Hypothesis Testing

Overview

Why

Hypothesis testing provides a statistical comparison of two samples. It provides an objective basis for concluding whether there is a difference between them. The procedure can be applied to two independent samples or to one sample taken from a large "whole" or constant.

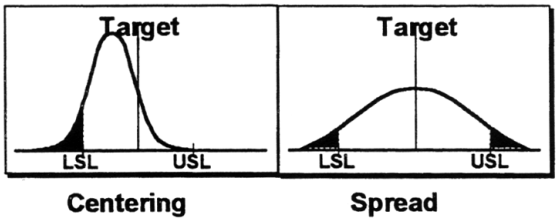
In the Analyze phase of a Six Sigma project, hypothesis testing helps compare samples in order to assess their meaning. During the Improve phase of a Six Sigma project, hypothesis testing can prove if a change you have made in the process has statistically altered it. Hopefully, the alteration will be an improvement.

How

1. The first step in hypothesis testing is to confirm that samples you test are stable, and that they have a normal distribution.
2. Set the test up. Most significantly, this involves a clear statement of the hypothesis. In addition, you need to identify the type of data available. Discrete data involves one hypothesis test, continuous data another.
3. Conduct the test, depending upon its nature. Here we will discuss four different applications, or problem statements, for hypothesis testing:

- Comparison of a sample mean against the target for the process;
- Comparison of two independent samples;
- Comparison of two dependent samples;
- Comparisons of three or more samples.

In all cases, the hypotheses are tested by comparing spread and centering of the two populations.



Setting the Test Up

A **Hypothesis** is a statement of assumptions. One can make a hypothesis about virtually anything, and one can statistically test it. For every hypothesis, there is an alternate hypothesis. Hypothesis testing is done by checking a hypothesis against its alternate. To do so, we first have to be very clear in their definition.

A **Null Hypothesis (H_o)** is a statement of the status quo. It presumes no change and no difference between the two samples. If you re comparing two machines, a null hypothesis will state there will be no difference in their operations when observed over time.

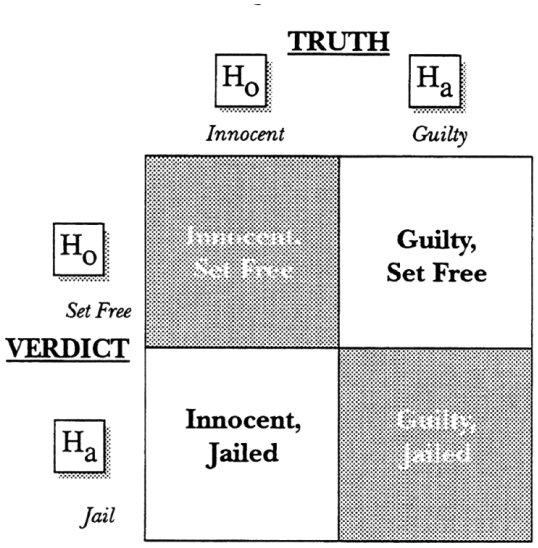
An **Alternative Hypothesis (H_a)** is a statement of difference. It is often a statement of something we want to prove. In the case of the observed operation of two machines, it states that there will be a difference.

A hypothesis test will provide the statistical basis for rejecting or not rejecting H_o. The test statistic (called a **P Value**) is used as a means of rejecting or not rejecting the null hypothesis.

Guilty or Innocent

To illustrate how the process works, and to provide some insight into the importance of a well thought out Alternate Hypothesis, we will use the American justice system. The most important presumption in justice in America is that a person accused of a crime is innocent until

proven guilty. This is a null hypothesis. It states that innocence is the norm and that proof is required before an alternative belief is accepted. That alternate belief, guilt, is our alternative hypothesis. It is to be proven or disproved by an assessment of evidence. Evidence is data. The problem with data — and with evidence—is that it does not always reflect the truth. It provides an indication of the truth, and it must be weighted carefully before making a final decision. Look at the diagram below:



Two choices can be right here, and two choices can be wrong. These are the only possible outcomes of the decision. If a jury makes a correct choice an innocent defendant will go free or a guilty one will go to jail. An incorrect choice means a guilty defendant goes free and an innocent one is jailed. Statisticians have given the two errors names. What's more, they have identified the likelihood of detecting a change.

