The Effective Use of Statistical Tests

Power and Type I and Type II Errors

When carrying out a statistical test about their data, it is often the case that researchers would like to show that the alternative hypothesis is true. That is, they would like to reject the null hypothesis. If a study is well designed, there will be a high probability of rejecting the null hypothesis when indeed it isn't true. Such a test is said to have high power. In this section, we will discuss more about the concept of power and the possibilities that a statistical test may result in an incorrect conclusion.

When carrying out a statistical test, the first step is to define our hypotheses. In this section, we will assume that our null hypothesis is that some parameter from the theoretical world, whether it's the probability of heads or the mean of a distribution of, say body temperatures, is equal to some hypothesized value. Our alternative hypothesis is that the parameter is not this hypothesized value. Depending on the situation we can use a one-sided or a two-sided alternative.

$$H_0 : \text{parameter} = \text{hypothesized value vs } H_a : \text{parameter} < \text{or} > \text{or} \neq \text{hypothesized value}$$

We then calculated a test statistic from our data under the assumption that H_0 is true. Assuming the null hypothesis is true, the p-value is the probability of observing the value of the test statistic that we got, or a value more extreme. Recall that small p-values give evidence against the null hypothesis.

The **significance level** of a test gives a cut-off for how small is small for a p-value, and it is denoted by the Greek letter α ("alpha"). If we have a significance level α, the p-value is then the smallest level of α at which the data are statistically significant. Sometimes the p-value is thus called the **observed level of significance**.

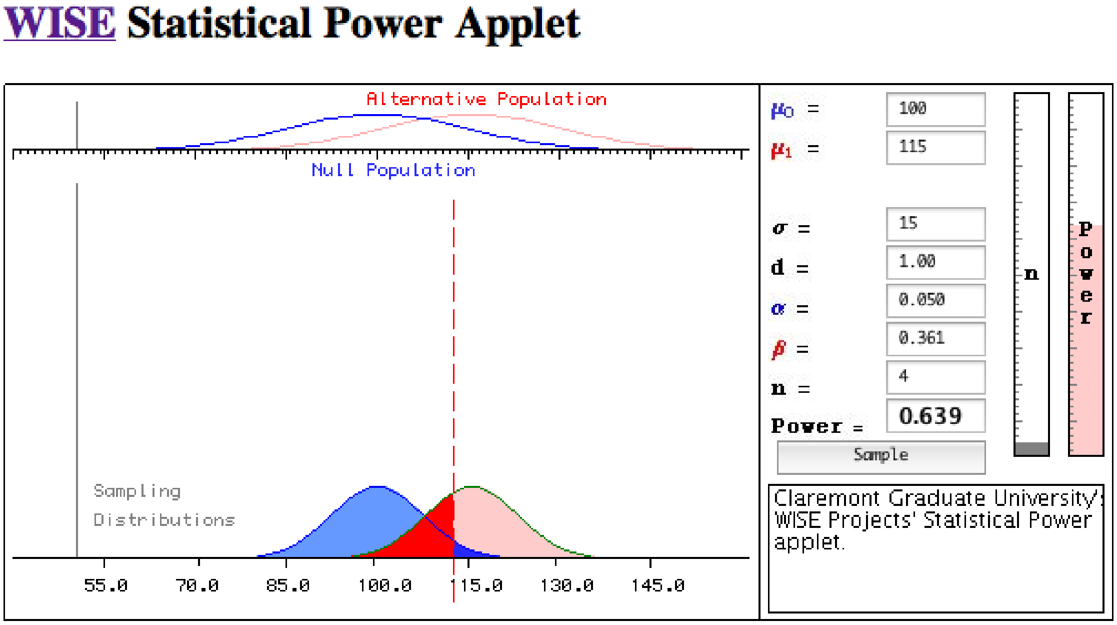
The significance level α also gives us a measure of how the test performs in repeated sampling. If H_0 is true and you use a significance level of $\alpha = 0.01$, and you carry out a test repeatedly with a different sample of the same size each time, you will reject H_0 (a wrong conclusion!) 1% of the time. The good news is that 99% of the time you would not reject the null hypothesis, which would be the right decision.

It may sound like a good idea to set the significance level to be very small, so that you do not mistakenly reject the null hypothesis. The problem is that if you set α too small, you may never reject H_0, even if the true value is very different from the null hypothesized value. What we need is a high probability that the test will reject H_0 when an actual alternative value of the parameter is true.

For a fixed significance level and a particular alternative value of the parameter being true, the **power** of a test is the probability of making a correct decision (by rejecting the null hypothesis) when the null hypothesis is false. The higher the power of a test, the more sensitive it is in detecting a false null hypothesis. As its name implies, power is a good thing and we want it to be high. So what do we need to do in order to get higher power? We will investigate this question using the WISE Statistical Power Applet.

Example 1

Below we see a screenshot of an applet created by the Web Interface for Statistics Education (WISE) project. By clicking here, you can access the applet and follow along with the example.



WISE Statistical Power Applet

μ_0 = 100

μ_1 = 115

σ = 15

d = 1.00

α = 0.050

β = 0.361

n = 4

Power = 0.639

Claremont Graduate University: WISE Projects' Statistical Power applet.

Figure 1: Screenshot of WISE Statistical Power Applet

For simplicity, this applet assumes that all distributions of data are normally distributed. The blue normal distribution at the top labeled the Null Population represents our theoretical world model under the assumption that the null hypothesis is true; that is $\mu_0 = 100$. To calculate a numerical value for power, we need to do so for a particular value of the alternative. Here, that alternative value is $\mu_1 = 115$, and the corresponding Alternative Population is shown as the red normal distribution.

