribution for the random variable X. Finally, the other aspect we use are the properties of the expectations and variance operators, as needs be. Thus, we are interested in:

$$E(\overline{X}) = \frac{1}{N} E\left[\sum_{i=1}^{N} x_i\right] = \frac{1}{N} \sum_{i=1}^{N} E\left(x_i\right) = \frac{1}{N} \sum_{i=1}^{N} \mu_x = \mu_x \tag{11}$$

where the second equality follows from the linearity of the expectations operator, so

it can pass through the summation operator (an expectation of a sum is always a sum of the expectations). The third equality followed from the 'identically distributed' part of iid, and in fact we see that only the assumption of a common mean - **not** a common variance - is required to establish the result. The next equality is just algebra, and it establishes that $E(\overline{X}) = \mu_x$, which says that the population mean of the sample mean is the population mean. In short, 'on average' $\overline{X}$ is equal to its population counterpart μ_x, even though in any *particular* sample we basically never have that $\overline{X}$ equals μ_x. We call this property that of *unbiasedness* when the expectation of some estimator is equal to its population counterpart. It is important to note that we did not require the full assumption of *iid* to establish the unbiasedness of the sample mean for its population counterpart the population mean - for example, σ_x^2 could vary across observations (and thus the assumption of 'identically distributed' fails) and the sample mean would still be unbiased - the same conclusion holds if the x_i observations are not independent, as the expectation of the sum being the sum of the expectations did not require independence of the sample observations.

The notion of unbiasedness is a general property of an estimator, and to talk in general about defining such properties, it is helpful to talk about some generic estimator, which we will denote as $\widehat{\theta}$, the 'hat' is common notation for an estimator of some true population parameter, which in our generic notation will be θ. With this generic notation in hand, we can now define the general notion of the *bias of an estimator* as:

$$Bias(\widehat{\theta}) \equiv E(\widehat{\theta}) - \theta \tag{12}$$

and it is important to recognize that an *unbiased* estimator has a zero bias *regardless of the true value of the population parameter* θ! In other words, bias is a property that describes an estimator regardless of the true value of the parameter. There are situations in applied settings where we may have come clue as to what the true value of θ is, or we can rule out some values of θ as being implausible, but that generally takes us into the realm of Bayesian statistics, where we formally incorporate that prior information on the parameter θ in the estimation and inference procedures. We don't take that approach in this course, but we do sometimes consider situations where we can do 'better' in settings where we have some extra or prior information on θ - like say we believe θ is equal or near 0 - although any new estimator that is designed to include that prior information is no longer *unbiased* because, as I emphasized above, and unbiased estimator has zero bias regardless of the true parameter value. In short, if you really do have strong extra or prior information, you probably want to depart from the class of unbiased estimators and consider some of these alternative approaches. Finally, while this is a redundant comment in light of all that I just said, the computation of the bias does not require knowledge of the true θ, which is something we never have or know anyway, but instead the bias is a function of the parameter θ, regardless of whatever value it takes on.

Another common estimator that you may also have encountered when you first learned about the sample average or the sample mean is the *sample variance*:

$$S_x^2 \equiv \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{X})^2 \tag{13}$$

Notice the similarity of the expression for the sample variance to the expression for the population variance that we discussed above:

$$Var(X) \equiv E\left\{[X - E(X)]^2\right\} \tag{14}$$

The similarity of replacing population expectations with sample averages - which one caveat that I will discuss in the next sentence - is something the economist/statistician Charles Manski has called 'the analogy principle', and it is a good point of departure for thinking of ways to construct estimators for population quantities. The caveat is that a pure average of the terms $(x_i - \overline{X})^2$ would divide by N as opposed to the divisor $N-1$ used in the definition of the sample variance S_x^2. The intuitive reason for subtracting 1 from the overall sample size N is that the *sample mean* $\overline{X}$ is used in the definition of S_x^2 as compared to the *population mean* $E(X)$ that is used in the definition of the population variance. In essence, having to *estimate* $E(X)$ by using $\overline{X}$ 'burns up' 1 "degree of freedom", or dof for short, from the overall degrees of freedom or sample size N. Put a little more technically, of the overall N terms involved in the sum $\sum_{i=1}^{N}(x_i - \overline{X})^2$ in the definition of S_x^2, only $N-1$ of them are independent if the x_i themselves are independent. Finally, another way to justify dividing by $N-1$ as opposed to N is to just verify that using $N-1$ leads to an unbiased estimator. Remember my advice from above, the first step in evaluating $E\left(S_x^2\right)$ is to substitute in for the definition of the estimator:

