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 is supported by the statistics evaluated against 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 generally represents the view that an observed result can be explained by pure chance) in favor of an *alternative hypothesis* (which represents a non-random reason for the observed result). 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-random 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 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%.

The probability that the observation can be explained by pure chance is known as the *P-value*, and based on a threshold known as the *significance level* that you decide for this value, 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.

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 central limit theorem makes the sample standard deviation close enough to the population standard deviation) or a *t-test* (which can be used with a small number of observations 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 not exceed 74. 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 sample mean is greater than a specified value. You could also perform a one-tailed test to determine whether or not the sample mean is less than the expected value, or you could perform a two-tailed test to determine whether or not the sample 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 p-value 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.