IEOR 151 – Lecture 3
Specifying Tests

1 Kidney Stone Treatment Example

1.1 Simpson's Paradox

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 counterintuitive 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.

1.2 Explanation of 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 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 is general have a lower treatment success rate because of its more complicated nature.

[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 Specification of a Two-Decision Problem

A two-decision problem is a situation in which there are two possibilities, and one of these possibilities must be chosen as being the more "likely" one given data about the possibilities. These types of problems are usually defined in terms of testing a null hypothesis that is accepted or rejected. There are a variety of approaches to evaluating such decisions problems, and we will focus on methods where the probability of rejecting a null hypothesis when it is true (type I error or false positive) is controlled to be below a significance level α.

2.1 Defining the Hypotheses

The null hypothesis is the choice that is to be believed by "default" unless data suggests otherwise. An example of a null hypothesis is that measurements from two groups $A_i \sim \mathcal{N}(\mu_a, \sigma_a^2)$ and $B_i \sim \mathcal{N}(\mu_b, \sigma_b^2)$ have the same mean $\mu_a = \mu_b$ with possibly different variances.

The alternative hypothesis is the "opposite" choice. There are three important classes of alternative hypothesis: one-tailed directional, two-tailed directional, and non-directional. For the example above, corresponding alternatives *could* be

1. One-tailed directional: $\mu_a < \mu_b$;
2. Two-tailed directional: $\mu_a < \mu_b$ or $\mu_a > \mu_b$;
3. Non-directional: $\mu_a \neq \mu_b$, which is equivalent to the two-tailed directional in this example.

The reason for a distinction between one-tailed and two-tailed tests is related to the idea of the *power* of a test. Before we can define what this means, we first have to discuss how to control type I errors.

2.2 Significance Level

A test statistic is a function of the observed data that is used to summarize the data being tested. For our example above when assuming that the variances are known to be $\sigma_a^2 = \sigma_b^2 = 1$ and the number of samples from the two groups are equal $n_1 = n_2 = n$, a possible test-statistic for comparing the means is

$$t = \sqrt{\frac{n}{2}} \left(\frac{1}{n} \sum_{i=1}^{n} A_i - \frac{1}{n} \sum_{i=1}^{n} B_i \right).$$

The probability of observing a test statistic that is as extreme or more extreme, under the distribution specified by the null hypothesis, is called the p-value. If a null hypothesis is rejected only when the p-value is smaller than the significance level α, then the probability of making a type I error (false positive or falsely rejecting the null hypothesis) is controlled to be smaller than α. On the other hand, if the p-value is greater than α, then the null is accepted.

The interpretation is subtle. Technically speaking, the null hypothesis cannot be proven to be true or untrue. Rejecting the null hypothesis is a statement that the probability of making the same measurements as were observed, under the null distribution, is "small". On the other hand, accepting the null hypothesis is a statement that there is not enough evidence to be able to reject the null hypothesis; it is not a statement that the null hypothesis is true. Real-world practice typically necessitates making a decision (accepting or rejecting the null), and this can obscure the formal meaning of the tests.

There is another subtle point that is important to note. In practice, the choice of the significance level is often a subjective decision. And if the choice of the significance level can be subjective, then the decision made by the hypothesis test can also be subjective. What this means is that risk (or cost) introduced by making an incorrect decision in either direction should be considered when choosing the significance level and when accepting or rejecting the null hypothesis. It can be the case that a null hypothesis should be rejected given the chosen significance level (e.g., $p = 0.049$ and $\alpha = 0.05$), but the risk (or cost) of making type I error is high enough that the null is instead accepted. The alternative situation can also occur.

2.3 Power

The power of a test is the probability that the null hypothesis will be rejected when the alternative is true. As an extreme example, suppose that the null hypothesis is always accepted. Then, the probability of making a type I error is always smaller than α, no matter what the value of α is. However, this test is not useful because it never rejects the null hypothesis. In this extreme example, the test has low power.

Strictly speaking, we cannot compute the power of a test under this framework because we do not prescribe a distribution for the alternative hypothesis. However, we can qualitatively affect the power through our choice of a one-tailed versus two-tailed directional alternative. Consider the test-statistic we defined above

$$t = \sqrt{\frac{n}{2}}\left(\frac{1}{n}\sum_{i=1}^{n} A_i - \frac{1}{n}\sum_{i=1}^{n} B_i\right) = \sqrt{\frac{n}{2}}\left(\frac{1}{n}\sum_{i=1}^{n} A_i - \mu + \mu - \frac{1}{n}\sum_{i=1}^{n} B_i\right),$$

Next note that the Central Limit Theorem implies that $\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} A_i - \mu) \xrightarrow{d} \mathcal{N}(0,1)$ and $\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} B_i - \mu) \xrightarrow{d} \mathcal{N}(0,1)$. But since A_i, B_i are independent, we can conclude by a standard result that these sequences jointly converge as

$$\begin{pmatrix} \sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} A_i - \mu) \\ \sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} B_i - \mu) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \mathcal{N}(0,1) \\ \mathcal{N}(0,1) \end{pmatrix},$$

where the components of the limiting distribution are independent (cf. the example for Slutsky's theorem from last lecture). A simple application of the Continuous Mapping Theorem gives that

$t \xrightarrow{d} 1/\sqrt{2}\mathcal{N}(0,2) = \mathcal{N}(0,1)$. The one-tailed and two-tailed alternative hypothesis involve computing p-values using either one or both tails of this limiting Gaussian distribution.

This shows the intuition behind the names one-tailed and two-tailed. When the alternative hypothesis is the one-tailed directional $\mu_a < \mu_b$, then we only need to consider the left side of the Gaussian distribution when computing p-values. If the alternative is the two-tailed directional $\mu_a < \mu_b$ or $\mu_a > \mu_b$, then we need to consider both sides of the Gaussian distribution when computing p-values. Qualitatively, using a two-tailed alternative when a one-tailed alternative could be used leads to a lower power because it allows for possibilities that are not relevant to the situation. However, using a one-tailed alternative when a two-tailed alternative should be used can cause the test to not meet the desired significance level; it is possible for the p-value to artificially decrease in this situation.