WHAT IT IS

Return to Table of Contents

Hypothesis testing is a statistical method for:

- drawing inferences about a population based on sample data from such population
- assessing the statistical significance of the difference between populations on a variable of interest based on sample data from such populations
- choosing among alternative courses of action

Hypothesis testing is the basic form of statistical inference. Its objective is to determine whether or not sample data support a belief (i.e. hypothesis) about the population(s) from which the sample(s) is drawn. For example, one may want to know if a proposed policy is equally supported by men and women. After drawing a random sample of the appropriate size, one would test the hypothesis that no difference exists in the extent to which males and females support the proposed policy. Hypothesis testing is based on known and predictable properties of the distribution of the random variable which underlies the variable being sampled. The normal distribution of the sample mean is the distribution most likely to be used on audits and evaluations, though the F-distribution and χ^2 distributions may also be encountered.

DEFINITIONS

The following terms figure prominently in hypothesis testing:

- **Statistical significance** occurs when the difference between an observed sample statistic and a hypothesized population parameter is considered too great to be attributable to chance. Establishing statistical significance, or the lack thereof, is the goal of hypothesis testing.

- **Null hypothesis** is denoted by H_0 and always asserts that no difference exists between a population parameter and a single value, generally the value of the variable of interest as given by the sample. The null hypothesis is always stated as a mathematical equality. For example, if an auditor/evaluator thought that the mean dollar value (μ) of errors made by entity staff when entering data into USAS was $100, the null hypothesis used by the auditor/evaluator would be stated as:

$$H_0: \mu = 100$$

- **Alternative hypothesis** is denoted by H_A and is the assertion that answers the question about the value of a variable of interest. The alternative hypothesis is always stated as a mathematical inequality. In the case of the mean dollar value of USAS data entry errors noted above, if the auditor/evaluator thought that the mean value of such errors was greater than $100, the alternative hypothesis would be stated as:

$$H_A: \mu > 100$$

Alternatively, if the auditor thought the mean dollar value of such errors was less than $100, the alternative hypothesis would be stated as:

$$H_A: \mu < 100$$

Finally, if the auditor/evaluator thought the mean dollar value of such errors was not $100, i.e. actually either less than or greater than $100, the alternative hypothesis would be stated as:

$$H_A: \mu \neq 100$$

Thus, the alternative hypothesis specifies that the population parameter is one of the following:

— greater than the value stated in the null hypothesis
— less than the value stated in the null hypothesis
— different from the value stated in the null hypothesis

- **Test statistic** is used to determine whether the auditor/evaluator should reject or not reject the null hypothesis. (Note that one should not use the term *accept* in the context of hypothesis testing.) The test statistic is the point estimator of the population parameter being tested. It expresses the value of the variable of interest obtained from the sample in standardized form.

- **Standardized test statistic** expresses the value of the variable of interest as estimated by the sample in terms of a known and predictable probability distribution. For example, the test statistic for a population mean based on a sample mean is:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

t = test statistic
$\bar{x}$ = sample mean
μ = population mean specified in the null hypothesis
s = sample standard deviation
n = sample size

Thus, this equation sets a relationship between the difference of the observed and hypothesized values ($\bullet - \mu$) and divides by the standard deviation of the sample mean, also called the "standard error of the mean." The output of this equation is a number of standard deviations.

- **Rejection region** is the range of values such that if the test statistic falls in that range, one rejects the null hypothesis. The limits of this range are set by tables which define a "critical value" above and/or below which one should reject the null hypothesis.

  In determining the rejection region, one should use a t-table if the population variance is unknown or if the sample size is 30 or less. If the sample size is larger than 30, use the z-table. Note that at n = 29, the values on the t-table and z-table begin to converge.

- **Degrees of freedom** is denoted by d.f. and is a concept applying to certain hypothesis tests. While the definition is rather complicated, it generally refers to how free one is to set the value of a variable once an observation has been drawn. Since drawing one observation removes one degree of freedom to set a value, the degrees of freedom associated with many hypothesis tests is some variation of the expression n - 1, where n is the sample size. In fact, n - 1 is the number of degrees of freedom for the t-test. (Note that the z-test does not involve degrees of freedom since it assumes that the population variance is known.) The calculation of the number of degrees of freedom associated with a particular hypothesis test will always be specified as part of defining the rejection region.