$$E\left(S_x^2\right) \equiv E\left\{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{X})^2\right\} \tag{15}$$

$$= \frac{1}{N-1}\sum_{i=1}^{N}E(x_i - \overline{X})^2 \tag{16}$$

due to the linearity of the expectations operator. For the next steps, it pays to do a little algebra by adding and subtracting the unknown population mean μ:

$$= \frac{1}{N-1}\sum_{i=1}^{N}E\left[(x_i - \mu) - (\overline{X} - \mu)\right]^2 \tag{17}$$

Now expand the square:

$$= \frac{1}{N-1}\sum_{i=1}^{N}\left\{E\left[(x_i - \mu)^2 + (\overline{X} - \mu)^2 - 2(x_i - \mu)(\overline{X} - \mu)\right]\right\} \tag{18}$$

now apply the summation operator to each of the three terms:

$$= \frac{1}{N-1}\left\{\left[\sum_{i=1}^{N}E(x_i - \mu)^2\right] + NE(\overline{X} - \mu)^2 - 2E\left[(\overline{X} - \mu)\sum_{i=1}^{N}(x_i - \mu)\right]\right\} \tag{19}$$

and the last term simplifies to:

$$= \frac{1}{N-1}\left\{\left[\sum_{i=1}^{N} E(x_i-\mu)^2\right] + NE(\overline{X}-\mu)^2 - 2NE(\overline{X}-\mu)^2\right\} \tag{20}$$

so we can bring the last two terms, multiplied by N, back under the summation operator to get:

$$= \frac{1}{N-1}\sum_{i=1}^{N}\left\{E(x_i-\mu)^2 - E(\overline{X}-\mu)^2\right\} \tag{21}$$

Now by the definition of the population variance, $E(x_i-\mu)^2 \equiv \sigma_x^2$, and by the definition of the variance of $\overline{X}$, i.e. $Var(\overline{X})$, we know $Var(\overline{X}) \equiv E(\overline{X}-\mu)^2$ (in words, this is the *population variance of the sample mean*). Now, as an aside, let's derive $Var(\overline{X})$:

$$Var(\overline{X}) \equiv Var\left\{\frac{1}{N}\sum_{i=1}^{N} x_i\right\} \tag{22}$$

and now bear in mind our rules about using the Variance operator, and the need to square constants brought outside the variance operator, etc:

$$Var(\overline{X}) \equiv \frac{1}{N^2} Var\left\{\sum_{i=1}^{N} x_i\right\} \tag{23}$$

$$= \frac{1}{N^2}\left\{\sum_{i=1}^{N} Var(x_i) + 2\sum_{i=1}^{N}\sum_{j>i}^{N} Cov(x_i,x_j)\right\} \tag{24}$$

Now if we invoke the iid assumption for the sample on x_i, then we have $Var(x_i) = \sigma_x^2 \; \forall$ i, and $Cov(x_i,x_j) = 0 \; \forall \; i \neq j$. Thus, under the iid assumption we can conclude:

$$Var(\overline{X}) \equiv \frac{\sigma_x^2}{N} \tag{25}$$

a result you may have seen in earlier courses. It shows that the inherent sampling variability in the sample mean shrinks as N, the sample size, grows. Intuitively, this just means that not only is the sample mean unbiased for the population mean, but as the sample size grows, it becomes an increasingly precise estimate of the population mean. The concept of an estimator having a vanishingly small inherent variability as the sample size increases is an important property, one that relates to the concept of consistency that I define in the last section of these notes.

