Statistical hypothesis testing

A **statistical hypothesis** is a hypothesis that is testable on the basis of observed data modelled as the realised values taken by a collection of random variables.[1] A set of data is modelled as being realised values of a collection of random variables having a joint probability distribution in some set of possible joint distributions. The hypothesis being tested is exactly that set of possible probability distributions. A **statistical hypothesis test** is a method of statistical inference. An alternative hypothesis is proposed for the probability distribution of the data, either explicitly or only informally. The comparison of the two models is deemed *statistically significant* if, according to a threshold probability—the significance level—the data would be unlikely to occur if the null hypothesis were true. A hypothesis test specifies which outcomes of a study may lead to a rejection of the null hypothesis at a pre-specified level of significance, while using a pre-chosen measure of deviation from that hypothesis (the test statistic, or goodness-of-fit measure). The pre-chosen level of significance is the maximal allowed "false positive rate". One wants to control the risk of incorrectly rejecting a true null hypothesis.

The process of distinguishing between the null hypothesis and the alternative hypothesis is aided by considering two conceptual types of errors. The first type of error occurs when the null hypothesis is wrongly rejected. The second type of error occurs when the null hypothesis is wrongly not rejected. (The two types are known as type 1 and type 2 errors.)

Hypothesis tests based on statistical significance are another way of expressing confidence intervals (more precisely, confidence sets). In other words, every hypothesis test based on significance can be obtained via a confidence interval, and every confidence interval can be obtained via a hypothesis test based on significance.[2]

Significance-based hypothesis testing is the most common framework for statistical hypothesis testing. 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.[3] The most common selection techniques are based on either Akaike information criterion or Bayes factor. However, this is not really an "alternative framework", though one can call it a more complex framework. It is a situation in which one likes to distinguish between many possible hypotheses, not just two. Alternatively, one can see it as a hybrid between testing and estimation, where one of the parameters is discrete, and specifies which of a hierarchy of more and more complex models is correct.

- Null hypothesis significance testing* is the name for a version of hypothesis testing with no explicit mention of possible alternatives, and not much consideration of error rates. It was championed by Ronald Fisher in a context in which he downplayed any explicit choice of alternative hypothesis and consequently paid no attention to the power of a test. One simply set up a null hypothesis as a kind of straw man, or more kindly, as a formalisation of a standard, establishment, default idea of how things were. One tried to overthrow this conventional view by showing that it led to the conclusion that something extremely unlikely had happened, thereby discrediting the theory.

Contents

The testing process

In the statistics literature, statistical hypothesis testing plays a fundamental role.[4] There are two mathematically equivalent processes that can be used.[5]

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 with known degrees of freedom, or a normal distribution with known mean and variance. If the distribution of the test statistic is completely fixed by the null hypothesis we call the hypothesis simple, otherwise it is called composite.
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. In the case of a composite null hypothesis, the maximal probability of the critical region is α.
8. Compute from the observations the observed value t_{obs} of the test statistic T.
9. 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 H_0 if the observed value t_{obs} is in the critical region, and to accept or "fail to reject" the hypothesis otherwise.

A common alternative formulation of this process goes as follows:

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 (the maximal probability of that event, if the hypothesis is composite).
3. Reject the null hypothesis, in favor of the alternative hypothesis, if and only if the *p*-value is less than (or equal to) the significance level (the selected probability) threshold (α).

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):

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

The difference between accepting the null hypothesis and simply failing to reject it is important. The "fail to reject" terminology highlights the fact that the a non-significant result provides no way to determine which of the two hypotheses is true, so all that can be concluded is that the null hypothesis has not been rejected. 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 is likely to be incorrect. 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.[6][7]

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.[8]

Interpretation

The *p*-value is the probability that a given result (or a more significant result) would occur under the null hypothesis (or in the case of a composite null, it is the largest such probability; see Chapter 10 of "All of Statistics: A Concise Course in Statistical Inference", Springer; 1st Corrected ed. 20 edition, September 17, 2004; Larry Wasserman). For example, say that a fair coin is tested for fairness (the null hypothesis). At a significance level of 0.05, the fair coin would be expected to (incorrectly) reject the null hypothesis in about 1 out of every 20 tests. The *p*-value does not provide the probability that either hypothesis is correct (a common source of confusion).[9]

If the *p*-value is less than the chosen significance threshold (equivalently, if the observed test statistic is in the critical region), then we say the null hypothesis is rejected at the chosen 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 chosen significance threshold (equivalently, if the observed test statistic is outside the critical region), then the evidence is insufficient to support a conclusion. (This is similar to a "not guilty" verdict.) The researcher typically gives extra consideration to those cases where the *p*-value is close to the significance level.

