Statistical hypothesis testing

(From Wikipedia)

A statistical hypothesis is a scientific hypothesis that is testable on the basis of observing a process that is modeled via a set of random variables.
A test result is called *statistically significant* if it has been predicted as unlikely to have occurred by sampling error alone, according to a threshold probability—the significance level. Hypothesis tests are used in determining what outcomes of a study would lead to a rejection of the null hypothesis for a pre-specified level of significance. In the Neyman-Pearson framework, the process of distinguishing between the null hypothesis and the alternative hypothesis is aided by identifying two conceptual types of errors (type 1 & type 2), and by specifying parametric limits on e.g. how much type 1 error will be permitted.

An alternative framework for statistical hypothesis testing is to specify a set of statistical models, one for each candidate hypothesis, and then use model selection techniques to choose the most appropriate model. The most common selection techniques are based on either Akaike information criterion or Bayes factor.

Statistical hypothesis testing is sometimes called confirmatory data analysis. It can be contrasted with exploratory data analysis, which may not have pre-specified hypotheses.

Variations and sub-classes

Statistical hypothesis testing is a key technique of both Frequentist inference and Bayesian inference. Statistical hypothesis tests define a procedure that controls (fixes) the probability of incorrectly *deciding* that a default position (null hypothesis) is incorrect. The procedure is based on how likely it would be for a set of observations to occur if the null hypothesis were true. There are other possible techniques of decision theory in which the null and alternative hypothesis are treated on a more equal basis.

One naive Bayesian approach to hypothesis testing is to base decisions on the posterior probability, but this fails when comparing point and continuous hypotheses. Other approaches to decision making, such as Bayesian decision theory, attempt to balance the consequences of incorrect decisions across all possibilities, rather than concentrating on a single null hypothesis. A number of other approaches to reaching a decision based on data are available. Hypothesis testing, though, is a dominant approach to data analysis in many fields of science. Extensions to the theory of hypothesis testing include the study of the power of tests, i.e. the probability of correctly rejecting the null hypothesis given that it is false.

The testing process

A commonly used process is:

1. There is an initial research hypothesis of which the truth is unknown.
2. The first step is to state the relevant **null and alternative hypotheses**. This is important as mis-stating the hypotheses will muddy the rest of the process.
3. Compute from the observations the observed value t_{obs} of the test statistic T.
4. Calculate the p-value. This is the probability, under the null hypothesis, of sampling a test statistic at least as extreme as that which was observed.
5. Reject the null hypothesis, in favor of the alternative hypothesis, if and only if the p-value is less than the significance level (the selected probability) threshold.

This process relies on extensive tables or on computational support. The explicit calculation of a probability is useful for reporting. The calculations are performed with appropriate software. Previously, some more elaborate process of manual analysis was followed.

Interpretation

If the *p*-value is less than the required significance level (equivalently, if the observed test statistic is in the critical region), then we say the null hypothesis is rejected at the given level of significance. Rejection of the null hypothesis is a conclusion. This is like a "guilty" verdict in a criminal trial: the evidence is sufficient to reject innocence, thus proving guilt. We might accept the alternative hypothesis (and the research hypothesis).

If the *p*-value is *not* less than the required significance level (equivalently, if the observed test statistic is outside the critical region), then the test has no result. The evidence is insufficient to support a conclusion. (This is like a jury that fails to reach a verdict.) The researcher typically gives extra consideration to those cases where the *p*-value is close to the significance level.

Whether rejection of the null hypothesis truly justifies acceptance of the research hypothesis depends on the structure of the hypotheses. Rejecting the hypothesis that a large paw print originated from a bear does not immediately prove the existence of Bigfoot. *Hypothesis testing emphasizes the rejection, which is based on a probability, rather than the acceptance, which requires extra steps of logic.*

"The probability of rejecting the null hypothesis is a function of five factors:

- whether the test is one- or two tailed,
- the level of significance,
- the standard deviation,
- the amount of deviation from the null hypothesis,
- and the number of observations."

These factors are a source of criticism; factors under the control of the experimenter/analyst give the results an appearance of subjectivity.

