INFO 3401: Information Exploration

Abe Handler
Department of Information Science
University of Colorado, Boulder

November 16, 2021

Contents

1 Preliminaries

To understand this document, you need to know a few preliminary things.

First, we write $X \sim F$ to say that the observations $X = X_1, X_2, ...X_N$ are drawn from the distribution F.

Example 1.1. For instance, if F is a uniform distribution over the numbers 1 to 6 (e.g. rolling a die) and $X \sim F$ then we would expect that we have roughly equal numbers of 1s, 2s, 3s, 4s, 5s and 6s in x.

Second, if we sample **with replacement** this means that we draw samples from some population or underlying distribution one-by-one. Each time we draw a sample, we replace the thing we just sampled.

Example 1.2. For instance, if we sample 10 times with replacement from a pack of cards, the probability of getting a 10 of diamonds is 1/52 each time we draw a card. If we sample without replacement and get a king of spades on the first draw, then the probability of getting a 10 of diamonds on the second draw is 1/51 (because we do not replace the kind of spades).

2 Hypothesis testing

Say you have an intervention that you think will have some outcome in the world. For instance, perhaps you think that changing the color of the "buy" button to red on your company's website will lead more people to purchase your company's products. Or maybe you think that providing free preschool will increase the number of days in which a student attends Kindergarten. How do you know if you are right? How do you know if the intervention works?

One common way to evaluate such claims is through **hypothesis testing**. The idea is to put forth a formal proposal about the effect of a change or intervention. And then to test if you actually observe the effect or change. To grasp hypothesis testing we need to understand three interrelated concepts: effect size (Section 3), significance (Section 4) and power (Section 5).

Usually, we experiment by collecting data for a **control group** and a **treatment group**. We assume that data points are randomly assigned to either of the two groups, and that only the treatment group gets the intervention.

Example 2.1. For instance, say we want to examine the effect of changing the color of the "buy" button to red. We might show half of users the red button. Those users would be in the treatment group; they are "treated" with our intervention. The remaining half of users would see the current button. Those users would be the control group.

This document assumes that experiments are **randomized**. That means that we assume that data points are placed into either the treatment or control groups at random. We discuss randomization at the end of the document (Section 6).

3 Effect size

Say you undertake an intervention. Effect size measures: how much did the intervention change things. Did things change by a little after the intervention? Or by a lot?

Example 3.1. For instance, say you change the buy button to red and this leads to 100 times more sales. This is a large effect.

Example 3.2. For instance, say you enroll 1000 kids in preschool at random (who would otherwise not go to preschool) and this leads to 0.4 more days in Kindergarten for the treatment group. This is a small effect.

It's OK to talk about "big" or "small" effects but often we want to give a more precise quantitative measurement of effect size. There are different ways to quantify effect size. One common measure of effect size is Cohen's d which is defined as

$$d = \frac{\overline{x}_1 - \overline{x}_2}{s} \tag{1}$$

where $\overline{x}_1$ refers to the mean of group 1 (treatment) and $\overline{x}_2$ refers to the mean of group 2 (control) and where for our purposes s refers to the standard deviation of the whole data set (both groups).[1] Recall that the standard deviation is the square root of the variance, and the variance is the expected squared deviation from the mean. Intuitively, the variance is a measure of how "spread out" the data is.

Example 3.3. For instance, say we observe the data $(1, 3)$ in the treatment group $\bar{x}_1$ and the data $(1, -7)$ in the control group $\bar{x}_2$. Thus $\overline{x}_1 = \frac{1+3}{2} = 2$ and $\overline{x}_2 = \frac{1+-7}{2} = -3$. The mean μ of the overall dataset is $\frac{1+3+1+-7}{4} = -.5$ and the standard deviation is $\sqrt{\sum(x_i - \mu)^2} = \sqrt{73}$ where x_i refers to an individual point in the overall dataset (e.g. 3 or 1). Thus $d = \frac{2--3}{\sqrt{73}} = \frac{5}{\sqrt{73}} = .53$, the Cohen's d.

The denominator in Cohen's d scales the difference in means by how spread out the data is. If the data is more spread out, a bigger difference in means is less meaningful than if the data is less spread out.

In general, the bigger the effect size, the easier it will be to tell that the difference between groups is unlikely to be observed by chance, i.e. that the difference is "statistically significant" (Section 4).

4 Significance

Say you observe that the mean of the treatment group is higher than the mean of the control group. Does that indicate that the treatment actually changed outcomes? Not necessarily. We need to test if an effect is greater than what you would observe by chance.