Some people find it helpful to think of the hypothesis testing framework as analogous to a mathematical proof by contradiction.[10]

In the Lady tasting tea example (below), Fisher required the Lady to properly categorize all of the cups of tea to justify the conclusion that the result was unlikely to result from chance. His test revealed that if the lady was effectively guessing at random (the null hypothesis), there was a 1.4% chance that the observed results (perfectly ordered tea) would occur.

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."[11] 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 are helpful in analyzing most collections of data. This is equally true of hypothesis testing which can justify conclusions even when no scientific theory exists. In the Lady tasting tea example, it was "obvious" that no difference existed between (milk poured into tea) and (tea poured into milk). The data contradicted the "obvious".

Real world applications of hypothesis testing include:[12]

- 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. For example, Lehmann (1992) in a review of the fundamental paper by Neyman and Pearson (1933) says: "Nevertheless, despite their shortcomings, the new paradigm formulated in the 1933 paper, and the many developments carried out within its framework continue to play a central role in both the theory and practice of statistics and can be expected to do so in the foreseeable future".

Significance testing has been the favored statistical tool in some experimental social sciences (over 90% of articles in the *Journal of Applied Psychology* during the early 1990s).[13] Other fields have favored the estimation of parameters (e.g. effect size). Significance testing is used as a substitute for the traditional comparison of predicted value and experimental result at the core of the scientific method. 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 the most heavily criticized application of hypothesis testing.

Cautions

"If the government required statistical procedures to carry warning labels like those on drugs, most inference methods would have long labels indeed."[14] This caution applies to hypothesis tests and alternatives to them.

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. The issue of data quality can be more subtle. In forecasting for example, there is no agreement on a measure of forecast accuracy. In the absence of a consensus measurement, no decision based on measurements will be without controversy.

The book *How to Lie with Statistics*[15][16] is the most popular book on statistics ever published.[17] It does not much consider hypothesis testing, but its cautions are applicable, including: 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. These are often dealt with by using multiplicity correction procedures that control the family wise error rate (FWER) or the false discovery rate (FDR).

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).

Examples

Human sex ratio

The earliest use of statistical hypothesis testing is generally credited to the question of whether male and female births are equally likely (null hypothesis), which was addressed in the 1700s by John Arbuthnot (1710),[18] and later by Pierre-Simon Laplace (1770s).[19]

Arbuthnot examined birth records in London for each of the 82 years from 1629 to 1710, and applied the sign test, a simple non-parametric test.[20][21][22] In every year, the number of males born in London exceeded the number of females. Considering more male or more female births as equally likely, the probability of the observed outcome is 0.5^{82}, or about 1 in 4,8360,0000,0000,0000,0000,0000; in modern terms, this is the *p*-value. Arbuthnot concluded that this is too small to be due to chance and must instead be due to divine providence: "From whence it follows, that it is Art, not Chance, that governs." In modern terms, he rejected the null hypothesis of equally likely male and female births at the $p = 1/2^{82}$ significance level.

Laplace considered the statistics of almost half a million births. The statistics showed an excess of boys compared to girls.[23][24] He concluded by calculation of a *p*-value that the excess was a real, but unexplained, effect.[25]

Lady tasting tea

In a famous example of hypothesis testing, known as the *Lady tasting tea*,[26] Dr. Muriel Bristol, a female colleague of Fisher claimed to be able to tell whether the tea or the milk was added first to a cup. Fisher proposed to give her eight cups, four of each variety, in random order. One could then ask what the probability was for her getting the number she got correct, but just by chance. The null hypothesis was that the Lady had no such ability. The test statistic was a simple count of the number of successes in selecting the 4 cups. The critical region was the single case of 4 successes of 4 possible based on a conventional probability criterion (< 5%). A pattern of 4 successes corresponds to 1 out of 70 possible combinations (p≈ 1.4%). Fisher asserted that no alternative hypothesis was (ever) required. The lady correctly identified every cup,[27] which would be considered a statistically significant result.

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 evidence for the prosecution is the defendant 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, H_0, is called the *null hypothesis*, and is for the time being accepted. The second one, H_1, is called the *alternative hypothesis*. It is the alternative hypothesis that one hopes to support.

The hypothesis of innocence is rejected only 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, an *error of the second kind* (acquitting a person who committed the crime), is more common.