At the moment, using these two results, we can plug them back into our earlier expression for $E\left(S_x^2\right)$ to get:

$$= \frac{1}{N-1}\sum_{i=1}^{N}\left\{\sigma_x^2 - \frac{\sigma_x^2}{N}\right\} \tag{26}$$

or

$$E\left(S_x^2\right) = \frac{1}{N-1}\sum_{i=1}^{N}\left\{\frac{(N-1)\sigma_x^2}{N}\right\} = \sigma_x^2 \tag{27}$$

In other words, the estimator of the sample variance we proposed, $S_x^2 \equiv \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{X})^2$ is unbiased, and thus this also helps explain why we divide by $N-1$, the appropriate degrees of freedom since we use the sample estimator $\overline{X}$ in place of the unknown μ_x, as opposed to N. Had we divided by N, the estimator would have been biased, although the bias would go to zero as the sample size N became asymptotically large.

Similar arguments establish that the estimator for the *sample covariance* $\widehat{\sigma}_{xy}$ which is unbiased is:

$$\widehat{\sigma}_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{X})(y_i - \overline{Y}) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{X})y_i = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{Y})x_i \tag{28}$$

where, just as was the case with the population covariance, the last two equalities demonstrate that only one of the two variables involved in a sample covariance needs to be deviated from its sample mean, and the resulting sample covariance is unchanged.

With the estimators for sample means, sample variances and sample covariances, we are in a position to start to consider hypotheses about the distributions and processes generating the data - an overall set of tasks called *statistical inference* (or just *inference*) - and we do that in the next lecture. For now, we consider one additional, and very much related concept to the notions of bias and variance of estimators called *Mean Squared Error* or MSE for short. MSE takes account of the fact that we may not always want to place ourselves purely in search of just unbiased estimators, since we may achieve unbiasedness only by resorting to estimators with large inherent variabilities (this is indeed a thorny problem in applied econometrics when people consider IV estimators which may be unbiased, but which may also have very large variances, making drawing inferences very difficult). MSE puts bias and variance on the same footing by squaring the bias (so as to penalize the *magnitude* of the bias and not its actual *value* - otherwise a negative bias would offset a necessarily positive variance for what would look like a very small MSE) and adding it to the variance:

$$MSE(\widehat{\theta}) \equiv \left[Bias(\widehat{\theta})\right]^2 + Var(\widehat{\theta}) = E\left[\left(\widehat{\theta} - \theta\right)^2\right] \tag{29}$$

Using the MSE criterion, we may be willing to accept a biased estimator if it has a small enough variance so as to dominate an unbiased estimator in an MSE sense. Furthermore, even for unbiased estimators, it is useful to verify the MSE (which is just the $Var(\widehat{\theta})$ if the estimator is unbiased) goes to zero as the sample size grows arbitrarily large. It turns out this is one way to establish what is called 'consistency', a concept we define in the next lecture, but which means the estimator lies in a vanishingly small neighborhood of the true parameter value as the sample size gets asymptotically large. For example, $MSE(\overline{X}) = \frac{\sigma_x^2}{N}$, and so we see the sample average is not only unbiased, but it also a consistent estimator of

the population mean. By contrast, taking say the third observation of a sample, x_3, which has the virtue of being unbiased for the mean, has $MSE(x_3) = \sigma_x^2$, which clearly does not go to zero as the sample size gets arbitrarily large. The simple reason for that is that only using the third observation of a sample of size N as the sample size grows clearly does not make very efficient use of the total sample.

MSE also allows us to consider estimators which, while biased, may be better under certain circumstances. For example, consider the estimator $\widehat{\theta} = c\overline{X}$ as an estimator of the population mean μ_x, where c is some constant chosen by the econometrician. Then, $Bias(\widehat{\theta}) = (c-1)\mu_x$, and so $\left[Bias(\widehat{\theta})\right]^2 = (c-1)^2\mu_x^2$, and $Var(\widehat{\theta}) = c^2\frac{\sigma_x^2}{N}$, so putting this all together, we have, $MSE(\widehat{\theta}) = (c-1)^2\mu_x^2 + c^2\frac{\sigma_x^2}{N}$ and we want to compare this to $MSE(\overline{X}) = \frac{\sigma_x^2}{N}$, and so

