Hypothesis Testing: Decision Making using Data

Overview

A defining characteristic of rational behavior is successful decision making. Since successful decision making requires optimal use of all relevant information, statistical methods are often needed. Consider the common situation in which we are required to decide between two competing descriptions of nature. Using scientific terminology, we refer to these competing statements about reality as *hypotheses* and term formal procedures for deciding between them *hypothesis tests*. Although we continually make decisions, it is rare for us to decide between two competing claims in a formal, structured manner. For most Americans, the only time they will do this is if they are selected for jury duty. Although most of you have not served as a juror, you probably have some knowledge of jury trials through media coverage of celebrity trials and movie depictions. Therefore, we will introduce hypothesis testing by discussing trial by jury.

Competing Hypotheses: H_o and H_a

In our hypothesis testing setup, the two competing hypotheses are treated differently. Therefore we need to be clear which hypothesis is which. We do this by giving each hypothesis a name. The two hypotheses are referred to as the *null* and *alternive hypotheses* and denoted by H_o and H_a, respectively. How are the two hypotheses treated differently? The null hypothesis (H_o) is initially assumed to be true while the alternative hypothesis (H_a) is initially assumed false. We say the burden of proof rests upon the alternative hypothesis. In a trial, one hypothesis is that the defendant is innocent; the other is that the defendant is guilty. Which is H_o? Recall that in our justice system, people are assumed innocent until proven guilty. Since H_o is initially assumed true and the burden of proof is on H_a, we see that in a U.S. trial, H_o is that the defendant is innocent and H_a is that she is guilty. Note that since we initially assume H_o is true, the trial is the government's attempt to prove H_a. This is true in general: the purpose of hypothesis testing (and experimentation to acquire relevant data) is to determine if there is sufficient evidence to assert H_a is true. Because of this, in scientific settings H_a is often termed the *research hypothesis*. Let's summarize the characteristics of H_o and H_a:

H_a: alternative or research hypothesis

1. Hypothesis for which we seek evidence
2. Initially assumed false

H_o: null hypothesis

1. Initially assumed true
2. Contradictory to H_a

Proving H_a by Contradicting H_o using Data

One aspect of hypothesis testing which can be confusing is that we use proof by contradiction. We prove H_a by disproving the competing, contradictory hypothesis H_o. We do this by providing data (evidence) which contradicts H_o and supports H_a. In a trial, for example, if the defendant is accused of stealing a painting, the fact that the painting was found in the defendant's room is contradictory to her assumed innocence (H_o) and supportive of her guilt (H_a). While this example may not seem too convoluted, in other scenarios the emphasis on providing evidence against H_o to prove H_a can be confusing. (Imagine electing a candidate to office by casting votes against the competing candidates. Don't try this in Florida!) An additional complication is that it is typically not possible to perfectly contradict H_o using data. In a trial it is not possible to show that H_o *defendant innocent* is false beyond all doubt. There are usually ways of reconciling the evidence against the defendant with the assumption of her innocence. So what if she possessed the stolen painting and her DNA was on it? Someone with matching DNA could have planted it in her room. (Recall that the probability of a randomly selected person's DNA matching that found at a crime scene is typically not zero but only very small.) In order to instill (reasonable) doubt in the jury regarding the conclusiveness of evidence against H_o, the defense will attempt to construct plausible scenarios, like the one above, to reconcile the evidence with their client's innocence. However, if the jury finds these scenarios too implausible, they reject H_o and declare H_a true (defendant guilty) *beyond all reasonable doubt*. In other words, if the plausibility of reconciling the evidence with the assumption that H_o is true (defendant innocent) is too small, then the jury rejects H_o and asserts H_a true (defendant guilty). The phrase "beyond all reasonable doubt" is the justice system's allowance for the fact that it is impossible to completely contradict H_o, i.e., to prove H_o false beyond all doubt. Thus all juries are instructed to convict if the evidence is sufficient to contradict H_o "beyond all reasonable doubt."

P-value: The Plausibility of Reconciling Data and H_o

In this course, we will test hypotheses concerning numerical characteristics (parameters) of populations and processes. For example, we might want to test that the average dimension of some component coming off a production line is 5. We would therefore test H_o: $\mu = 5$ vs. H_a: $\mu \neq 5$. The distinguishing feature of our hypothesis testing procedure, as opposed to that of a jury trial, is that our procedure will be objective. While the outcome of the jury trial is subjective, i.e., it depends on the participants, e.g., the jury, judge, and lawyers, our procedure will be objective, the outcome will not depend on who does the hypothesis test. We can achieve this objectivity because our evidence against H_o consists of numerical data. Using numerical data to test our hypotheses allows us to calculate a number (p-value) describing the plausibility of reconciling the data with the assumption that H_o is true and to numerically define what is meant by "reasonable doubt."

