Microsoft

Data Science Essentials

Hypothesis Testing

Hypothesis testing is a core skill in statistics. You use hypothesis tests to evaluate data and determine whether or not a hypothesis could be supported by the dataset.

Hypothesis Tests and P-Values

A hypothesis test uses statistics to answer a yes or no question about some data set, and the result tells you whether to reject a *null hypothesis* (which often represents the view that conditions have not changed) in favor of an *alternative hypothesis* (which represents a reason for the observed result different than the null hypothesis). For example, suppose a cupcake store sells chocolate and vanilla cupcakes. You might suspect that each customer will have a preference for a particular flavor, and that more customers might prefer one flavor (for example, chocolate) over the other (vanilla). The null hypothesis (which we label H_0) for your test is that customers will choose chocolate or vanilla in equal numbers (in other words, there is a 50% probability of the customer choosing vanilla, and a 50% probability that their preference will be chocolate). The alternative hypothesis (H_1) is that there is some non-equal preference on the choice of flavor, so that the probability of a particular flavor (say. chocolate) being chosen is not 50%.

These hypotheses can be expressed as:

$$H_0: P = 0.5$$
$$H_1: P \neq 0.5$$

Given a suitably large sample of cupcake sales data, you can determine the actual number of chocolate cupcakes sold compared to the total sales, and work out how probable that result is if the null hypothesis is true. For example, suppose the sample data includes 100 sales; 70 of which were for chocolate cupcakes, and 30 of which were for vanilla cupcakes. If each cupcake sold has an even 50% probability of being either chocolate or vanilla (as stated in the null hypothesis), then based on a binomial distribution, the probability of 70 out of 100 cupcake sold being chocolate flavored is approximately 0.0023%. Note that the hypotheses are always about the population, and not about the sample. (The population mean is unknown, but the sample mean is not, thus we do not need to create hypotheses about the sample.)

The probability to observe what we did (or something more extreme) under the null hypothesis is known as the *P-value*. Based on a pre-determined threshold known as the *significance level* that you decide; you can reject the null hypothesis or not. In most cases, a value of around 0.05 (or 5%) is chosen as the significance level, and in this case the P-value is much lower than this; so the null hypothesis can be rejected in favor of the alternative hypothesis. If the P-value is lower than the significance level, we reject the null hypothesis.

Types of Test

There are numerous types of hypothesis test that you can conduct, depending on the type of data and the alternative hypothesis you are trying to validate. Many tests are focused on evaluating the mean of a given dataset and comparing it to an expected value.

Single-Sample T-Tests and Z-Tests

Suppose our cupcake store expects to sell an average of 75 or more cupcakes per day. You could record actual sales figures over a period of time and perform a test to determine whether the mean sales figure is greater than 74. Depending on the volume of sample data available, you can perform a *z-test* (which you should use for normal distributions with a known population standard deviation, or for data sets with more than 30 independent observations – in which case the sample standard deviation is close enough to the population standard deviation) or a *t-test* (which can be used with a small number of observations and when the population standard deviation is not known).

The result of the z-test or t-test includes a p-value, which you can use to determine whether or not to reject the null hypothesis. In this case, the null hypothesis is that the mean sales amount will be 75 or more, and the alternative hypothesis is that average daily sales will be less than 75. This can be expressed like this:

H_0: P >=75
H_1: P < 75

This is an example of a one-tailed test, in which we are testing whether or not the population mean could be greater than a specified value. You could also perform a one-tailed test to determine whether or not the population mean is less than the expected value, or you could perform a two-tailed test to determine whether or not the population mean varies from the expected value in either direction.

Two-Sample Tests

In addition to single-sample tests, you can perform tests that compare two samples. For example, suppose you want to test the hypothesis that on average, chocolate cupcakes weigh more than vanilla cupcakes. To test this hypothesis, you can individually weigh a set of chocolate cupcakes and a set of vanilla cupcakes, and then conduct a t-test that compares the mean weight of each set. The resulting pvalue will indicate the significance of the difference in mean weights.

Comparing the mean weights of two different cupcake flavors is an example of an unpaired test. The individual observations (the measured weights of each cupcake) are independent – you could even include more vanilla cupcakes than chocolate cupcakes (or vice-versa) without affecting the outcome of the test. However, some two-sample tests are paired tests in which there is a dependency between the observations in the two datasets. For example, suppose you wanted to test the hypothesis that the daily average sales figure of chocolate cupcakes is higher than that of vanilla cupcakes. In this case, the two sets of observations must be paired so that the first observation in each sample is the total favor-specific sales for the first day, the second observation is the total favor-specific sales for the second day, and so on.

Confidence Intervals

Without having access to the total population of the data, you need to be able to determine whether the sample mean ($\bar{x}$) is likely to approximate the population mean (μ). A *confidence interval* is a way to express the probability that a true population parameter falls within an interval of a statistic of the data.

The confidence interval for the mean μ would be centered at $\bar{x}$, and the size of the confidence interval would depend on the sample standard deviation of points and the total number of sample points.