$$MSE(\widehat{\theta}) < MSE(\overline{X}) \text{ when } (c-1)^2\mu_x^2 + c^2\frac{\sigma_x^2}{N} < \frac{\sigma_x^2}{N} \tag{30}$$

or

$$(c-1)^2\mu_x^2 + (c^2-1)\frac{\sigma_x^2}{N} < 0 \tag{31}$$

In the case where $\mid c \mid < 1$, then $(c-1) < 0$, and so if we divide the above expression by $(c-1)$ we need to flip the inequality sign to get:

$$(c-1)\mu_x^2 + (c+1)\frac{\sigma_x^2}{N} > 0 \tag{32}$$

or

$$(1+c)\frac{\sigma_x^2}{N} > (1-c)\mu_x^2 \tag{33}$$

In general this is difficult to verify, but we see in the particular case where $\mu_x = 0$, even though the estimator $\widehat{\theta}$ is biased, as I emphasized above, the MSE of $\widehat{\theta}$ is smaller than that of $\overline{X}$, simply because multiplying $\overline{X}$ by a constant $\mid c \mid < 1$ has less of an impact on the bias than it does on shrinking the variance of $\widehat{\theta}$ towards 0 in a manner that depends on c^2. In fact, that is why this type of estimator is called a *shrinkage* estimator in the statistics literature, because it takes advantage of the fact that while the impact on the bias is only linear in c, the effect on the variance is quadratic. In terms of MSE, obviously the above expression is more likely to hold the closer is μ_x to 0.

Even if we don't want to work formally with the MSE - applied people will often just talk in non-analytical terms about the 'bias variance tradeoff' - it does help remind us that restricting ourselves to only unbiased estimators can be unduly restrictive, especially if we end up with estimators with large inherent variances as a result. Sometimes it is worth giving up a small amount of bias if it means we can gain in precision as a result, and that is the tradeoff in choosing among estimators that the MSE criterion summarizes.

In the next sub-section, I also use the *asymptotic* behavior of the MSE of an estimator to say something about its *consistency* (both of these terms are defined in the next set of lecture notes). So-called 'convergence in mean square' is a sufficient condition for the concept of

consistency that I discuss next, and convergence in mean square is defined to mean that the MSE goes to zero as the sample size goes to infinity. The intuitive idea is that, for our purposes, a consistent/convergence in MSE estimator is one that has a distribution that 'collapses' around its true value as the sample size grows infinitely large. I mention this here, since this will also be a way in which you will use the MSE criterion in problem sets and exams.

0.3 Large Sample Properties of Estimators

[For background, see Greene, pages 126-127 or Engle's Chapter in the *Handbook of Econometrics*, or Lindgren's book *Statistical Theory* pages 280 to 281. For a textbook approach relevant for this class, see Chapter 2 of Johnston and DiNardo, especially pages 53-56.]

While we prefer to consider the 'finite sample' behavior of estimators based on the Expectations and Variance operators, often even the simplest of estimators in econometrics involve ratios, squares etc of random variables, and as we discussed in the first set of lecture notes, the expectations operator does not pass through functions like say ratios of random variables. To make headway, we instead turn to ascertaining the large sample properties, since the operator involved there (to be defined next) *does* pass through non-linear functions of random variables. We call this approach the 'asymptotic behavior' or 'asymptotic approximation' behavior, since we use the asymptotic behavior that we solve for to approximate the behavior of the estimators we work with in our finite samples.