	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.

Philosopher's beans

The following example was produced by a philosopher describing scientific methods generations before hypothesis testing was formalized and popularized.[28]

> Few beans of this handful are white.
> Most beans in this bag are white.
> Therefore: Probably, these beans were taken from another bag.
> This is an hypothetical inference.

The beans in the bag are the population. The handful are the sample. The null hypothesis is that the sample originated from the population. The criterion for rejecting the null-hypothesis is the "obvious" difference in appearance (an informal difference in the mean). The interesting result is that consideration of a real population and a real sample produced an imaginary bag. The philosopher was considering logic rather than probability. To be a real statistical hypothesis test, this example requires the formalities of a probability calculation and a comparison of that probability to a standard.

A simple generalization of the example considers a mixed bag of beans and a handful that contain either very few or very many white beans. The generalization considers both extremes. It requires more calculations and more comparisons to arrive at a formal answer, but the core philosophy is unchanged; If the composition of the handful is greatly different from that of the bag, then the sample probably originated from another bag. The original example is termed a one-sided or a one-tailed test while the generalization is termed a two-sided or two-tailed test.

The statement also relies on the inference that the sampling was random. If someone had been picking through the bag to find white beans, then it would explain why the handful had so many white beans, and also explain why the number of white beans in the bag was depleted (although the bag is probably intended to be assumed much larger than one's hand).

Clairvoyant card game

A person (the subject) is tested for clairvoyance. They are shown the reverse of a randomly chosen playing card 25 times and asked which of the four suits it belongs to. The number of hits, or correct answers, is called *X*.

As we try to find evidence of their clairvoyance, for the time being the null hypothesis is that the person is not clairvoyant.[29] The alternative is: the person is (more or less) clairvoyant.

If the null hypothesis is valid, the only thing the test person can do is guess. For every card, the probability (relative frequency) of any single suit appearing is 1/4. If the alternative is valid, the test subject will predict the suit correctly with probability greater than 1/4. We will call the probability of guessing correctly *p*. The hypotheses, then, are:

- null hypothesis **:** $H_0 : p = \frac{1}{4}$ (just guessing)

and

- alternative hypothesis **:** $H_1 : p > \frac{1}{4}$ (true clairvoyant).

When the test subject correctly predicts all 25 cards, we will consider them clairvoyant, and reject the null hypothesis. Thus also with 24 or 23 hits. With only 5 or 6 hits, on the other hand, there is no cause to consider them so. But what about 12 hits, or 17 hits? What is the critical number, *c*, of hits, at which point we consider the subject to be clairvoyant? How do we determine the critical value *c*? With the choice *c*=25 (i.e. we only accept clairvoyance when all cards are predicted correctly) we're more critical than with *c*=10. In the first case almost no test subjects will be recognized to be clairvoyant, in the second case, a certain number will pass the test. In practice, one decides how critical one will be. That is, one decides how often one accepts an error of the first kind – a false positive, or Type I error. With *c* = 25 the probability of such an error is:

$$P(\text{reject } H_0 \mid H_0 \text{ is valid}) = P(X = 25 \mid p = \tfrac{1}{4}) = \left(\tfrac{1}{4}\right)^{25} \approx 10^{-15},$$

and hence, very small. The probability of a false positive is the probability of randomly guessing correctly all 25 times.

Being less critical, with *c*=10, gives:

$$P(\text{reject } H_0 \mid H_0 \text{ is valid}) = P(X \geq 10 \mid p = \tfrac{1}{4}) = \sum_{k=10}^{25} P(X = k \mid p = \tfrac{1}{4}) = \sum_{k=10}^{25} C(25,k)(1-\tfrac{1}{4})^{(25-k)}(\tfrac{1}{4})^{k} \approx 0.0713,$$

(where C(25,k) is the binomial coefficient 25 choose k). Thus, *c* = 10 yields a much greater probability of false positive.

Before the test is actually performed, the maximum acceptable probability of a Type I error (α) is determined. Typically, values in the range of 1% to 5% are selected. (If the maximum acceptable error rate is zero, an infinite number of correct guesses is required.) Depending on this Type 1 error rate, the critical value *c* is calculated. For example, if we select an error rate of 1%, *c* is calculated thus:

$$P(\text{reject } H_0 \mid H_0 \text{ is valid}) = P(X \geq c \mid p = \tfrac{1}{4}) \leq 0.01.$$

From all the numbers c, with this property, we choose the smallest, in order to minimize the probability of a Type II error, a false negative. For the above example, we select: $c = 13$.