Use and importance

Statistics in hypothesis testing can justify conclusions even when no scientific theory exists. The data may contradict the "obvious".

Real world applications of hypothesis testing include:

- Testing whether more men than women suffer from nightmares
- Establishing authorship of documents
- Evaluating the effect of the full moon on behavior
- Determining the range at which a bat can detect an insect by echo
- Deciding whether hospital carpeting results in more infections
- Selecting the best means to stop smoking
- Checking whether bumper stickers reflect car owner behavior
- Testing the claims of handwriting analysts

Statistical hypothesis testing plays an important role in the whole of statistics and in statistical inference.

At the core of the scientific method is comparison of predicted value and experimental result. When theory is only capable of predicting the sign of a relationship, a directional (one-sided) hypothesis test can be configured so that only a statistically significant result supports theory. This form of theory appraisal is a heavily criticized application of hypothesis testing.

Cautions

The successful hypothesis test is associated with a probability and a type-I error rate. The conclusion *might* be wrong. The conclusion of the test is only as solid as the sample upon which it is based. The design of the experiment is critical. A number of unexpected effects have been observed including:

- The Clever Hans effect. A horse appeared to be capable of doing simple arithmetic.
- The Hawthorne effect. Industrial workers were more productive in better illumination, and most productive in worse.
- The Placebo effect. Pills with no medically active ingredients were remarkably effective.

A statistical analysis of misleading data produces misleading conclusions (Example – Ban Bread). The issue of data quality can be more subtle. Many claims are made on the basis of samples too small to convince. If a report does not mention sample size, be doubtful.

Hypothesis testing acts as a filter of statistical conclusions; only those results meeting a probability threshold are publishable. Economics also acts as a publication filter; only those results favorable to the author and funding source may be submitted for publication. The impact of filtering on publication is termed publication bias. A related problem is that of multiple testing (sometimes linked to data mining), in which a variety of tests for a variety of possible effects are applied to a single data set and only those yielding a significant result are reported.

Those making critical decisions based on the results of a hypothesis test are prudent to look at the details rather than the conclusion alone. In the physical sciences most results are fully accepted only when independently confirmed. The general advice concerning statistics is, "Figures never lie, but liars figure" (anonymous).

An Analogy - Courtroom trial

A statistical test procedure is comparable to a criminal trial; a defendant is considered not guilty as long as his or her guilt is not proven. The prosecutor tries to prove the guilt of the defendant. Only when there is enough charging evidence the defendant is convicted.

In the start of the procedure, there are two hypotheses H_0: "the defendant is not guilty", and H_1 : "the defendant is guilty". The first one is called *null hypothesis*, and is for the time being accepted. The second one is called *alternative (hypothesis)*. It is the hypothesis one hopes to support.

The hypothesis of innocence is only rejected when an error is very unlikely, because one doesn't want to convict an innocent defendant. Such an error is called *error of the first kind* (i.e., the conviction of an innocent person), and the occurrence of this error is controlled to be rare. As a consequence of this asymmetric behaviour, the *error of the second kind* (acquitting a person who committed the crime), is often rather large.

	H_0 is true Truly not guilty	**H_1 is true Truly guilty**
Accept Null Hypothesis Acquittal	Right decision	Wrong decision Type II Error
Reject Null Hypothesis Conviction	Wrong decision Type I Error	Right decision

A criminal trial can be regarded as either or both of two decision processes: guilty vs not guilty or evidence vs a threshold ("beyond a reasonable doubt"). In one view, the defendant is judged; in the other view the performance of the prosecution (which bears the burden of proof) is judged. A hypothesis test can be regarded as either a judgment of a hypothesis or as a judgment of evidence.

Terms-

Statistical test
: A procedure whose inputs are samples and whose result is a hypothesis.

Region of acceptance
: The set of values of the test statistic for which we fail to reject the null hypothesis.

Region of rejection / Critical region
: The set of values of the test statistic for which the null hypothesis is rejected.

Critical value
: The threshold value delimiting the regions of acceptance and rejection for the test statistic.