When testing hypotheses using random data, it is usually not possible to completely contradict H_o. For example, suppose a friend tells us he can flip a fair coin in such a way

that 90% of the tosses are heads. We decide to test his claim by testing H_o: $\pi = 0.90$ vs. H_a: $\pi < 0.90$ using the data provided by 10 tosses. Since we are skeptical of his claim, we are conducting the experiment to prove him wrong. Suppose 0 of the 10 tosses are heads. While this provides substantial evidence against H_o, it doesn't prove it to be false. It is possible, although unlikely, to get 0 heads in 10 tosses if H_o is true. If fact, since he is not claiming that 100% of his tosses will be heads, there is no sequence of coin tosses which could completely contradict H_o. We can always reconcile the data with his claim by appealing to chance. If we repeat the experiment and H_o is true, the probability we will get data at least as contradictory to H_o as our original data, i.e., the probability of 0 heads in 10 tosses, is $(0.1)^{10}$ or 10^{-10}. Although very small, this probability is not 0. Of course, the fact that this probability is very small means that reconciling this data with H_o by appealing to chance is highly implausible. Thus, on the basis of the data (0 heads in 10 tosses) we conclude that H_o is false "beyond all reasonable doubt." The probability we just computed and discussed is an example of what is called a p-value:

P-value: Probability, **assuming H_o true**, of repeating experiment and getting data at least contradictory to H_o (and favoring H_a), as actual data.

Note that the p-value provides a measure of the plausibility of reconciling the data (0 heads in 10 tosses) with the assumption that H_o is true (probability of heads is 90%) by appealing to chance. If H_o is true, we have one chance in 10^{10} of observing data like what we got. Thus either something incredibly improbable occurred or else H_o is false. We conclude H_o is false.

Our coin tossing example illustrates the essential steps of hypothesis testing:

1. Identify H_o and H_a
2. Acquire suitable data
3. Compute the p-value
4. Reject H_o and assert H_a true if the p-value is "small enough"

Details of these steps will vary depending on the hypothesis testing scenario. For example, deciding how small the p-value must be for us to reject H_o, i.e., quantifying "small enough" or "beyond all reasonable doubt," requires weighing the costs associated with the two types of errors we can commit:

Type I Error: Rejecting H_o when H_o is true

Type II error: Failing to reject H_o when H_o is false

In a U.S. trial, convicting an innocent person is a type I error; failing to convict a guilty person is a type II error. Clearly there are costs associated with either error. In the U.S., we highly value individual rights so we attach a high cost to making a type I error. Therefore we require a preponderance of evidence ("evidence beyond all reasonable doubt") before we

convict someone. By making it difficult to convict someone, we make it less probable we will make type I errors, i.e., convict innocent persons. However, by making it difficult to convict someone, we also make it more probable we will make type II errors, i.e., acquit guilty persons. We have names for the probabilities of committing these errors:

α**:** Probability of committing a type I error

β**:** Probability of committing a type II error

In almost all cases, as in the trial example, we face a trade-off when attempting to minimize these error probabilities. For a fixed amount of data (sample size n) if we decrease α, we increase β. Likewise, if we decrease β, we increase α. The only way to simultaneously decrease α and β is to increase the amount of information (sample size n). Since acquiring data is usually expensive, we can never fully escape the trade-off between the two error probabilities. Thus we determine a suitable trade-off using the Neyman-Pearson approach: we weigh the costs associated with the two types of error and determine the maximum type I error probability we can tolerate. We then conduct the hypothesis test so that the probability of a type I error does not exceed this maximum value, which is called the *significance level*, or α, of the test. The rationale of the Neyman-Pearson approach is this: By using the largest type I error probability we can tolerate, we keep the probability of a type II error as low as possible. In other words, using a smaller α than necessary will cause β to be larger than necessary.

Having discussed how to determine what α to use, we are faced with one final question: How do we conduct the test so that the type I error probability does not exceed this α? The answer is simple. We use the following rule:

Hypothesis Testing Decision Rule: If the p-value $\leq \alpha$, reject H_o; otherwise, fail to reject H_o.

Summary: Hypothesis Testing Steps

Incorporating the preceding discussion of the α-β trade-off and the decision rule above into our hypothesis testing steps, we get the following hypothesis testing summary:

1. Identify H_o and H_a
2. Determine the significance level α
3. Acquire suitable data
4. Compute the p-value
5. If the p-value $\leq \alpha$ reject H_o and assert H_a true; otherwise, fail to reject H_o.