14.03 Recitation 2 Statistics and RCM*

Andrea Manera

October 30, 2021

Contents

I. Probability

A. Frequentist statistics through a simple example

We will look by example at basic frequentist statistics. In this framework, we consider each observation like the replication of an experiment. Suppose that an experiment can have either outcome A or outcome B. Then, the frequentist view is that the probability of the outcome A will be the fraction of times A is realized if we repeat the experiment an infinite number of times. This notion of probability will be very useful to keep in mind when we will look at p-values later on.

To clarify this view of probability, take the example of coin tossing. We all know that if a coin is well balanced and we flip it, we have the same probability of getting head H or tail T. Suppose you did

*I thank previous years' TAs (including my previous self) for sharing their notes. All mistakes are due to my changes.

not know this fact, and you wanted to estimate the probability p that it occurs. Frequentist statistics tells you that you should toss a coin (the "experiment" here) an infinite number of times. In that case, you will find that exactly half of the times you will get H and exactly half T, i.e. you will correctly estimate $p = 1/2$.

If you repeat the experiment only a finite number of times, however, you might get far from this result, but you will get closer to it as your sample of experiments (number of coin tosses) increases. This a key assumption of frequentist statistics: the probability of getting H or T is **fixed** (at 1/2), but the number we actually estimate fluctuates around 1/2 unless we can observe an infinite amount of events. We therefore say that the uncertainty around the parameters we estimate is *sampling uncertainty*, i.e. due to the sample we extract.

For example, if we extract a sample of just 2 coin tosses, we can make very large mistakes in estimating p. For example we could observe two heads in a row, and estimate that the probability of getting head is 1, in other cases we could get two tails, and estimate that the probability of getting head is 0. With such a small sample size, sampling uncertainty is very high. However, if extract a sample of 100 coin tosses, we really have to be unlucky to observe a streak of 100 heads or 100 tails! (the probability of observing such sequences is $.5^{100}$!). Therefore, we could be more confident that the probability p we estimate is more exact. Sampling uncertainty has decreased with the sample size.

The above result is proven theoretically. Sampling uncertainty decreases as the number of observations (number of experiments conducted) increases. The intuition can be sharpened thinking about the frequentist view of statistics: a larger number of observations "approximates infinity better," so the result on your estimates for p (which would be exactly correct if you had infinite experiments) becomes more reliable as your sample increases.

Let us now delve more deeply into the fascinating realm of definitions (...)

B. Preliminary: Set Notation

Let Ω be a generic set with generic element ω. The following definitions will be very useful in what follows.

- Element in: $\omega \in \Omega$ which tells us that some element ω is a part of set Ω
- Subset of : $A \subseteq B$ which tells us that all $\omega \in A$ are also in B.
- Union : $A \cup B = \{\omega | \omega \in A \text{ or } \omega \in B\}$
- Intersection : $A \cap B = \{\omega | \omega \in A \text{ and } \omega \in B\}$

C. Defining Probability

Consider a random experiment. A concrete example would be the employment status of some individual at some time t. Define the following:

- Let Ω be the set of all possible outcomes or sample space. In our case, Ω collects all the states of the world that can occur at time t. Denote ω the generic outcome belonging to this set.
- Let $\mathscr{E}$ be the event space which gives the combinations of outcomes we may be interested in. In our case, we are interested only in the sets of ω where the individual is either employed E, unemployed U, or non employed, N. Formally we have e.g. U is the subset of outcomes $\omega \in \Omega$ such that an individual is unemployed.[1] Examples are states of the world in which there is a deep recession. Therefore, events are subsets of the sample space Ω.
- We can define a probability measure $P(\cdot)$ as a function that maps subsets of Ω onto the interval $[0,1]$ such that the following hold:
  1. $P(\Omega) = 1$,"the probability that **anything** happens is 1";
  2. $P(\emptyset) = 0$,"the probability that **nothing** happens is 0";
  3. Monotonicity. If a subset B of Ω is contained in a subset A of Ω it holds that $P(B) \leq P(A)$. "Smaller sets have a smaller measure."
  4. Additivity. for any two empty-intersection subsets $A, B \in \Omega$ (i.e. $A \cap B = \emptyset$), and $C = A \cup B$, it holds $P(C) = P(A) + P(B)$.

  The last property tells you that if you want to measure the probability of a set C, $P(C)$, you can split it into non-overlapping parts A, B and then measure: $P(C) = P(A) + P(B)$.

As discussed above, in the frequentist interpretation, we have that e.g. $P(A)$ is the frequency of times that we would observe an outcome $\omega \in A$ if we were to observe realization of events from Ω forever. In our example, if you could observe the individual infinite times for infinite parallel dimensions (realizations ω), the probability that she is unemployed $P(U)$ would be the fraction of these infinite times that we see her working.

A *random variable* $X : \Omega \to \mathbb{R}$ is a function that maps our outcomes to numbers. An example in our case would be an indicator for whether the individual is unemployed. For example, define:

$$X \begin{cases} 1 & \text{if unemployed} \\ 0 & \text{if employed} \end{cases}$$

This is also known as a "dummy variable", that coarsely describes quantitatively a qualitative fact. Here another interesting example would be a variable that tells us the income of the individual in

[1] The subset of outcomes need not be a singleton, i.e. there might be more than just one ω such that the individual is unemployed.

various states of the world.

For each random variable we can define a *cumulative distribution function* (CDF) $F_X(x)$ as follows:

$$F_X(x) = P(X \leq x)$$

.