Note that both the null and alternative populations have the same spread and that is quantified by the value of $\sigma = 15$. We are interested in testing the mean, so we will keep the standard deviation the same in the null and alternative populations.

The shaded blue and red normal distributions at the bottom are our sampling distributions for the mean, $\bar{X}$, or the average of our sample data. Recall that the sample mean or average is an unbiased estimator, so these distributions are centered at the same values as our theoretical world models. However, these sampling distributions have less spread.

Our alternative here is one-sided, a greater than alternative, so we are testing

$$H_0 : \mu = 100 \text{ vs } H_a : \mu > 100$$

Since the truth is that the mean is 115, to calculate power we use the red sampling distribution for the alternative population, or alternative theoretical model. We still reject H_0 if the test statistic is to the right of the vertical bar and the probability that we correctly reject is shaded in pink. Our power, the probability we correctly reject the null hypothesis, is 0.639 for this situation.

Clicking Sample randomly selects a sample of size 4 from the alternative population. We can see the randomly chosen sample at the top, and the resulting mean among the sampling distributions. The sample statistics and the conclusion of the test can be found in the bottom right output window. If you clicked Sample 10 times, you should expect to get $6.39 \approx 6$ "Reject H_0" conclusions.

Example 2

In the WISE Statistical Power Applet, enter in the information $\mu_0 = 100, \mu_1 = 130, \sigma = 15, \alpha = 0.05, n = 4$. What happens to the power of the test as the alternative value moves further away from our hypothesized null value of the mean (μ_1 moves further away from μ_0)? Since we are holding everything else constant, we should be more likely to reject the null hypothesis since the true alternative value is far from it. The power in this case is calculated to be 0.991 which is quite high!

Example 3

In the WISE Statistical Power Applet, enter in the information $\mu_0 = 100, \mu_1 = 115, \sigma = 15, \alpha = 0.10, n = 4$. What happens to the power of the test when the criteria for rejecting H_0 is less strict (α is increased)? We will reject if the p-value is less than 0.1 or 10%, so we are rejecting more often. As a result, the probability that we correctly reject the null hypothesis, our power, should be larger. In this case, the power is calculated to be 0.764, which is larger than our original value of 0.639.

Example 4

In the WISE Statistical Power Applet, enter in the information $\mu_0 = 100, \mu_1 = 115, \sigma = 10, \alpha = 0.05, n = 4$. What happens to the power of the test when we decrease σ, the standard deviation of the population? Less variability should give us more precise estimates. Increased precision is a good thing, and this has increased our power to 0.912.

Example 5

In the WISE Statistical Power Applet, enter in the information $\mu_0 = 100, \mu_1 = 115, \sigma = 15, \alpha = 0.05, n = 25$. What happens to the power of the test when we increase n, the sample size? Our sampling distributions are now much more tightly clustered around the mean, reflecting the smaller amount of variability we have in our estimator of the mean. As a result our power goes up, computed in this case to be approximately 1.000.

To summarize, we have seen 4 different ways we can increase the power of a test

1. The power is higher the further the alternative value is away from the null hypothesized value.
2. A higher significance level α gives higher power.
3. Less variability gives higher power.
4. The larger the sample size, the great the power.

To *determine the sample size* needed for a study for which the goal is to get a significant result from a test, set α and the desired power, decide on an alternative value that is practically interesting, estimate σ, and calculate the sample size necessary to give the desired power.

There are two types of errors we can make in statistical testing:

- **Type I error: reject H_0 when it is true:** This happens with probability α. In the courtroom, a Type I error is analogous to an innocent person being falsely convicted.
- **Type II error: fail to reject H_0 when H_a is true:** This happens with probability β ("beta"). In the courtroom, a Type II error is analogous to a criminal is erroneously freed. In relation to power, $\beta = 1 - \text{Power}$.

In practice we would like to have low probability of both Type I and Type II errors, but lowering the probability of one type of error, generally speaking, increases the probability of the other type of error. Increasing the sample size can decrease the probability of both types of error. This is analogous to having more evidence in a criminal trial, leading us to be more likely to make the correct verdict. But it is often a balancing act between the two types of error, and which is more serious can depend on the context. A medical analogy can illustrate this point.

Example 6

Suppose we have a diagnostic medical test for determining whether a patient has a disease or not. For this example, we have the following null and alternative hypothesis

$$H_0 : \text{patient does not have the disease vs } H_a : \text{patient does have the disease}$$

and therefore, we have the errors

$$\text{Type I error} = \alpha = P(\text{test says patient has disease when he does not})$$
$$\text{Type II error} = \beta = P(\text{test says patient does not have disease when he does})$$

If a decrease in α increases β and vice versa, it is necessary to decide which type of error has more serious consequences, so that you can sacrifice making one type of the error in order to lessen the chances of making the more serious error. But which type of error is more serious?

If the disease is treatable and the patient will die without the treatment, then the Type II error would be more serious, as an incorrectly diagnosed ill patient would miss out on the treatment. But if the treatment had its own serious side effects and the disease was not fatal, a Type I error might be more serious.

In medicine, all tests have associated false positive and false negative rates, with a trade-off between the two. How to manage that trade-off needs to be determined based on the consequences for the patients. The same is true of statistical tests, where we have to consider the implications of making a wrong decision. If choosing to reject the null hypothesis, which is typically the status quo, will incur a great deal of expense, then we won't want to do that unnecessarily. So we won't want to make a Type I error and we will insist on having a very small p-value, very strong evidence against the null hypothesis, before we are willing to reject it. In the next section, we will consider some more advice about carrying out statistical tests in practice.