Definition: *Consistency* - Let $\widehat{\theta}$ denote a generic estimator of some population moment or parameter θ, then we say $\widehat{\theta}$ is a consistent estimator of θ if $\lim_{N\to\infty}$ Prob $\left\{\left|\widehat{\theta}-\theta\right| > \varepsilon\right\} = 0$ $\forall \ \ \varepsilon > 0$. Thus 'consistency' (and in 558, I am defining this via the Weak Law of Large Numbers - in more advanced courses, you will spend a lot of time distinguishing among various notions of consistency and converge) is intuitively like a version of unbiasedness, in that as the sample size gets large, the bias in $\widehat{\theta}$ goes to zero. However, this is only intuition, and while it is essentially correct intuition, to be precise consistency does not imply unbiasedness, nor does unbiasedness imply consistency. As an example of an unbiased estimator that is not consistent, consider using the third observation in any *iid* sample of data, since the sample is *iid*, $E(\widehat{\theta}) = E(X_3) = \mu$, and so this estimator is unbiased. But this estimator is not consistent because its variance remains at σ^2 as the sample size goes to infinity, and so it remains possible, no matter how large the sample, to find an ε such that the consistency definition is not satisfied. Also note we can define an estimator which, while consistent, is not unbiased (though this example is a bit contrived): construct the estimator $\widehat{\theta}$ such that Prob$(\widehat{\theta} = \theta) = \frac{N-1}{N}$ and such that Prob$(\widehat{\theta} = N) = \frac{1}{N}$. You can verify that Prob$(\widehat{\theta} = \theta) \to 1$ as $N \to \infty$, and so the estimator is consistent, but as $E(\widehat{\theta}) = \frac{N-1}{N}\theta + 1$, the estimator is biased, even as $N \to \infty$. While this example is contrived, in that if we actually *knew* θ then there would be no need to estimate it, the two examples together serve to show:

$$\text{Unbiasedness} \quad \not\Rightarrow \quad \text{Consistency} \tag{34}$$

$$\text{Consistency} \quad \not\Rightarrow \quad \text{Unbiasedness} \tag{35}$$

For the purposes of this class, however, it is *sufficient* to think of consistency as the joint requirement of unbiasedness, or at least asymptotic unbiasedness, along with the variance of the estimator going to zero as the sample size goes to infinity. This notion is properly referred to as 'Convergence in Mean Square', since it implies the Mean Squared Error (MSE) goes to zero as the sample size gets infinitely large, as opposed to just 'consistency' although, like I said, for our purposes we can treat these terms as equivalent.

In terms of notation, if the estimator is consistent, we denote this as: plim $\widehat{\theta} = \theta$, where the notation 'plim' is an abbreviation for 'probability limit' and it is implicitly understood that the limit is the limit as $N \to \infty$. So the particularly useful feature of the plim operator as opposed to the expectation operator is that if $g(\cdot)$ is some non-linear function that is continuous at plim $\widehat{\theta}$, then plim $g(\widehat{\theta}) = g($ plim $\widehat{\theta})$ - i.e. the plim of a function of an estimator is the function of the plim of the estimator, so the plim 'passes through' the non-linear function $g(\cdot)$. This general result is known as Slutsky's Theorem. In addition, we can say that plim $(\widehat{\theta}_1\widehat{\theta}_2) =$ plim $\widehat{\theta}_1$ plim $\widehat{\theta}_2$ irrespective of whether $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are independently distributed. Finally, we we mention above, plim $\frac{\widehat{\theta}_1}{\widehat{\theta}_2} = \frac{\text{plim } \widehat{\theta}_1}{\text{plim } \widehat{\theta}_2}$ provided plim $\widehat{\theta}_2 \neq 0$.

Finally, while consistency tells us something about the limiting value of an estimator, a central limit theorem (CLT) tells us about the limiting *distribution*, although usually this stronger condition requires that we say something about limiting the degree of dependence across observations, as well as requiring that the observations are drawn from an identically distributed random variable. For example, for an *iid* sample, a simple central limit theorem tells us that $\overline{X}$ is distributed asymptotically as $N(\mu, \frac{\sigma^2}{N})$, though a simple glance at a book like Halbert White's *Asymptotic Theory for Econometricians* will reveal that such results hold for non-*iid* settings as well. The main point is that a CLT tells you not only that plim $\overline{X} = \mu$, but it also tells you its asymptotic variance, as well as the distribution that it follows. Thus while we often use plim results as a point of departure for the asymptotic behavior of estimators, we ultimately resort to CLTs to obtain the full distributional aspects that the estimator follows, at least in large samples. Finally, as I discuss later in the course, since the variance of $\overline{X}$ goes to zero asymptotically, we usually instead discuss the limiting distribution of $\sqrt{N}\left(\overline{X} - \mu\right) \sim N(0, \sigma^2)$ so that the asymptotic variance does not go to zero.