In addition, for discrete random variables we can define a *probability mass function*

$$f_X(x) = P(X = x)$$

and for continuous variables a probability *density* function

$$f_X(x) = \frac{d}{dx} F_X(x)$$

D. Conditional Probability

Frequently in this class we will use the concept of conditional probability. This is simply the probability of an event after we restrict our attention to only a partial subset of events. In particular we say the probability of event A given event B is given by:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

For example, the probability that an individual is unemployed is higher if we restrict our attention from the whole labor force to the set of unemployed and not employed people only.

II. Statistics

A. Expectation and Variance

Sometimes we want to know the "average value" of a random variable X in which case we can compute the expectation:

$$\begin{aligned} E[X] &= \sum_{x_i} P(X = x_i) x_i \\ &= \int_{-\infty}^{\infty} x f(x) dx \end{aligned}$$

We might also not just be interested in the "average value" but how often the distribution departs from this average value, the "variance". We can define this as: $Var(X) = E[(X - E[X])^2]$ and it's handy to define the standard deviation as the square root of this value.

B. Conditional Expectation

Another object of interest is the average realization of the random variable x in a subset of events ("conditioning on a subset of events"). This object is called *conditional expectation*, and it is defined in the same way as the mean, but using conditional probabilities.

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

C. Inference

All the above discussion is nice, but 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 mean 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.[2]
- 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 $p = P(|T| \leq t^*|\mu = 0)$.[3] Then we have

[2] 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.

[3] 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.

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.

III. Review of the Rubin Causal Model

A. Objects of Interest

- X_i: treatment actually administered to an individual i. $X_i = 1$ usually refers to treatment being administered, while $X_i = 0$ to no treatment. Therefore all individuals with $X_i = 1$ are part of the *treatment group*, while those with $X_i = 0$ constitute the *control group*;
- Y_{ij} the outcomes of some individual i, after receiving treatment j
- $T_i = (Y_i|X_i = 1) - (Y_i|X_i = 0)$: treatment effect for individual i

Denote the treatment group by $\mathbf{T}$ and the control by $\mathbf{C}$. In the notation above, what we observe in the data is:

- $Y_{j1}|X_j = 1$, i.e. the outcomes of individuals j after receiving treatment, for the group $j \in \mathbf{T}$ that has been treated ($X_j = 1$);
- $Y_{k0}|X_k = 0$, $k \in \mathbf{C}$ i.e. the outcomes of individuals k after NOT receiving treatment, for the group k that has NOT been treated ($X_k = 0$);

This is where selection issues come in. If you worry that receiving the treatment (the value of variable X) is not random, you will also worry that outcomes will be systematically different across treatment and control *groups*, *independently of the nature of the treatment!*

This is the selection bias that we have seen in class:[4]

$$E_{\mathbf{T}}(Y_{j1}|X_j = 1) - E_{\mathbf{C}}(Y_{k,0}|X_k = 0) = \left[E_{\mathbf{T}}(Y_{j1}|X_j = 1) - E_{\mathbf{T}}(Y_{j,0}|X_j = 1)\right] + \left[E_{\mathbf{T}}(Y_{j,0}|X_j = 1) - E_{\mathbf{C}}(Y_{k,0}|X_k = 0)\right] \tag{1}$$

The last term in brackets is the selection effect (or bias), the systematic difference in outcomes between the two groups absent treatment. Therefore, if we want to obtain the average treatment-on-the-treated

[4] Here I use superscripts $\mathbf{C}, \mathbf{T}$ to stress that averages are taken over control and treatment group. This is exactly the same as the notation in class:

$$E(Y_1|X = 1) - E(Y_0|X = 0) = [E(Y_1|X = 1) - E(Y_0|X = 0)] + [E(Y_1|X = 0) - E(Y_0|X = 0)]$$

effect ATT:

$$E(Y_1|X=1) - E(Y_0|X=1)$$

we have to ensure that the bias term is 0 if we want to use the above difference in means (1).
Instead, if we want to use $E_{\mathbf{T}}(Y_{j1}|X_j=1) - E_{\mathbf{C}}(Y_{k,0}|X_k=0)$ to compute the average treatment effect ATE:

$$E(Y_1) - E(Y_0)$$

IN addition to a 0 selection effect, we would need to also assume

$$E_{\mathbf{T}}(Y_{j1}|X_j=1) = E_{\mathbf{C}}(Y_{k1}|X_k=0)$$

i.e. that the outcomes of the two groups would be the same if administered treatment. Randomization, if successfully carried out, ensures both approximately apply, since the treatment and control groups are very similar to each other.

B. Examples

Miguel & Kramer 2004 found deworming medicine increased probability of attending school for a sample of Kenyan school children by 7pp. Let's think through with RCM.

- What are potential outcomes here?
  - Y_{i1} is school attendance for individual i if they take the deworming medicine
  - Y_{i0} is school attendance for individual i if they do not take the deworming medicine
- What is the ATT?
  - $E[Y_{i1} - Y_{i0}|X_i = 1]$ is the average effect on school attendance for individuals who will choose to take the drug.
- What is the ATE?
  - $E[Y_{i1} - Y_{i0}]$ is the average effect on school attendance for all individuals
- Which one is bigger and why?
  - The ATT is almost certainly bigger as individuals who have worms are more likely to take the medicine. While individuals without worms likely wouldn't benefit from the medicine and are unlikely to take it.