The Process of Statistical Tests

The Structure of Statistical Tests

A statistical test is a little bit like a proof by contradiction in mathematics. We start by assuming something and then we do some work under that assumption, and we end up with a contradiction, or something that is obviously wrong; for us, obviously wrong will mean unlikely in a probability sense. As a result we conclude that there must be something wrong with our original assumption.

In statistics, tests are used to answer direct questions about the theoretical world and the goal is to determine whether the data provide enough evidence for us to believe a claim about the population or theoretical world. Statistical tests are often called *hypothesis tests*, or *tests of significance*.

In this section, we will outline some of the steps and structure of a statistical test, and we will introduce some of the vocabulary that is used. We will begin by carefully examining each step of a statistical test.

Step 1: Determine the null hypothesis and the alternative hypothesis
The first step of a statistical test is to formulate a hypothesis to test, and specify an alternative to that hypothesis. Throughout this section we are going to use a court case as an analogy to a statistical test. Many legal systems around the world are based on the presumption of innocence, often referred to as *innocent until proven guilty*. We presume that our starting point is that anyone charged with a crime is innocent, and the burden of proof is on the prosecution to prove otherwise.

In statistics the presumption of innocence corresponds to what we call the **null hypothesis**. Written as H_0 ("H-naught"), the null hypothesis is the status quo. It states that nothing is happening, or there is no relationship, or there are no differences. We assume it is true throughout and see if our data are likely under that assumption or if they contradict it.

The **alternative hypothesis** corresponds to showing that the defendant is guilty. Sometimes the alternative hypothesis is called the research hypothesis because it is what the research wants to show. The alternative hypothesis is written as H_a or H_A or H_1. Like the prosecutor who wants us to conclude that the defendant is guilty, we typically want to prove that this researcher alternative hypothesis is true. The way to make that conclusion is to see if we have enough evidence to rule out that any differences from the null hypothesis could be due to chance. Therefore, an important first step in any statistical test is to carefully define what our null hypothesis is and what the alternative is.

Example 1
Suppose our goal is to show that facial plastic surgery reduces the mean perceived age, calculated as the perceived age before minus after surgery. Our null hypothesis would be that nothing is going on, so we would assume that the mean perceived age change is 0. Our alternative hypothesis would be that the perceived age change is positive.

This is an example of a **one-sided alternative**; we are only interested if the mean perceived age change is positive, meaning patients look younger. We can also have **two-sided alternatives**. In our beer cap example, if we are interested in if it is 50/50, we would be interested in whether the chance of getting reds is greater than a half or if it is less than a half. So then we would use it a two sided alternative, corresponding to not equal to a half. In most situations two-sided tests are appropriate, and if you are ever in doubt use a two sided test.

Step 2: Collect the data and calculate a test statistic
In a court case we need evidence collected by the police and prosecution. Similarly, the second step in our statistical test is also to put together the evidence. In statistics the evidence is provided by our data. In order to make the data useful, we have to summarize it into a statistic called the test statistic. The **test statistic** is a numerical summary of the data that is formulated assuming that the null hypothesis holds. This is important to remember in statistical testing: we make an assumption, namely our null hypothesis H_0, and then work as if it were true.

Example 2
If we assume that a face lift does not lower the mean change in the perceived age of a patient, then we construct the test statistic under the assumption that the mean age change is zero.

Step 3: Determine the p-value - how unlikely the test statistic is if the null hypothesis were true

Once the evidence has been collected and summarized for the judge and jury, they have the responsibility of careful deliberation. They need to decide if the evidence is overwhelming enough to reject the presumption of innocence beyond a reasonable doubt. If the evidence is not beyond a reasonable doubt, that means that chance or the natural variability we expect is a reasonable explanation for anything we see in our data that is not consistent with what we would expect if the null hypothesis is true. To decide whether the results could just be due to chance, we ask the question: "If the null hypothesis is really true, how likely would it be to observe a test statistic of this magnitude or even larger, just by chance?"

In statistics, the tool for deliberation is a p-value. A **p-value** transforms a test statistic into a probabilistic scale. It is a number between 0 and 1, that quantifies the strength of the evidence against the null hypothesis. The smaller the p-value, the more unlikely it is that we observe our data assuming the null hypothesis is true, and therefore the stronger the evidence we have against the null hypothesis.

A p-value is not a measurement of how likely it is that the null hypothesis is true. The null hypothesis either is true or is not true. We may not know which it is, but we cannot put a probability value on it because it is not random. What a p-value does tell you is how likely the observed data would be, if the null hypothesis were true.

Step 4: Make a conclusion based on the p-value and the context of the problem
To reach a final verdict, our judge or jury has one of 2 choices:

1. They could decide that the evidence is not strong enough to convincingly rule out that the defendant is innocent and they would conclude not guilty. In statistics, not strong enough evidence means the p-value was not small, and we conclude that our data are consistent with the null hypothesis. We cannot conclude H_0 is exactly true, but we cannot reject it either.
2. Our jury could decide that they are willing to rule out the possibility that an innocent person stacked up this much evidence against them, and then they would reject the presumption of innocence and conclude that he or she is guilty. The corresponding situation in statistics is a small p-value. We have data that are unusual enough that we are willing to rule out that the null hypothesis is true, and by chance we got the observed data that we did. In this case we have a small p-value and we reject the null hypothesis, concluding that we have sufficient evidence against it and that the alternative hypothesis must be true. In this case, we say we have a **statistically significant** result.

In Figure 1 we have a general guideline for conclusions based on different p-values:

P-Value	Strength of evidence against the null hypothesis
P-value < 0.001	Very strong
0.001 < P-value < 0.01	Strong
0.01 < P-value < 0.05	Moderate
0.05 < P-value < 0.1	Weak
P-value > 0.1	None

Figure 1: General guideline of strength of evidence against H_0 based on p-value

When we obtain a larger value for our p-value, the evidence against the null hypothesis is not as strong. And if we get a p-value greater than 0.1, that means that if the null hypothesis were true, what we have observed in our data is something that happens 10% of the time. Then we conclude that our data are consistent with the null hypothesis.

Depending on the context of the problem, you might be willing to accept a larger or a smaller p-value. If concluding that you have a statistically significant result is going to incur a lot of expense or possible discomfort for people, then you will want a smaller p-value.

In summary, here are the four steps of a statistical test:

1. Determine the null hypothesis and the alternative hypothesis.
2. Collect the data and calculate a test statistic, assuming H_0 is true.
3. Calculate the p-value.
4. Make a conclusion based on the p-value and the context of the problem.