EC325 Poverty and Discrimination
What You Should Know and Understand about Statistics for this Class

Students come to this class with very different statistical backgrounds. This document is intended to give you the necessary tools to get up to speed on what statistical significance really mean and how to interpret a t-statistic and regression coefficients. I have chosen key videos from Khan Academy and created some videos in a similar style (although much lower production value) where I could not find any. You may choose to follow all the videos or just fill in gaps.

**Random variables** - The really key point is that many things that you would not think of as random can be thought of as random from a statistical perspective.

https://www.khanacademy.org/math/probability/random-variables-topic/random_variables_prob_dist/v/random-variables

**Mean** - (this is a very basic review; I very much doubt anyone needs it)


**Variance** -

https://www.khanacademy.org/math/probability/descriptive-statistics/variance_std_deviation/v/variance-of-a-population

Note: there are multiple Kahn Academy videos explaining how to use the sample variance to estimate the population variance. This is worth knowing but not essential to the course.

**Standard deviation** -


Note: The video above also reviews mean and variance. So unless these concepts are new to you (which I doubt), you can skip the videos on mean and variance.

**Normal distribution** - The really key point, not fully brought out in the video, is that normal distributions are fully described by their mean and standard deviation. 95% of observations are within 1.96 standard deviations from the mean; 90% are within 1.64 standard deviations of the mean. Similarly, we can look up the likelihood that an observation is more than any particular distance from the mean.


Note: There are lots of additional videos that help you explore the normal distribution.

**Law of Large Numbers** - If you sample enough times from a distribution, your sample mean gets very close to the mean of the population from which you are sampling

Central Limit Theorem - If you take mean of a lot of independent random variables, and if you do this over and over again, the distribution of your sample mean will be approximately normally distributed.


Note that even in cases where we know the true distribution cannot be normal, the normal approximation is often very accurate. If we flip a fair coin 100 times, the true mean number of heads is 50. We can show that standard deviation is 5. Using the central limit theorem and therefore the normal approximation, we predict that number of heads will be greater than 40 and less than 60, 95% of the time. Actual number of 94.3%

Standard error of estimate - The Khan Academy video focuses on the standard error of the mean. Many statistics are only estimates and have standard errors that are largely ignored. For example, the poverty rate is almost always reported as a precise number but has a standard error of poverty rate of about .14 for the country as a whole.


Confidence interval -

https://www.khanacademy.org/math/probability/statistics-inferential/confidence-intervals/v/confidence-interval-1
https://www.khanacademy.org/math/probability/statistics-inferential/confidence-intervals/v/confidence-interval-example

Statistical Significance - This is just a expression for unlikely to be due to chance. It does not mean that the relation is causal. *And it does not mean it is important.*

https://www.khanacademy.org/math/probability/statistics-inferential/hypothesis-testing/v/hypothesis-testing-and-p-values

We will almost always use two-tailed tests, but it is useful for you to know about one- v two-tailed tests.

https://www.khanacademy.org/math/probability/statistics-inferential/hypothesis-testing/v/one-tailed-and-two-tailed-tests

`t-statistic` - This is really just to make sure that you are not confused when we talk about t-statistics instead of z-statistics. In most cases we deal with, they will be identical for all practical purposes.

https://www.khanacademy.org/math/probability/statistics-inferential/hypothesis-testing/v/z-statistics-vs--t-statistics

Simple Regression - how to interpret the output from a simple regression showing the relation between two variables. You should understand what coefficients mean and how the concepts of confidence interval, t-statistics and statistical significance apply to them.

http://www.showme.com/sh/?h=LNxjMSu
Controlling for other factors - You should understand what is meant by controlling for other factors, holding other factors constant or holding other factors equal mean. These are all synonyms.

http://www.showme.com/sh/?h=Dlyfjv6

Interpreting regression equations - You should be able to read a regression equation. However, this is primarily preparation for the next video.

http://www.showme.com/sh/?h=anJiBbk

Interpreting tables of regression (and similar) coefficients - You should be able to read tables like the ones presented and described in this clip.

http://www.showme.com/sh/?h=TEVOuxc

Summary (Key Points)
We will often want to know whether some intervention (making condoms available in schools) has an effect of some outcome (the proportion of female students who get pregnant).

Sometimes we will only be shown whether the effect of the intervention was statistically significant at the .05 (or 5%) level. If it is, then the difference between the proportion of girls getting pregnant in schools distributing condoms and those not distributing condoms is larger (in absolute value) than would happen by chance 95% of the time. Only in 5% of the cases (or less) would it happen by chance.

At other times we will be shown the t-statistic for the hypothesis that the pregnancy rates at the two types of the schools was the same. If the t-statistic is greater than 1.96 or less than -1.96 than the probability that the difference happened by chance is less than 5% or .05. If the t-statistic is greater than 1.64 or less than -1.64, the probability is than 10% or .1.

Important Caveats:
1. Statistical significance does not mean importance. You flip a coin 1,000,000 times. It comes up heads 501,000 times. If the coin flips are fair, a deviation of this magnitude from 500,000 heads should happen less than 5% of the time. We translate this by the phrase, “We can reject at the 5% (or .05) level the hypothesis that the coin is fair. But if you are betting over something unimportant, the fact that the coin seems to come up heads 50.1% of the time is unimportant.

2. Statistical significance does not tell you causality. If there is a difference between the two types of schools, it may be the distributing condoms affects the probability of pregnancy. But it may also that the prevalence of pregnancy in the school affects the decision to make condoms available. Or some other factor may influence both.

3. We only have an estimate of the probability that something would happen by chance. Nothing dramatic happens as the t-statistic passes from 1.95 to 1.97 or the probability passes from .049 to .051. As the absolute-value of the t-statistics gets closer to 0 (the p-value gets closer to 1), it becomes increasingly likely that the difference we observe could have happened by chance, and thus we become increasingly less confident that the difference we observe is not just an accident of our particular sample.