In one instance, identified in the lower left quadrant of our four block, the error has been to jail an innocent person. Applied to the terminology of our hypotheses, the error was to accept the alternate hypothesis when, in reality, the null hypothesis was the truth.

Evidence (the data) pointed to a change in the status quo that did not, in reality exist. This type of error is called an α error. When doing a statistical test, α can be chosen. It should be chosen with full understanding of the implication of making this type of error. Typically, α is set at .05.

The other type of error (as represented in the upper right quadrant of the four block) is called a β error, A β error means that we have accepted the null hypothesis when, in fact, the alternate hypothesis was true. β is not as easy to set as α. It is dependent on both α and sample size. A typical value for β is 0.10.

The hypothesis testing we will be dealing within this module centers on α errors.

Your goal in hypothesis testing is to reject the null hypothesis. Like the prosecuting attorney in a criminal trial, that requires proof. You want to demonstrate a difference in samples.

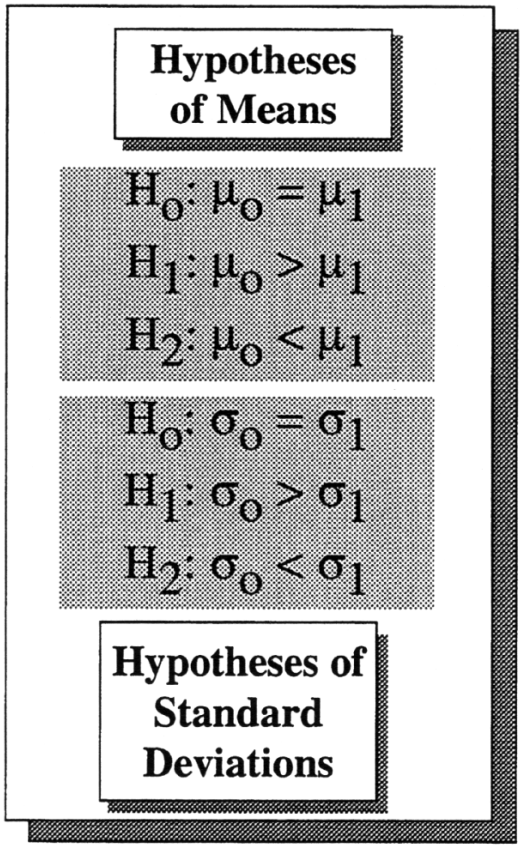
Conduct the Test

Once you have established a null and alternate hypothesis, the next step is to compare the data. Successful application of hypothesis testing is based on the following recognized statistical assumptions:

- The data are independent both within each sample and among samples;
- The sample data come from normal distributions;
- The data come from processes that are stable.

Histograms and plot charting over time will provide the answers to these questions. Refer to the module on Process Capability for how to develop a histogram and to the module on Regression Analysis for scatter plotting.

Once you are satisfied these assumptions have been met, you can test for one of the hypotheses illustrated below.



Application 1: Comparing a Single Population with a Norm

A typical question for this application might be: Does the mean for a sample meet a target for my process?

The null hypothesis for this application would be:
$H_O \Rightarrow \mu 1 = T$ where T is the target.

The alternate hypothesis is:
$H_a \Rightarrow \mu 1 \neq T$ where T is the target.

Before testing for the mean, it must first be established that the samples have the same variance (their standard deviation is equal).

The test is termed a **One Sample T-Test.** It uses a standardized **Acceptance Region** formula. The derivation of this formula and its validity are outside the scope of this discussion. Just accept it as presented.

Let's look at an example. Thirty pieces of data were collected on a grinding operation for steam turbine buckets. The mean of the sample was .96953. The standard deviation for both sample and process was .00017. The desired target for the sample was .96960. The question is whether the process is on target. An acceptable Type I error is 5%, or alpha = .05.

The null hypothesis states that the process is on target. The alternate hypothesis states that it is not.

In the formula, H_o states that μ=.96960 = T where T is the target. H_α states that $\mu \neq$. 96960. Other relevant values are:

- Sample Size = n = 30
- α = .05, α / 2 = .025. The alpha value is divided by 2 because this is a two sided test where the mean could be off in either direction to reject H_o.
- Xbar = .96953

The acceptance Region formula is:

$$UCL = X + t_{\alpha/2,\, n-1} \frac{\hat{\sigma}}{\sqrt{n}}$$

UCL equals the upper confidence level. A Target higher than this value means that the null hypothesis is rejected. Because we are also testing below the mean, a lower confidence level must be established. This is done by applying the same formula, except that we subtract the $t_{x}/_{x}$ value. This value comes from a T Distribution table (see the end of the module).