  Under no circumstances should the number of data points minus the number of variables ever fall below 30. Optimally, this quantity should never fall below 60. (See the modules on Sampling and Questionnaires/Surveys for more information.)

- **Significance level** is the probability of rejecting a null hypothesis when it is true. It is denoted by the Greek letter α and is typically set at .10, 05, .02, or .01. Note that (1 - α = confidence level). The significance level is also the probability of a Type I error.

- **Confidence level** is the percentage of times one would expect that the sample represents the population. The confidence level of a given hypothesis test is chosen by the auditor/evaluator. It is generally set at .90, .95, .98, or .99. Note that (1 - confidence level = α).

- **Confidence interval** refers to the interval of values around the test statistic in which one expects to find the true population parameter, subject to the confidence level of the hypothesis test. Confidence intervals are generally notated as follows:

$$\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}$$

$\bar{x}$ = sample mean	s = sample standard deviation
t = test statistic	n = sample size
α = significance level	μ = population mean stated in null hypothesis

Thus, this inequality states that the population mean lies between a Lower Confidence Limit [$\bar{x}$ - (test statistic times standard error of the mean)] and an Upper Confidence Limit [$\bar{x}$ + (test statistic times standard error of the mean)]. Note that α, the significance level, figures prominently in setting this range of values in which one expects the population mean to fall.

- **Type I and Type II errors** refer to the probability of guessing wrong when performing a hypothesis test, as outlined in the following table:

	H_0 is True	**H_0 is False**
Reject H_0	Type I Error (α)	Correct Decision
Do Not Reject H_0	Correct Decision	Type II Error (β)

Thus, a Type I error is the probability of rejecting the null hypothesis when it is true and is denoted by the Greek letter α. A Type II error is the probability of not rejecting the null hypothesis when it is false and is denoted by the Greek letter β.

One wants both α and β to be as small as possible. Unfortunately, an inverse relationship exists between them such that as α gets smaller, and as the hypothesis test becomes more stringent, β gets larger.

- **One-tail hypothesis test** puts all of the significance level α at one end of a distribution. A one-tail test applies if the alternative hypothesis is either:

H_A: population parameter < X
or
H_A: population parameter > X

The first two hypotheses stated in the definition of *alternative*

hypothesis above would use a one-tail test.

- **Two-tail test** puts one-half of the significance level α at each end of a distribution. A two-tail test applies if the alternative hypothesis is:

$$H_A: \text{population parameter} \neq X$$

The last hypothesis stated in the definition of *alternative* hypothesis above would use a two-tail test.

- **p-value** is the smallest value of α that would lead to rejection of the null hypothesis. Generally, the smaller the p-value, the more likely is it that one can reject the null hypothesis.

WHEN TO USE IT

Hypothesis testing is appropriate when the goal is to test an assumption about population parameters based on samples from such populations. Hypothesis testing is often used to assess the probability that a management assertion about a population or condition is correct.

However, for hypothesis testing to work properly, one must be certain that data are relatively normally distributed. That is, the distribution of data should be:

- mound-shaped
- fairly symmetrical
- free of significant skew and kurtosis

The less these conditions hold, the less reliable are tests of hypothesis and, thereby, the conclusions one may draw from such tests.

HOW TO PREPARE IT

The ordered steps of hypothesis testing follow below:

- Determine the null hypothesis H_0 and the alternative hypothesis H_A.
- Specify the test statistic.
- Specify α, and set up the rejection region.
- Calculate the test statistic.
- State conclusions of the hypothesis test and interpret the results.