Example 4.1. For instance, say you randomly give 10 students an SAT prep course and compare their scores to another 10 students who did not take the SAT prep course. Say that the students who enrolled in the SAT course scored 40 points higher than the control group. If the variance of SAT scores is 200 points, this means that lots of students will get scores that are hundreds of points above the mean, and that others will get scores that are hundreds of points below the mean. We would expect that in any group of 20 students, some will get high scores and others will get low scores. How do we know that the students who happened to enroll in the course didn't just happen to get high scores on the SATs? Maybe this result just happened because of random variation in SAT scores; maybe the course did not actually help.

In order to formally test if we observe a result that is greater than what we would expect by chance we need a few definitions.

[1] Different ways of defining Cohen's d sometimes use different denominators; for instance considering the standard deviations of each different group. For our purposes it is OK to simplify and just use the standard deviation of the dataset.

- H_0 the **null hypothesis**. For our purposes, the null hypothesis is that there is no effect from the treatment.
- H_1 the **alternative hypothesis**. For our purposes, the alternative hypothesis is that there is an effect from the treatment.

The basic procedure in hypothesis testing is to compute a **test statistic** T and see how likely it would be to observe T, if H_0 is true. If T is very improbable, we say that H_0 is unlikely to be true, because the chance of observing T under H_o is low.

Example 4.2. Say you have an initial hypothesis that the mean weight for a dog is 50 pounds, with a standard deviation of 10 pounds. Then you measure 10 dogs chosen at random from the kennel which each weigh around 120 pounds. The chances of seeing 10 dogs that weight around 120 pounds is very low if the mean weight for dogs is 50 pounds, so you might reject your initial hypothesis.

The test statistic T is the output from a function from data. More generally, a **statistic** is a function from data. The mean is one statistic. But any function can be a statistic. We can quantify how unlikely it is that we observe T under H_0 using a **p-value**. A p-value specifies the probability of seeing our observed test statistic T (or an even higher value of T) under H_0. A little more formally, a p-value is $p(T \geq T_{obs}$ where T_{obs} is the observed test statistic.

Example 4.3. Let $max : [x_1, x_2...x_N] \rightarrow x_j$, where x_j is in $[x_1, x_2...x_N]$ and x_j is greater than or equal to all other points in $[x_1, x_2...x_N]$. Less formally, the statistic *max* picks out the highest number in a dataset.

Example 4.4. Say an INFO instructor claims that they are a psychic, and can pick out exactly 4 real photos from a collection of 8 photos. The there are 8 ways to pick the first photo, 7 ways to pick the second photo, 6 ways to pick the third photo and 5 ways to pick the fourth photo. So there are a total of $8*7*6*5$ ways to pick the photos. The instructor could choose the 4 photos in any order, and by the same logic, there are $4*3*2*1$ ways to order the selected photos. Thus there are $\frac{8*7*6*5}{4*3*2*1} = 70$ ways to pick the photos. If the instructor is not a psychic, and is guessing randomly (which is the null hypothesis), they have a $\frac{1}{70}$ chance of picking the 4 correct photos. Hence the probability of getting all 4 right under H_0 is $p = \frac{1}{70}$. This is called **Fisher's exact test**.

4.1 Bootstrap hypothesis testing

Recall that a p-value is $p(T \geq T_{obs})$ where T_{obs} is the observed test statistic based on a sample $\boldsymbol{x} = [x_1, x_2, ...x_n]$. Say we could draw infinite data from an underlying distribution F. If we could draw infinite data, it would be really, really easy to compute the p-value. All we would need to do is keep drawing samples from F, compute a test statistic T' for each of our samples, and just observe how often T' is greater than T.

In reality, we can't keep sampling from an underlying distribution F forever. There is almost always some limit in how much data we can collect. Therefore, instead of sampling forever from F, we use our original sample $\boldsymbol{x}$ is an approximation or estimate of F. Because $\boldsymbol{x}$ is an approximation or estimate of F, sampling from $\boldsymbol{x}$ is like sampling from $\hat{F}$, an approximation of F. This is the basic idea behind bootstrap hypothesis testing. [2] Note that the distribution of data points you observe (i.e. the distribution of $\boldsymbol{x}$) is sometimes called the empirical distribution.

