Some Stats Concepts you should know

If you understand how to do hypothesis testing and how to interpret regression coefficients, you are well-prepared to take econometrics. However, we will even review these topics in econometrics class, if only quickly. Below are some more basic statistics concepts that you should understand before proceeding in this course. Many of these will be useful in the first problem set.

- Population
  - The idea of a "population" is a central concept in statistics. Its characteristics are usually considered unknown, as it may be too costly (or in some cases impossible)[1] to find them out for certain. Instead, we will attempt to use samples, along with statistical techniques, to estimate its characteristics.
- Sample
  - Simple random sample: each member of the population has an equal probability, *a priori*, of being in the sample.
  - Stratified random sample: within defined subgroups (e.g., by gender or race) a random sample is taken, but subgroups may not be sampled at the rate at which the groups are represented in the population. Used, e.g., to get larger samples of rare subgroups that are of interest to the survey takers.
- Dataset types – are distinguished by what constitutes the "unit of observation." See chapter 1 of Wooldridge textbook for more extended discussion of data types.
  - Time series data – unit of observation is a time period (e.g., quarterly reports of U.S. GDP; annual sales figures for a single company)
  - Cross section data – unit of observation is an "entity" (person, company). The data contain information on each entity from roughly the same time (e.g., the 1990 U.S. Census of Population.)
  - Pooled cross-section data– combined cross-section data from several years. Distinguished from panel data by the fact that the "entities" need not be exactly the same in the different years. For example, a collection of telephone surveys of people's political preferences taken each year, where different people were surveyed each year but the same questions were asked, would be a "pooled cross-section."
  - Panel data – the unit of observation is usually an entity x year. Roughly the same variables on the same entities are obtained in several years. For example, the National Longitudinal Survey of Youth, which we will use for a research project in this class, interviewed kids aged 14-19 in 1979 and followed up with additional surveys every year or 1-2 years since then.

[1] Why impossible? Some populations are infinite (the set of all possible coin tosses) and some are unobservable (such as the "counterfactual" outcome in the Rubin Causal model).

- Random Variable - numerical outcome of a random trial; examples: number from the roll of a di, survey responses from a randomly drawn sample, and estimators.
    - Discrete -- takes on a finite number of values (as in the roll of a di)
    - Continuous – can take on any value over a specified range. In practice what we often loosely refer to variables as "continuous" if they take on a large number of values. For example, we often call years of education a continuous variable even though it typically takes on only roughly 20 values.
- Estimator - a procedure for estimating a population parameter or value with a sample. (For example, "add up all the data and divide by the number of observations," = $\bar{x}$, the sample mean). Estimators are considered a random variables because (or in contexts where) the sample is randomly chosen. For example, the value we get for a sample average will depend on who happens to be chosen for our sample, analogous to randomness of the outcome of rolling a di.
    - Estimate - a number obtained from applying an estimator to a particular sample. This is NOT random.
    - Standard Error - the estimated standard deviation (see below) of an estimator. It measures how much we expect the estimate would vary from one similarly taken sample to another.
        - You may recall that the standard error of the sample mean, $\bar{x}$, is $s/\sqrt{N}$, where s is the standard deviation in the data and N is the sample size.
        - So for example, in a survey of 2,500 consumers on annual holiday spending with a mean of \$750 and a standard deviation of \$1200, the standard error on the mean would be \$1200/50 = \$24. This means in repeated random samples of 2,500 consumers, we would expect $\bar{x}$ to vary by about \$24 from one sample to another.
- Expected value - E[x] = population average
    - Recall that expected value is a linear operator; that is, for an random variables X and Y, and for any numbers a, b, and c, E[aX + bY + c] = aE[X] + bE[Y] + c. For example if the mean income in some population is \$50,000, and you give everybody \$1000 (= c) the mean income becomes \$51,000. If you then convert everyone's income to Euros at 1.00€/\$1.33 (=a) , mean income becomes 51,000/1.33 = 38,476€
        - Note that c is just a number -- it is not random -- so its expected value is the number itself.
- Unbiased - said of estimators: when the expected value of the sample estimator is equal to the population parameter. The sample mean is unbiased $E[\bar{x}]=\mu$, where μ the population average, if the data come from a random sample.
    - Upward biased: if the expected value of the estimator is above the true value of the population parameter, e.g., $E[\bar{x}]>\mu$. Downward biased is the opposite.

- Variance: the expected squared deviation of a random variable from its mean. If the population mean of a random variable X is μ, the variance is defined as $E[(X-\mu)^2]$.
  - The sample analog of the variance in a dataset, also known as an estimator (see above) of the variance, replaces μ with the sample average, and adds up the data and divides by N-1: $\frac{1}{N-1}\sum_{i=1}^{N}(X_i-\bar{X})^2$, where N is again the sample size and *i* indexes the observations of the dataset. This also happens to be an unbiased (see above) estimator of the variance (whereas dividing by N instead of N-1 produces a downward biased estimator).
  - It is common to denote the population variance using σ^2, where σ is the lowercase Greek letter sigma, which represents the standard deviation
  - Standard deviation -- square root of the variance. This also measures how much variation there is in the data, but it is more useful because it is measured in units of the original data (as opposed to squared units with the variance). The standard deviation of a population is often denoted σ.
  - Note that the common usage of "the standard deviation" refers to the sample estimator concept in a particular dataset. But in statistics and econometrics there is also a population concept defined in terms of random variables. This is why it is meaningful to talk about things like "the standard deviation of an estimator" even though in practice we only typically have one sample. [2]
- Covariance a general measure of relatedness of two random variables analogous to the variance. If the population mean of a random variable X is μ_X and Y is μ_Y, then the covariance is defined as $E[(X-\mu_X)(Y-\mu_Y)]$.
  - The sample estimator of the covariance in a dataset replaces the μ's with the sample averages, and adds up the data and divides by N-1: $\frac{1}{N-1}\sum_{i=1}^{N}(X_i-\bar{X})(Y_i-\bar{Y})$.
  - The covariance does not have very meaningful units, and its magnitude is hard to interpret. But the sign tells you whether X and Y are positively or negatively related.
- You may recall that the correlation, a number between -1 and +1, standardizes the covariance by dividing by the standard deviation of each variable. It measures the strength, but not the magnitude (slope) of any linear relationship between X and Y.[3] It has no units, and is typically denoted with an "r" or "R" if it is a sample estimate, and a ρ when it is a population concept.

