Recitation 2 - Statistics and Causal Inference *

Nagisa Tadjfar

14.03/003 Fall 2025

Expectations and Random Variables

A random variable arises when we assign a numerical value to each event that might occur. For example, an event can be the result of three coin tosses, then $X =$ of Heads is a random variable. Each random variable has a corresponding probability distribution which is the likelihood that each possible value is assumed. In this example:

- $Pr(X = 0) = \frac{1}{8}$
- $Pr(X = 1) = \frac{3}{8}$
- $Pr(X = 2) = \frac{3}{8}$
- $Pr(X = 3) = \frac{1}{8}$

The expected value of X is the average value of X weighted by the likelihood of its possible values:

$$\underbrace{E[X]}_{\text{Unconditional Expectation}} = \sum_{x_i} Pr(X = x_i)$$

In this example, $E[X]$ is simply $\frac{12}{8}$. If the random variable is continuous, then the expected value is:

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx$$

We might also be interested in how often a random variable "departs" from its expected value. This is the *variance* of X:

$$Var(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2$$

*Thanks to Professor Autor and Jon Cohen for sharing materials from previous years.

Note that $(X - E[X])^2$ is a function of the random variable X and so is itself a random variable and therefore has an expected value. When we transform random variables (i.e. take some function of random variables), its variance may change. Adding a constant to a random variable does not affect its variance, but multiplying a random variable by a constant does. Let k and c be constants, then:

$$Var(kX + c) = k^2 Var(X)$$

Conditional Expectation

The conditional expectation of a random variable is the average realization of the random variable in a *subset of possible events*. For some subset of events where $Y = y$, or in other words, given $Y = y$:

$$E[X|Y = y] = \sum_{x_i} x_i P(X = x_i|Y = y) = \int_{-\infty}^{\infty} x f(x|Y = y) dx$$

In general, the conditional and unconditional expected values need not be the same, i.e. $E[X|Y = y] \neq E[X]$. For example, let X be a random variable corresponding to the value of a die, and A is another variable that takes the value of 1 if X is even and $A = 0$ otherwise, and B is another random variable that takes the value of 1 if X is prime and $B = 0$ otherwise. We can think of the possible events as:

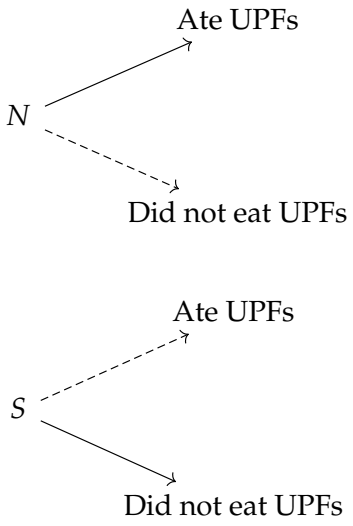
X	1	2	3	4	5	6
A	0	1	0	1	0	1
B	0	1	1	0	1	0

In this example, $E[A] = \frac{1}{2}$ and $E[A|B = 1] = \frac{1}{3}$. For the first, you can visually see that A has six possible values, and half the time it is 1 and the other half the time it is 0. For the latter, you only consider the columns where $B = 1$ and you can see that in that subset of three events, a third of the time $A = 1$ and two thirds of the time $A = 0$. Random variables are important in economics because we generally think of most of the outcomes we measure or otherwise estimate to be realizations of random variables governed by some underlying probability distribution we may not always know or that we assume.

Potential Outcomes and Randomized Control Trials

Now we go back to discussing potential outcomes using the expectations notation we just discussed. Suppose you are interested in measuring the causal effect of eating ultra-processed foods on some health outcome, e.g. cholesterol, denoted by Y. Nagisa loves

processed foods and has been consuming them daily for the last year. Salo, on the other hand, dislikes them and hasn't eaten any this past year. You measure each of their cholesterol levels today and find that $Y_{Nagisa} = 330$ and $Y_{Salo} = 200$. While these numbers are interesting, they are not as informative as we might hope. Implicitly, each of these is a realization of the *potential outcomes*. Specifically, we only observe each of Nagisa and Salo today in 1 of 2 possible states of the world:



Let's denote the path of having eaten UPFs as $U = 1$ and the path of not having eaten UPFs as $U = 0$. Then, with this notation, we observe only $Y_{N,U=1}$ and $Y_{S,U=0}$ and could in principle compute $Y_{N,U=1} - Y_{S,U=0}$, but what we are really after is $Y_{N,U=1} - Y_{N,U=0}$ and/or $Y_{S,U=1} - Y_{S,U=0}$. Each of these expressions tells us *exactly* how UPFs affect Nagisa and Salo's cholesterol levels relative to "counterfactual" Nagisa and Salo who took the other path. Unfortunately, unless we have a portal to a parallel universe, we have no way of ever knowing the values of $Y_{N,U=0}$ and $Y_{S,U=1}$. Why is this an issue? Suppose that, the "truth" is actually as follows:

- $Y_{N,U=1} = 330$
- $Y_{N,U=0} = 320$
- $Y_{S,U=0} = 210$
- $Y_{S,U=0} = 200$

The naive difference between the cholesterol measures you took would give 130, when in reality we can see that the true effect of the UPFs on cholesterol is 10 for both Nagisa and Salo. The difference between 10 and 130 is called the *selection bias*. We can gain

some insight on selection bias and how it relates to our naive raw difference by adding and subtracting $Y_{N,U=0}$ on the right hand side:

$$Y_{N,U=1} - Y_{S,U=0} = \underbrace{Y_{N,U=1} - Y_{N,U=0}}_{\text{Causal Effect}} + \underbrace{Y_{N,U=0} - Y_{S,U=0}}_{\text{Selection Bias}}$$

Using the numbers we have above for the "truth", $Y_{N,U=0} - Y_{S,U=0} = 320 - 200 = 120$ is simply capturing how much higher Nagisa's cholesterol would be compared to Salo's *even if Nagisa did not eat UPFs*. You might think that this issue is resolved if instead of sampling two people, we surveyed a large number of people. Unfortunately, selection bias will carry through even when working with large groups.

For example, suppose you surveyed $M + N$ MIT students, M of which ate UPFs and N of which did not. The difference in means of the first group's cholesterol levels and that of the second group is simply:

$$\frac{1}{M}\sum_{i=1}^{M} Y_{i,U=1} - \frac{1}{N}\sum_{i=1}^{N} Y_{i,U=0}$$

The average causal effect we are after is actually $Avg[Y_{i,U=1} - Y_{i,U=0}]$ taken over *the entire population*, but we only observe the average $Y_{i,U=1}$ among individuals who were already *choosing to each UPFs*. Similarly, we only observe the average $Y_{i,U=0}$ among individuals who were already *choosing not to each UPFs*. We have reason to believe that these two samples are not equal, i.e. those choosing to consume UPFs and those who don't may have different underlying health habits besides eating UPFs that affect their cholesterol levels. When working with groups, the difference in group means is now the average causal effect + selection bias where selection bias in this case is defined as the difference in the average $Y_{U=0}$ between the two groups compared. What can actually help address selection bias and help us estimate causal effects is *randomization* along with the law of large numbers ('LLN'). Suppose now that instead, we randomly pick M individuals from Boston and assign them to a treatment group in a study, where they are required to consume UPFs for a year. Similarly, we pick another N individuals randomly from Boston and assign them to the control group, where they are required to avoid consuming UPFs for a year. We introduce a new dummy variable, D, where $D_i = 1$ if an individual is in the treatment group and $D_i = 0$ if an individual is in the control group. For convenience, we will also assume that everyone has the same treatment effect that has the value T:

$$Y_{i,U=1} = Y_{i,U=0} + \underbrace{T}_{\text{Treatment Effect}} \quad \forall i$$

The observed difference in average cholesterol levels between the treatment and the

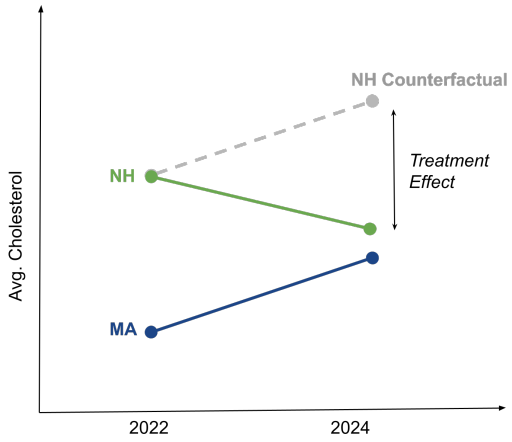
control group at the end of the study that we measure, $E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$, can be decomposed into:

$$\underbrace{E[Y_{i,U=0} + T|D_i = 1] - E[Y_{i,U=0}|D_i = 1]}_{\text{Avg Treatment Effect on Treated}} + \underbrace{E[Y_{i,U=0}|D_i = 1] - E[Y_{i,U=0}|D_i = 0]}_{\text{Selection Bias}}$$

