IEOR 165 – Lecture 15
Null Hypothesis Testing

> The first principle is that you must not fool yourself; and you are the easiest person to fool. So you have to be very careful about that. – Richard Feynman

1 Kidney Stone Treatment Example

In a study[1] comparing the effectiveness of two classes of treatments for kidney stones, the following success rates for each class of treatment were obtained:

Stone size	**Open surgery**	**Percutaneous nephrolithotomy**
$<$ **2cm**	81/87 (93%)	234/270 (87%)
$\geq$ **2cm**	192/263 (73%)	55/80 (69%)
Overall	273/350 (78%)	289/350 (83%)

Table 1: The numbers of successful treatments and total treatments are shown, with the success rate given in parenthesis.

What is interesting about this example is that there is a counter-intuitive result. Percutaneous nephrolithotomy has a higher success rate when all stone sizes are grouped together, but open surgery has a higher success rate when comparing based on stone size. This result shows the need for careful consideration of data when making comparisons.

1.1 Explanation of Simpson's Paradox

The natural question to ask is: Why does this odd behavior occur in the kidney stone treatment example? When comparing the treatments with the aggregated data, an assumption is implicitly being made that the decision for which treatment to use does not depend upon the size of the kidney stones. It turns out that this implicit assumption is incorrect, because open surgery was more often used for larger kidney stones; however, larger kidney stones in general have a lower treatment success rate because of its more complicated nature. This counter-intuitive result is sometimes called Simpson's paradox, though it is actually a manifestation of the more general statement that "correlation does not equal causation".

[1] C. Charig, D. Webb, S. Payne, J. Wickham, "Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy", *Br Med J (Clin Res Ed)*, vol. 292, no. 6524, pp. 879–882, 1986.

2 Null Hypothesis Testing

In many contexts (e.g., scientific, engineering, policy making, etc.), we must make a single decision amongst a set of possible choices. Hypothesis testing is a set of *multiple* frameworks for data-driven decision making. In this course, we will cover two such frameworks. The first is called *null hypothesis testing*. The idea is as follows:

1. We begin with a base assumption about our system.
2. Data is measured from the system.
3. Then using measured data, we compute the probability of making observations that are as (or more) extreme than was measured. This probability is computed under the base assumption about the system, and this probability is called the p-value.
4. We make a decision on the basis of this probability. The reasoning is that if the above probability is small, then it is unlikely that we would have observed the measured data if the base assumption were true.

There is a subtle fundamental tension underlying this framework (and actually present within all of hypothesis testing), which continues to cause significant controversy with hypothesis testing. The tension is that we must make a *decision* regarding truth, but the framework of null hypothesis testing does not relate to this concept; instead, we make decisions on the basis of trying to invalidate some base assumption about the system. Moreover, the quantity we use to make this decision is *not* the probability that the null hypothesis is true. Rather, the quantity is the probability of observing extreme events under the assumption that the null is true.

3 Coin-Flipping Example

To make the discussion more concrete, consider the following scenario. We are given a specific coin, and we would like to decide if the probability of getting heads on a single coin flip is exactly 50%.

3.1 Formulating the Null Hypothesis

The first step of this process is that we must convert the possible decisions about the coin into a mathematical formulation. Specifically, we must convert the scenario into a base assumption about the system. In this case, one possible base assumption is that

- the distribution of a single coin flip is given by a Bernoulli distribution;
- the result of each coin flip is (mutually) independent;
- the probability of heads is exactly 50%.

3.2 Designing the Experiment

The second step of this process is that we must conduct an experiment to gather data that we will use to make our decision. Part of this step involves the design of the experiment. In our particular example, the design is the answer to the question: How many times should we flip the coin? If we flip the coin only once, then we will not have a lot of data to make our decision; and so we might be less confident with our decision. If we flip the coin one-billion times, then we will have a lot of data to make our decision; but doing so many experiments can be prohibitive because of cost or time constraints we might have. We would prefer to choose an intermediate value. For this example, suppose we decide to do $n = 100$ flips.

3.3 Computing the p-Value

The third step is that we need to compute the probability of making observations that are as (or more) extreme than was measured. To do this, observe that after our experiment the data will consist of a sequence of flip results. So perhaps the flips were

$$H, H, T, H, T, T, H, T, T, H, \ldots$$

To compute this p-value, we have to use the properties of our base assumption. The key insight in this case is that the distribution for the total number of heads given n trials is a binomial distribution. So in fact, the raw data is not directly used to compute the p-value. Instead, we take the raw data and count the number of heads. For our example, suppose we have 40 heads total. For this example, the situations that are as (or more) extreme are having 0 to 40 heads or having 60 to 100 heads. This probability is given by

$$p = \sum_{k \in \{0,1,\ldots,40,60,61,\ldots,100\}} \binom{100}{k} 0.5^k \cdot 0.5^{100-k}.$$

Using MATLAB, this value is equal to $p = 0.0569$.

3.4 Making a Decision

The final step is that we must make a decision. Many textbooks use the terminology that we either *accept* or *reject* the null hypothesis. If we accept the null hypothesis, mathematically this means that we do not have enough data to invalidate the hypothesis; however, in terms of decisions this is a decision that the null hypothesis is true. This mismatch between the mathematics and the corresponding decision is one source of controversy about hypothesis testing. Similarly, if we reject the null hypothesis, mathematically this means that the data (or more extreme data) would be observed under the assumption of the null hypothesis with low probability; however, in terms of decisions this a decision that the null hypothesis is false. Furthermore, the framework in many textbooks for accepting or rejecting the null is to use a significance level α. If the p-value is above (below) the significance level, then we accept (reject) the null.

There is an additional source of controversy in this final step. What significance level α should we use to make decisions? The smaller the significance level, the more stringent the test is in terms of requiring greater amounts of evidence to reject the null. In many sciences, it is customary to use $\alpha = 0.05$ or $\alpha = 0.01$. (In particle physics, it is common to choose a significance level of roughly $\alpha = 5 \times 10^{-7}$.) One might ask what is the significance of the significance levels $\alpha = 0.05$ or $\alpha = 0.01$? And the answer is that these values are somewhat arbitrary and have become common because of tradition.

Given this arbitrariness, another way to make decisions is to examine the risk of accepting or rejecting the null, and take the risk of each possible decision and the p-value into account when making the decision. In the case of coin flips, if we are going to use the coin for deciding who serves first in a volleyball match then there is a low risk for incorrectly accepting the null; and so we we would accept the null since the p-value ($p = 0.0569$) is fairly large. However, the risk is context dependent. In another scenario, perhaps we are using coin flips to determine who must be deployed in a war. There is a high risk for incorrectly accepting the null, and so we would reject the null since the p-value is fairly small.

Moreover, there is often a third choice that is available beyond the conventional "accept" or "reject" decisions: In some situations, we can make a decision to collect more data and conduct additional analysis before we make a final decision regarding the system. However, it is important to keep in mind that the decision to collect more data can carry risks itself. In addition to the fiscal costs of conducting additional experiments, there are also costs that can be incurred by delaying the final decision.