SAMPLE APPLICATION

A program manager asserts that a newly enacted Federal Government reporting requirement makes it impossible for her staff to process as many job applications as in previous years. She says that in previous years, her staff was able to process an average of 100 such applications per week. A random sample is taken of 15 weeks' output during the period following enactment of the new reporting requirement. This sample yields the data listed below. Note that $n < 30$ to keep calculations relatively simple:

93, 103, 95, 101, 91, 105, 96, 94, 101, 88, 98, 94, 101, 92, 95

Assuming that weekly output of applications is normally distributed, does

sufficient evidence exist to conclude that staff productivity has statistically and significantly decreased following promulgation of the new regulation?

- Determine the null hypothesis H_0 and the alternative hypothesis H_A:
  - The null hypothesis H_0 is that average output has not changed, i.e.

$$H_0: \mu = 100$$

  - The alternative hypothesis is that average output has gone down, i.e.

$$H_A: \mu < 100$$

- Specify the test statistic. Since the sample size is less than 30 and the population variance is not known, the test statistic is the t-statistic, as follows:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

- Specify α, and set up the rejection region:
  - Here α is arbitrarily set at .05. Thus, the significance level is .05 and the confidence level is $(1 - \alpha)$ or .95
  - The rejection region is:

$$t < -t_{\alpha,d.f.}$$

Note that this quantity will be negative since the alternative hypothesis tests the assertion that the true population parameter, average weekly applications processed, is less than 100, i.e. toward the negative end of the distribution. Since α = .05 and d.f. = (n - 1) = (15 - 1) =14, this expression becomes:

$$t < -t_{.05,14.}$$

The t-table states that the value of this expression is -1.761. Thus, if the value of the test statistic is less than -1.761, one should reject the null hypothesis.

- Calculate the test statistic. To do this, one must know that the sample mean ($\bar{x}$) is 96.47, the sample size (n) is 15, and that s, the sample standard deviation, is 4.85. (See the module on Descriptive Statistics for more information on how to calculate these measures.) Given these numbers, the value for the test statistic is:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{96.47 - 100}{\frac{4.85}{\sqrt{15}}} = -2.82$$

- State conclusions of the hypothesis test and interpret the results. Since -2.82 is less than -1.761, one should reject the null hypothesis and state that sufficient evidence exists to conclude that the average number of job applications processed per week has decreased during the time of the new Federal Government reporting requirement. In this instance, it would be wise to use a similar technique to check the program manager's assertion that the average weekly output from previous years was, in fact, 100 applications per week since 100 is the standard used in the hypothesis test above.

A NOTE ON DRAWING CONCLUSIONS AND INTERPRETING RESULTS

Ultimately, only two possible conclusions exist for a hypothesis test:

- Reject the null hypothesis, in which case one concludes that sufficient evidence exists to indicate that the alternative hypothesis is true.
- Do not reject the null hypothesis, in which case one concludes that insufficient evidence exists to indicate that the alternative hypothesis is true.

Thus, the null hypothesis can be rejected or not rejected. As noted earlier, one should not say *accepted* instead of *not rejected*. More definitive conclusions can be drawn if the null hypothesis is rejected. The conclusion would be that the null hypothesis is false, at a particular level of confidence, i.e. beyond a reasonable doubt.

If a null hypothesis is not rejected, the same degree of certainty is not obtained. One can assume that null hypothesis is reasonable, but the level of certainty could be specified only if one knows the true population parameter, the mean weekly output in the sample application above. This is rarely possible to obtain since surveying entire populations is expensive and time-consuming. Nonetheless, one way to judge the accuracy of a non-rejected null hypothesis is to calculate the confidence interval around the estimated population parameter. As a rule, the wider this interval, the less certainty exists about the conclusion.

Thus, in final analysis, the probability of Type I and Type II errors prevent hypothesis tests from being 100 percent accurate. While this possibility of error exists, it is fairly common practice to assume a hypothesis is true if it is not

rejected. Still, care should always be taken when drawing conclusions from and stating inferences based on hypothesis tests.

ADVANTAGES

- Quantitative means of testing management assertions
- Useful in extrapolating from sample to population

DISADVANTAGES

- Careful interpretation is necessary
- Nature of test allows remaining uncertainty