The law of large numbers guarantees that the value of a sample average can be brought "as close as we like" to the population average from which it is drawn as we increase sample size. This means that each of the two terms in our expression for selection bias gets closer and closer to the same value so the selection bias gets closer and closer to zero. Suppose the underlying population of Boston is distributed such that the mean cholesterol level of the population is μ_0 if they didn't eat UPFs. Because we randomly sample $M + N$ people in Boston and randomly assign M of them to be in the treatment group (i.e. $D_i = 1$) and the remaining N are therefore in the control group (i.e. $D_i = 0$), the individuals in both treatment and control are drawn from the *same population*. The LLN then guarantees that as $M \to \infty$, the sample mean of the treatment group $\frac{1}{M}\sum_{i=1}^{M} Y_{i,U_i=0,D_i=1} \to \mu_0$ and as $N \to \infty$, the sample mean of the control group $\frac{1}{N}\sum_{i=1}^{N} Y_{i,U_i=0,D_i=0} \to \mu_0$. So for large enough M and N, selection bias $\to 0$. Conceptually, this is saying that if you randomly pick out two groups of people, we would expect the two groups to be the same in every way, including in their average $Y_{U=0}$. In practice when running randomized control trials, it is still considered best practice to do a "balance check" where you compare averages of different variables (e.g. age, gender, ethnicity, income) that you would expect to be the same across two random samples.

Difference-in-differences

Differences-in-differences (DiD) is a methodology that allows us to, under some assumptions, act as if we have parallel worlds. DiD is useful to estimate the treatment effect T when we can't conduct an RCT for various reasons (e.g. costs or ethical concerns). Suppose that starting in 2023, NH banned UPFs while MA did not and you observe average cholesterol levels of NH residents in both 2022 (when UPFs were still legal) and 2024 (when UPFs were no longer legal). You also observe average cholesterol levels in MA in both 2022 and 2024. Consider the figure below:



Notice that in this example, the cholesterol levels in NH start higher than in MA in 2022 and also end higher in 2024. Naively comparing cholesterol levels in 2024 between NH and MA would lead you to incorrectly conclude that banning UPFs raises cholesterol levels. Moreover, cholesterol levels in MA clearly follow a "time trend" changing between 2022 and 2024 (in this simple example we assume that nothing else is happening in either state during this time except this ban and the time trend). Instead, we are interested in comparing the change cholesterol levels between 2022 and 2024 in MA to the change in cholesterol levels in NH between 2022 and 2024. Based on the notation from lecture, what in this figure corresponds to α_{NH} and α_{MA}? What about $\delta_{t,NH}$ and $\delta_{t,MA}$? What do we need to assume about these variables and how they relate to one another between MA and NH?

Inference and Confidence Intervals

In general we do not observe *populations* (e.g. the universe of people in the US) but just *random samples* taken from these populations. This means that we will only be able to compute statistics that summarize the data we collected. An example is the sample average. However, often want to do more than just compute averages. One thing we commonly want to know is whether a mean is equal to zero, for example if we want to know if the treatment effect is zero. We typically do this using a "t-test" on a sample mean. The basic logic is as follows:

- Suppose I assume that true mean of our data is zero.
- I can compute the t-statistic, that has a known distribution conditional on some value of the mean. This means that I can tell what is the probability that I observe

certain values of t.[1]

- If I see a value that is unlikely if the data has indeed a zero mean, this suggests the original assumption that the mean is zero is wrong, and I can therefore reject my original hypothesis.

The above procedure is an application of the central limit theorem. The CLT tell us that if we have a sequence of independent and identically distributed or *i.i.d*. random variables, where $E[X_i] = \mu$ and $Var[X_i] = \sigma^2 < \infty$ then

$$\sqrt{n}\left(\frac{\sum_{i=1}^{n} X_i}{n} - \mu\right) \rightarrow^d N(0, \sigma^2)$$

Where $N(\cdot)$ is the normal distribution. That is that we know the distribution of sample averages converges to the normal distribution. And if that's true, we know that $T = \frac{\bar{x}-\mu}{\sqrt{s^2/n}}$ is distributed approximately as Student-t distribution $t(n-1)$ where s^2 is the sample variance.
So in order to test my hypotheses that the mean is zero, I assume $\mu = 0$. Then compute $t^* = \frac{\bar{x}}{\sqrt{s^2/n}}$. If t^* takes on an unlikely value given the distribution, it weighs against my hypotheses. And in particular to make concrete how "unlikely" I compute a p-value which is $p = P(|T| \geq t^*|\mu = 0)$.[2] Then we have rules of thumb about what "unlikely" is. Say usually $p = 0.05$ which occurs when $t^* \approx 2$.

A useful rule of thumb to test significance is $|\bar{x} \pm 2 * \sqrt{s^2/n}| > 0$. This means that the (approximate) 5% *confidence interval*:

$$[\bar{x} - 2 * \sqrt{s^2/n}, \bar{x} + 2 * \sqrt{s^2/n}]$$

does not overlap with 0.

[1] This distributions of the statistic encodes the sampling uncertainty surrounding it. In particular they tell us the frequency with which we will observe a value of the statistic if we sample infinitely many times from the population and repeat the computation of the statistic.

[2] The p-value therefore tells us that if we extracted infinite samples from a population with mean 0, and computed a t-stat for each of those, we would have that only 5% of the values we computed would fall above the value t^*. This is clearly a frequentist definition.