T Distribution Value

The T Distribution value is derived by considering two components. The first is the alpha value, or in this case, the alpha value divided by two. Remember, we are dividing alpha by two because we are testing on both sides of the mean. This value is .025. The second component of the T Distribution value is n-1, which is referred to as **degree of freedom (dof).** The dof value is the sample size minus 1. In our example, that is 30

$t_{\alpha 2,\, n-1}$

– 1 = 29. Look at the T Distribution table at the end of the playbook. The left hand column is labeled dof. Go down the column until you reach 29. The bold numbers at the head of the table are real alpha values. .025 is at the third column from the right. Follow this column down and the 29th row across, and you will see they intercept at 2.045. This is our T Distribution value for this example.

Complete the Equation

$$LCL = \bar{X} + t_{\alpha/2,\, n-1}\frac{\hat{\sigma}}{\sqrt{n}} \quad < T < \quad UCL = \bar{X} + t_{\alpha/2,\, n-1}\frac{\hat{\sigma}}{\sqrt{n}}$$

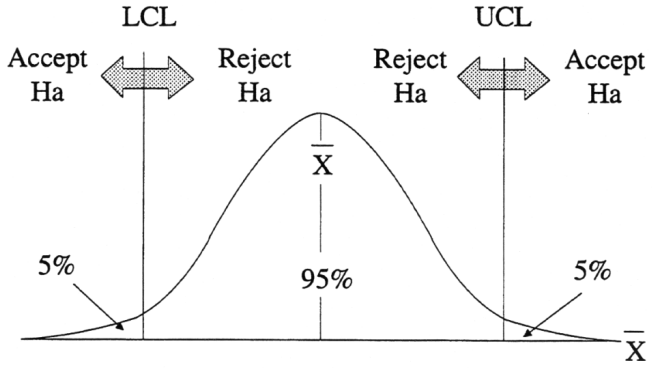
.96953 - 2.045* .00017/sq root of 30 < T <.96953 + 2.045* .00017/sq root of 30
.96946 < T < .96959

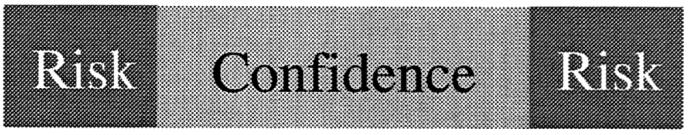
Accept or Reject the Null Hypothesis

Compare the results of the equation with the target value T = .96960. The target is NOT in the range defined by the equation, therefore $\mu_1 \neq$ T.

The process is not centered. We reject the null hypothesis. The sample does not fall within the target.

Use of the T Distribution





$$LCL = X - t_{\alpha/2,\, n-1}\frac{\hat{\sigma}}{\sqrt{n}} \quad < T < \quad UCL = X - t_{\alpha/2,\, n-1}\frac{\hat{\sigma}}{\sqrt{n}}$$

P Value

The P Value is the probability of making an α error. As noted, earlier, the generally accepted value for P is 0.05. So, for any P Value less than 0.05, we reject the null hypothesis in favor of the alternate hypothesis. Computation of this value is quite complicated, and it is best left up to a computer. Minitab is most often used and is accessible through a Black Belt.

Application 2: Comparing Two Populations

This application has two sub-categories. You may have a situation where you must compare **two independent samples,** for example eight widgets that are randomly divided into two piles of four and then subjected to the process. Another possibility is to have **two paired samples.** This would be the case if four sheets of material were each cut into two halves. Each is handled differently. With paired samples, you can safely assume that the spread of the two is the same. That is, they have the same variation, or standard deviation. With independent samples, the variation must be tested. Do not forget that before testing, make sure there is a normal distribution and stability of process.