Example 4.5. If F is a uniform distribution over the numbers 1 to 6, F puts equal weight on the numbers between 1 and 6. If you sample from F, you will probably get a sample with roughly uniform numbers of 1s, 2s, 3s, 4s, 5s and 6s. If you then sample from $\boldsymbol{x}$ you will be likely get roughly equal number of 1s, 2s, 3s, 4s, 5s and 6s.

Example 4.6. Say you work for a marketing company interested in what people think about Chipotle. You ask Chipotle customers to rate Chipotle on a 5-point scale. Each time a customer completes a survey, Chipotle will give them free chips. F is the underlying distribution of how all Chipotle customers will rate Chipotle. You can't really observe F because there is some cost to collecting the data; Chipotle can't give everyone free chips forever. But if you give 5000 people the survey you can observe $\hat{F}$, a good approximation of F. This is a funny example, but data is almost always scarce. For instance, to collect data you might have to travel to a rainforest to count species, pay undergraduates to do a psych experiment, launch a risky feature, etc.

Say we have data $\boldsymbol{z} \sim F$ and $\boldsymbol{y} \sim G$ and we wish to know if $F = G$, that is, are the two distributions the same? To test this we will compute $T_{obs} = \hat{\boldsymbol{z}} - \hat{\boldsymbol{y}}$. Intuitively, we wanted to test if F and G were different, we would draw tons and tons of samples from F, compute a test statistic T' for each of the B samples and just record how frequently T' is bigger than T_{obs}. That would get us our p-value. If T' is rarely bigger than T_{obs} (i.e. p is small) we might reject H_0. But often there is a cost to collecting data. Instead of drawing from F forever we will keep sampling from $\hat{F}$, our estimate of the null distribution. Concretely, here is the procedure.

- Draw B samples of size $N + M$ from $\boldsymbol{x}$ where the first N observations in each sample are $\boldsymbol{z}^*$ and the second N observations are $\boldsymbol{y}^*$.
- Compute T' for each sample, where $T' = \boldsymbol{z}^* - \boldsymbol{y}^*$
- Observe what fraction of the B samples have a $T' > T_{obs}$. That is the p-value.
- If p is less than some predetermined α, reject H_0. You can conclude that $F \neq G$.

[2] This presentation comes from Efron and Tibshirani's *Introduction to the Bootstrap*.

Example 4.7. F is a normal distribution with mean 4 and standard deviation 1. $X \sim F$ and $x = [4.1, 3.9, 5.2]$

Example 4.8. Chipotle raises their prices by 10% at a store in Broomfield. F is the distribution of 5-point satisfaction judgments from Chipotle customers. $\boldsymbol{x} \sim F$ is a sample of N satisfaction judgments from Chipotles across Colorado. Chipotle solicits judgments for the Broomfield store to measure if people will give similar judgments, after the Broomfield store raises prices. $\boldsymbol{x} \sim G$ is a sample of M surveys from the expensive Chipotle in Broomfield. The question is: does F=G? You could answer via bootstrap hypothesis testing.

4.2 Permutation test

One simple and intuitive way to check for significance without making many assumptions is through a permutation test. Suppose we observe two samples. The first sample is $[x_1, x_2...x_K]$ and the second sample is $[x_K, x_{K+1}...x_N]$. If both samples come from the same distribution, then there is nothing special about which points happen to fall in the first sample and which fall in the second sample; in truth we just have one big draw from an underlying distribution $[x_1, x_2...x_K, x_K, x_{K+1}...x_N]$. The fact that we drew the points $[x_1, x_2...x_K]$ for the first sample, and then drew the points $[x_K, x_{K+1}...x_N]$ for a second sample is totally arbitrary. We might just as well have well drawn the points $[x_N, x_5...x_{K+1}]$ and then $[x_6, x_{K+1}...x_N]$; we are just drawing two samples of total size N from one distribution.[3]

The permutation test returns the probability of observing some test statistic, if the data just comes from one big distribution. Concretely, we obtain this probability as follows:

- Compute T from the two samples.
- Compute all random permutations of the data. For each permutation, compute T'.
- Observe the probability that T' is greater than T. That is the p value.

Table 1 includes an example permutation test.

5 Power

6 Randomization

[3] This presentation of permutation tests is drawn from Wasserman's *All of Statistics*.

permutation	T stat	probability
(1,9,3)	2	1/6
(9,1,3)	2	1/6
(1,3,9)	7	1/6
(3,1,9)	7	1/6
(3,9,1)	5	1/6
(9,3,1)	5	1/6

Table 1: A toy permutation test example from Wasserman