Statistical hypothesis testing:

A statistical hypothesis, sometimes called confirmatory data analysis, is an hypothesis that is testable on the basis of observing a process that is modeled via a set of random variables. A statistical hypothesis test is a method of statistical inference. Commonly, two statistical data sets are compared, or a data set obtained by sampling is compared against a synthetic data set from an idealized model. A hypothesis is proposed for the statistical relationship between the two data sets, and this is compared as an alternative to an idealized null hypothesis that proposes no relationship between two data sets. The comparison is deemed statistically significant if the relationship between the data sets would be an unlikely realization of the null hypothesis 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.

The process of distinguishing between the null hypothesis and the alternative hypothesis is aided by identifying two conceptual types of errors, type 1 and 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.

Confirmatory data analysis can be contrasted with exploratory data analysis, which may not have pre-specified hypotheses.

The testing process:

In the statistics literature, statistical hypothesis testing plays a fundamental role. The usual line of reasoning is as follows:

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. The second step is to consider the statistical assumptions being made about the sample in doing the test; for example, assumptions about the statistical independence or about the form of the distributions of the observations. This is equally important as invalid assumptions will mean that the results of the test are invalid.

4. Decide which test is appropriate, and state the relevant test statistic *T*.

5. Derive the distribution of the test statistic under the null hypothesis from the assumptions. In standard cases this will be a well-known result. For example, the test statistic might follow a Student's t distribution or a normal distribution.

6. Select a significance level (α), a probability threshold below which the null hypothesis will be rejected. Common values are 5% and 1%.

7. The distribution of the test statistic under the null hypothesis partitions the possible values of T into those for which the null hypothesis is rejected—the so-called critical region—and those for which it is not. The probability of the critical region is α.

8. Compute from the observations the observed value t_{obs} of the test statistic T.

Decide to either reject the null hypothesis in favor of the alternative or not reject it. The decision rule is to reject the null hypothesis $H0$ if the observed value t_{obs} is in the critical region, and to accept or "fail to reject" the hypothesis otherwise.

An alternative process is commonly used:

1. Compute from the observations the observed value $t$obs of the test statistic T.

2. 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.

3. 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.

The two processes are equivalent. The former process was advantageous in the past when only tables of test statistics at common probability thresholds were available. It allowed a decision to be made without the calculation of a probability. It was adequate for classwork and for operational use, but it was deficient for reporting results. The latter process relied on extensive tables or on computational support not always available. The explicit calculation of a probability is useful for reporting. The calculations are now trivially performed with appropriate software.

The difference in the two processes applied to the radioactive suitcase example (below):

1. The Geiger-counter reading is 10. The limit is 9. Check the suitcase."

2. "The Geiger-counter reading is high; 97% of safe suitcases have lower readings. The limit is 95%. Check the suitcase."

The former report is adequate, the latter gives a more detailed explanation of the data and the reason why the suitcase is being checked.

It is important to note the difference between accepting the null hypothesis and simply failing to reject it. The "fail to reject" terminology highlights the fact that the null hypothesis is assumed to be true from the start of the test; if there is a lack of evidence against it, it simply continues to be assumed true. The phrase "accept the null hypothesis" may suggest it has been proved simply because it has not been disproved, a logical fallacy known as the argument from ignorance. Unless a test with particularly high power is used, the idea of "accepting" the null hypothesis may be dangerous. Nonetheless the terminology is prevalent throughout statistics, where the meaning actually intended is well understood. The processes described here are perfectly adequate for computation. They seriously neglect the design of experiments considerations. It is particularly critical that appropriate sample sizes be estimated before conducting the experiment. The phrase "test of significance" was coined by statistician Ronald Fisher.

Null hypothesis

In inferential statistics, the null hypothesis is a general statement or default position that there is no relationship between two measured phenomena, or no association among groups. Testing (accepting, approving, rejecting, or disproving) the null hypothesis—and thus concluding that there are or are not grounds for believing that there is a relationship between two phenomena (e.g. that a potential treatment has a measurable effect)—is a central task in the modern practice of science; the field of statistics gives precise criteria for rejecting a null hypothesis. The null hypothesis is generally assumed to be true until evidence indicates otherwise. In statistics, it is often denoted H_0 (read "H-nought", "H-null", "H-oh", or "H-zero"): $H_0: \mu_1 = \mu_2$.

Alternative hypothesis

In statistical hypothesis testing, the alternative hypothesis (or maintained hypothesis or research hypothesis) and the null hypothesis are the two rival hypotheses which are compared by a statistical hypothesis test.

In the domain of science two rival hypotheses can be compared by explanatory power and predictive power. $H_A \; or \; H_1: \mu \neq \mu_0$

In the case of a scalar parameter, there are four principal types of alternative hypothesis:

1. Point. Point alternative hypotheses occur when the hypothesis test is framed so that the population distribution under the alternative hypothesis is a fully defined distribution, with no unknown parameters; such hypotheses are usually of no practical interest but are fundamental to theoretical considerations of statistical inference and are the basis of the Neyman–Pearson lemma.

2. One-tailed directional. A one-tailed directional alternative hypothesis is concerned with the region of rejection for only one tail of the sampling distribution.

3. Two-tailed directional. A two-tailed directional alternative hypothesis is concerned with both regions of rejection of the sampling distribution.

4. Non-directional. A non-directional alternative hypothesis is not concerned with either region of rejection, but, rather, it is only concerned that null hypothesis is not true.

Statistical significance

In statistical hypothesis testing, a result has statistical significance when it is very unlikely to have occurred given the null hypothesis. More precisely, a study's defined significance level, α,is the probability of the study rejecting the null hypothesis, given that it were true; and the p-value of a result, p, is the probability of obtaining a result at least as extreme, given that the null hypothesis were true. The result is statistically significant, by the standards of the study, when $P < \alpha$. The significance level for a study is chosen before data collection, and typically set to 5% or much lower, depending on the field of study.

In any experiment or observation that involves drawing a sample from a population, there is always the possibility that an observed effect would have occurred due to sampling error alone. But if the p-value of an observed effect is less than the significance level, an investigator may conclude that the effect reflects the characteristics of the whole population, thereby rejecting the null hypothesis. This technique for testing the statistical significance of results was used in genetics as far back as the 18th century, and entered widespread use in other fields in the early 20th century. The term significance does not imply importance here, and the term statistical significance is not the same as research, theoretical, or practical significance. For example, the term clinical significance refers to the practical importance of a treatment effect.In below normal distribution curve for a two-tailed test, the rejection region for a significance level of α=0.05 is partitioned to both ends of the sampling distribution and makes up 5% of the area under the curve (white areas).