Comparing Independent Samples

Comparing Variation

The test for variation is called the Levene's Test. Its null hypothesis is:

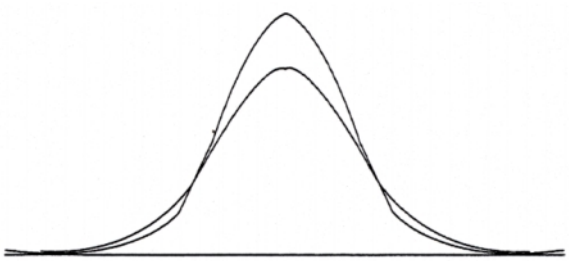
$$H_o \Rightarrow \sigma_A = \sigma_B.$$

The alternate hypothesis is:

$$H_a \Rightarrow \sigma_A \neq \sigma_B.$$

Computation is quite complicated, and it is best left up to a computer. Minitab is most often used and is accessible through a Black Belt. If the Minitab output shows a P value of less than alpha, then reject the null hypothesis.

Are the Variances the Same?



Comparing Means

The applicable test here is referred to as the **Two Sample Test for Equal Means.**

The null hypothesis for this application would be:

$$H_o \Rightarrow \mu_A = \mu_B.$$

The alternate hypothesis is:

$$H_a \Rightarrow \mu_A \neq \mu_B.$$

Another possible test is a 1-Way ANOVA in which the alternate hypothesis is $\mu_{A>}\mu_B$. That is not discussed here.

The Two Sample Test for Equal Means has some new terminology:

- **t critical** (t_c) is the maximum allowable value for T for accepting the null hypothesis.
- **degrees of freedom (dof)** is a term from the single sample test. Here it is defined differently as $n_1 + n_2 - 2$. Where n_1 is the sample number of the first sample and n_2 is the number in the second sample.

T is Test Statistic. It is a tedious calculation that looks like this:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\sqrt{\frac{(n_1 - 1)\widehat{\sigma}_1^2 + (n_2 - 1)\widehat{\sigma}_2^2}{n_1 + n_2 - 2}}}$$

Fortunately, the computer program (Minitab) will compute this for you. If the test statistic is greater than t critical, the null hypothesis is rejected.

To conduct the test, first find t critical. Use the T Distribution table to locate the value. In a case where the two samples each number 10, the dof =

(10 + 10-2) = 18. The alpha value is .025. The reason for this is the same as in the single sample test. The t critical is 2.101.

Second, calculate the test statistic. To accept the null hypothesis, the test statistic must be under this number.

Comparing Paired Samples

A **One Sampled T Test** (or Paired T-Test) is used to compare the means of the two samples. It works by comparing the differences between the two samples and eliminating the effects of other variables. For example, if you wanted to test two fan blades running on the same engine, you would sample each at the same multiple data points and take the difference in results. That difference is the Dbar.

The null hypothesis for this application would be:

$$H_o \Rightarrow \text{Dbar} = 0$$

The alternate hypothesis is:

$$H_a \Rightarrow \text{Dbar} \neq 0$$

The null hypothesis is rejected if $T \geq t_c$

The following example will illustrate how the test works:

Four engines are tested with old and new fan blades. The data is provided below. Did the new fan blades improve performance?

FAN

Engine	Old	New	Difference
1	45881.1	45862.5	18.6
2	45928.9	45972.4	-43.5
3	46011.7	45936.2	75.5
4	45944.7	45916.1	28.6

Dbar = 19.8 Ibf S_d = 48.95 Ibf

TS: T = Dbar /(S_d / sq. root of n) = 19.8 / 48.95 / 2 = .80

$t_c = t_{a/2,\ n-1} = t_{.025,\ 4-1} = 3.182$

.80 is not greater than or equal to 3.182, therefore the null hypothesis is not rejected.

Application 3: Comparing Three or More Populations

1 - Way ANOVA allows you to compare the centering of multiple populations or samples. Though ANOVA is short for "Analysis of Variance" it is really a comparison of means. You must test for variance before testing for means, just as you would if you had two independent samples—once it has been established that variance is the same test for means.

ANOVA involves arithmetically decomposing the total observed variation into two components. One component represents the response variation strictly attributable to the independent variable. The other represents the response variation strictly attributable to the independent variable. The other represents residual variation. ANOVA evaluates the probability of equal population means. If the probability exceeds a given threshold value, the alternate hypothesis of statistically significant difference is accepted. In other words, the variation in means is not a result of chance.