Size
: For simple hypotheses, this is the test's probability of *incorrectly* rejecting the null hypothesis. The false positive rate. For composite hypotheses this is the supremum of the probability of rejecting the

null hypothesis over all cases covered by the null hypothesis. The complement of the false positive rate is termed **specificity** in biostatistics.

Significance level of a test (α)

It is the upper bound imposed on the size of a test. Its value is chosen by the statistician prior to looking at the data or choosing any particular test to be used. It the maximum exposure to erroneously rejecting H_0 he/she is ready to accept. Testing H_0 at significance level α means testing H_0 with a test whose size does not exceed α. In most cases, one uses tests whose size is equal to the significance level.

p-value

The probability, assuming the null hypothesis is true, of observing a result at least as extreme as the test statistic.

Statistical significance test

A predecessor to the statistical hypothesis test. An experimental result was said to be statistically significant if a sample was sufficiently inconsistent with the (null) hypothesis. This was variously considered common sense, a pragmatic heuristic for identifying meaningful experimental results, a convention establishing a threshold of statistical evidence or a method for drawing conclusions from data. The statistical hypothesis test added mathematical rigor and philosophical consistency to the concept by making the alternative hypothesis explicit.

Conservative test

A test is conservative if, when constructed for a given nominal significance level, the true probability of *incorrectly* rejecting the null hypothesis is never greater than the nominal level.

Exact test

A test in which the significance level or critical value can be computed exactly, i.e., without any approximation. In some contexts this term is restricted to tests applied to categorical data and to permutation tests, in which computations are carried out by complete enumeration of all possible outcomes and their probabilities.

A statistical hypothesis test compares a test statistic (*z* or *t* for examples) to a threshold. The test statistic (the formula found in the table below) is based on optimality. For a fixed level of Type I error rate, use of these statistics minimizes Type II error rates (equivalent to maximizing power). The following terms describe tests in terms of such optimality:

Most powerful test

For a given *size* or *significance level*, the test with the greatest power (probability of rejection) for a given value of the parameter(s) being tested, contained in the alternative hypothesis.

Uniformly most powerful test (UMP)

A test with the greatest *power* for all values of the parameter(s) being tested, contained in the alternative hypothesis.

Common test statistics

One-sample tests are appropriate when a sample is being compared to the population from a hypothesis. The population characteristics are known from theory or are calculated from the population.

Two-sample tests are appropriate for comparing two samples, typically experimental and control samples from a scientifically controlled experiment.

Paired tests are appropriate for comparing two samples where it is impossible to control important variables. Rather than comparing two sets, members are paired between samples so the difference between the members becomes the sample. Typically the mean of the differences is then compared to zero. The common example scenario for when a paired difference test is appropriate is when a single set of test subjects has something applied to them and the test is intended to check for an effect.

Z-tests are appropriate for comparing means under stringent conditions regarding normality and a known standard deviation.

A *t-test* is appropriate for comparing means under relaxed conditions (less is assumed).

Tests of proportions are analogous to tests of means (the 50% proportion).

Chi-squared tests use the same calculations and the same probability distribution for different applications:

- *Chi-squared tests* for variance are used to determine whether a normal population has a specified variance. The null hypothesis is that it does.
- *Chi-squared tests* of independence are used for deciding whether two variables are associated or are independent. The variables are categorical rather than numeric. The null hypothesis is that the variables are independent. The numbers used in the calculation are the observed and expected frequencies of occurrence.
- *Chi-squared goodness of fit tests* are used to determine the adequacy of curves fit to data. The null hypothesis is that the curve fit is adequate. It is common to determine curve shapes to minimize the mean square error, so it is appropriate that the goodness-of-fit calculation sums the squared errors.

F-tests (analysis of variance, ANOVA) are commonly used when deciding whether groupings of data by category are meaningful. *Eg.* if the variance of test scores of the left-handed in a class is much smaller than the variance of the whole class, then it may be useful to study lefties as a group. The null hypothesis is that two variances are the same – so the proposed grouping is not meaningful.