---

[2] Note that when we do so, we are considering the situation before we have collected the sample; X_i represents what we might get from a random draw from the population, not the actual data.

[3] The linear distinction is important: in extreme cases two variables could even be perfectly related but have zero correlation (if that relationship was nonlinear)! (For example, if $y = x^2$, y and x would be perfectly related. However, y and x would have zero correlation: a straight slope fitted between y and x would have zero slope. Draw a picture to see why.)

  - In a bivariate (one Y, one X) linear regression **only** R^2 is literally the squared correlation between Y and X. In a multivariate regression this interpretation does not hold.
- In general, for random variables X and Y and numbers a, b, and c, $Var(aX + bY + c) = a^2Var(X) + b^2Var(Y) + 2abCov(X,Y)$. Note that:
  - Since c is just a number, it does not affect the variance. If you gave everybody in the room $5, it would raise mean wealth in the room, but not affect the variation in wealth.
  - If X and Y are independent, as is assumed to be for different observations in a random sample (in a simple random sample, the data are "independent and identically distributed" or "iid") then the covariance term disappears, so $Var(aX + bY + c) = a^2Var(X) + b^2Var(Y)$
  - Why are the constants a and b squared? Recall that the variance is the expected value of the squared deviations. So if you multiply all the data by 2, the variance goes up by a factor of four, not 2. The standard deviation goes up by a factor of 2.
- The covariance of a linear combinations of random variables:
  - You probably don't need to know this, but FYI, $Cov(aX_1 + bX_2, cY_1+dY_2) = acCov(X_1, Y_1) + adCov(X_1, Y_2) + bcCov(X_2, Y_1) + bdCov(X_2, Y_2)$ for random variables X_1, X_2, Y_1, Y_2 and constants a, b, c, d. This can be derived from the definition of covariance
  - Also, note from the definition of covariance that the covariance of a variable with itself is the variance: $Cov(X,X) = Var(X)$.
- Standardizing a random variable means subtracting off the mean and dividing by the standard deviation; if X has a mean of μ and a standard deviation of σ, then standardized X is $\frac{X-\mu}{\sigma}$.
  - This results in a new random variable which has a mean of zero and a standard deviation of 1.
- Cumulative Density Function (or distribution function in the case of a discrete random variable) or "CDF" -- measures the probability that a random variable takes on a value below a specified value. Often denoted with a capital letter function and a lowercase argument, as in G(x). Note that the argument is a number, not a random variable. For random variable X, $G(x) = Pr(X<x)$.
  - Probability Density Function (distribution in the case of a discrete random variable) or "PDF." The derivative of the CDF. Note in the case of a continuous random variable, the PDF does not measure a probability, since the probability that a continuous random variable takes on any particular value is zero.

  - The CDF probabilities associated with a standard -- mean zero, variance one -- normal random variable are shown in Table G.1. on page 831 of the text.
- Any linear transformation of a normally distributed random variable is also normally distributed. This allows us to transform a variable and look up probabilities that it takes on values in particular ranges using standard tables, such as those in Appendix G on page 831.
  - E.g., if X is normally distributed with mean 3 and variance 4, then Pr(X<-1) = $\Pr\left(\frac{X-3}{2} < \frac{-1-3}{2}\right) = \Pr(Z < -2)$, where Z is (often) used as a symbol for a standard normal random variable. According to the table on page 831, this probability is 0.0228. (Do you see this? How would you instead calculate Pr(X>-1)?)
  - The sum of independent (see below for definition), normally distributed random variables is also normally distributed
- | = "given that" or "conditional on" as in Pr(purple-people eater|one-eye, one-horn) = probability of being a purple people eater given that you have one eye and one horn, or E[drinks last weekend|fraternity member] = expected number of drinks consumed last weekend by a fraternity member.
- Two random variables are <u>independent</u> if the probability that one takes on any particular value is unrelated to the value the other variable takes on.[4]
  - Observations in a simple random sample are independent. If I survey people at random, the answers one person gives to questions will be on average unrelated to the other respondents' answers.
  - In linear regression, we are often talk about a weaker condition, so-called "mean independence": E[u|X] = 0. This condition says the expected value of random variable u (which is this example is zero) is unrelated to the value random variable X takes on.
  - If two variables are independent or even just mean independent, they also have zero correlation and zero covariance.

<u>For further reference, see also:</u>

- Wooldridge Textbook Chapter 1 and Appendices A, B, and C
- Materials from my stats class – especially problems and tests (available on Canvas)

[4] Technically, the condition is written as $g_{xy}(X,Y) = g_x(x).g_y(y)$ where $g_{xy}(X,Y)$ is the joint PDF – integrated over a range, it gives the joint probability that X is in the specified range at the same time as Y is in the specified range – and $g_x(x)$ and $g_y(y)$ are the PDFs of X and Y, respectively (technically called the "marginal" PDFs).