The null hypothesis for this application would be:

$$H_o \Rightarrow \mu_1 = \mu_2 = \mu_3 = \mu_4$$

The alternate hypothesis is:

$$H_a \Rightarrow \textbf{Any } \mu_i \neq \mu_j$$

The test begins by constructing an ANOVA table incorporating the data from all the populations. The table looks something like the example below:

The numbers in the parentheses in the table correspond to the following:

1. Factor source of variation
2. Residual or "error source of variation"
3. Total source of variation
4. Factor sum-of-squares, or the sum-of-squares which can be assigned to the factor, or "independent variable"
5. Residual or "error" sum-of-squares which can be assigned to the uncontrolled sources of variation or background noise

Generalized ANOVA Table

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (dof)	Mean Square (MS)	MS Ratio (F_{calc})	Critical MSR (F_{crit})
(1) **Between**	(4) SS_B	(7) g-1	(10) SS_B/dfg	(12) MS_B/MS_W	(13) F_{crit}
(2) **Within**	(5) SS_W	(8) g(n-1)	(11) SS_W/dfw		
(3) **Total**	(6) SS_T	(9) ng - 1			

6. Total sum-of squares, or the sum-of-squares which can be assigned to the factor and background variables
7. Between level degrees of freedom
8. Within level degrees of freedom
9. Total degrees of freedom
10. Mean square of the factor
11. Mean square of the residual
12. Mean square ratio
13. Critical F value or the threshold value

When using Minitab, the results will provide a P value. If it is a greater than alpha, do not reject the null hypothesis.

T-Distribution

dof	.400 .600	.300 .700	.200 .800	.100 .900	.050 .950	.025 .975	.010 .990	.005 .995
1	0.325	0.727	1.376	3.078	6.314	12.706	31.821	63.657
2	0.289	0.617	1.061	1.886	2.920	4.303	6.965	9.925
3	0.277	0.584	0.978	1.638	2.353	3.182	4.541	5.841
4	0.271	0.569	0.941	1.533	2.132	2.776	3.747	4.604
5	0.267	0.559	0.920	1.476	2.015	2.571	3.365	4.032
6	0.265	0.553	0.906	1.440	1.943	2.447	3.143	3.707
7	0.263	0.549	0.896	1.415	1.895	2.365	2.998	3.499
8	0.262	0.546	0.889	1.397	1.860	2.306	2.896	3.355
9	0.261	0.543	0.883	1.383	1.833	2.262	2.821	3.250
10	0.260	0.542	0.879	1.372	1.812	2.228	2.764	3.169
11	0.260	0.540	0.876	1.363	1.796	2.201	2.718	3.106
12	0.259	0.539	0.873	1.356	1.782	2.179	2.681	3.055
13	0.259	0.538	0.870	1.350	1.771	2.160	2.650	3.012
14	0.258	0.537	0.868	1.345	1.761	2.145	2.624	2.977
15	0.258	0.536	0.866	1.341	1.753	2.131	2.602	2.947
16	0.258	0.535	0.865	1.337	1.746	2.120	2.583	2.921
17	0.257	0.534	0.863	1.333	1.740	2.110	2.567	2.898
18	0.257	0.534	0.862	1.330	1.734	2.101	2.552	2.878
19	0.257	0.533	0.861	1.328	1.729	2.093	2.539	2.861
20	0.257	0.533	0.860	1.325	1.725	2.086	2.528	2.845
21	0.257	0.532	0.859	1.323	1.721	2.080	2.518	2.831
22	0.256	0.532	0.858	1.321	1.717	2.074	2.508	2.819
23	0.256	0.532	0.858	1.319	1.714	2.069	2.500	2.807
24	0.256	0.531	0.857	1.318	1.711	2.064	2.492	2.797
25	0.256	0.531	0.856	1.316	1.708	2.060	2.485	2.787
26	0.256	0.531	0.856	1.315	1.706	2.056	2.479	2.779
27	0.256	0.531	0.855	1.314	1.703	2.052	2.473	2.771
28	0.256	0.530	0.855	1.313	1.701	2.048	2.467	2.763
29	0.256	0.530	0.854	1.311	1.699	2.045	2.462	2.756
30	0.256	0.530	0.854	1.310	1.697	2.042	2.457	2.750
40	0.255	0.529	0.851	1.303	1.684	2.021	2.423	2.704
60	0.254	0.527	0.848	1.296	1.671	2.000	2.390	2.660
120	0.254	0.526	0.845	1.289	1.658	1.980	2.358	2.617
∞	0.253	0.524	0.842	1.282	1.645	1.960	2.326	2.576