Radioactive suitcase

As an example, consider determining whether a suitcase contains some radioactive material. Placed under a Geiger counter, it produces 10 counts per minute. The null hypothesis is that no radioactive material is in the suitcase and that all measured counts are due to ambient radioactivity typical of the surrounding air and harmless objects. We can then calculate how likely it is that we would observe 10 counts per minute if the null hypothesis were true. If the null hypothesis predicts (say) on average 9 counts per minute, then according to the Poisson distribution typical for radioactive decay there is about 41% chance of recording 10 or more counts. Thus we can say that the suitcase is compatible with the null hypothesis (this does not guarantee that there is no radioactive material, just that we don't have enough evidence to suggest there is). On the other hand, if the null hypothesis predicts 3 counts per minute (for which the Poisson distribution predicts only 0.1% chance of recording 10 or more counts) then the suitcase is not compatible with the null hypothesis, and there are likely other factors responsible to produce the measurements.

The test does not directly assert the presence of radioactive material. A *successful* test asserts that the claim of no radioactive material present is unlikely given the reading (and therefore ...). The double negative (disproving the null hypothesis) of the method is confusing, but using a counter-example to disprove is standard mathematical practice. The attraction of the method is its practicality. We know (from experience) the expected range of counts with only ambient radioactivity present, so we can say that a measurement is *unusually* large. Statistics just formalizes the intuitive by using numbers instead of adjectives. We probably do not know the characteristics of the radioactive suitcases; We just assume that they produce larger readings.

To slightly formalize intuition: radioactivity is suspected if the Geiger-count with the suitcase is among or exceeds the greatest (5% or 1%) of the Geiger-counts made with ambient radiation alone. This makes no assumptions about the distribution of counts. Many ambient radiation observations are required to obtain good probability estimates for rare events.

The test described here is more fully the null-hypothesis statistical significance test. The null hypothesis represents what we would believe by default, before seeing any evidence. Statistical significance is a possible finding of the test, declared when the observed sample is unlikely to have occurred by chance if the null hypothesis were true. The name of the test describes its formulation and its possible outcome. One characteristic of the test is its crisp decision: to reject or not reject the null hypothesis. A calculated value is compared to a threshold, which is determined from the tolerable risk of error.

Definition of terms

The following definitions are mainly based on the exposition in the book by Lehmann and Romano:[4]

Statistical hypothesis
: A statement about the parameters describing a population (not a sample).

Statistic

A value calculated from a sample without any unknown parameters, often to summarize the sample for comparison purposes.

Simple hypothesis
Any hypothesis which specifies the population distribution completely.

Composite hypothesis
Any hypothesis which does *not* specify the population distribution completely.

Null hypothesis (H_0)
A hypothesis associated with a contradiction to a theory one would like to prove.

Positive data
Data that enable the investigator to reject a null hypothesis.

Alternative hypothesis (H_1)
A hypothesis (often composite) associated with a theory one would like to prove.

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.

Power of a test ($1 - \beta$)
The test's probability of correctly rejecting the null hypothesis when the alternative hypothesis is true. The complement of the false negative rate, β. Power is termed **sensitivity** in biostatistics. ("This is a sensitive test. Because the result is negative, we can confidently say that the patient does not have the condition.") See sensitivity and specificity and Type I and type II errors for exhaustive definitions.

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. ("This is a specific test. Because the result is positive, we can confidently say that the patient has the condition.") See sensitivity and specificity and Type I and type II errors for exhaustive definitions.

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 is the maximum exposure to erroneously rejecting H_0 that they are 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. In case of a composite null hypothesis, the worst case probability.

Statistical significance test
A predecessor to the statistical hypothesis test (see the Origins section). 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. The term is loosely used for the modern version which is now part of statistical hypothesis testing.

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

Variations and sub-classes

Statistical hypothesis testing is a key technique of both frequentist inference and Bayesian inference, although the two types of inference have notable differences. 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. Note that this probability of making an incorrect decision is *not* the probability that the null hypothesis is true, nor whether any specific alternative hypothesis is true. This contrasts with other possible techniques of decision theory in which the null and alternative hypothesis are treated on a more equal basis.

One naïve Bayesian approach to hypothesis testing is to base decisions on the posterior probability,[30][31] 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 via decision theory and optimal decisions, some of which have desirable properties. 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. Such considerations can be used for the purpose of sample size determination prior to the collection of data.

History

Early use

While hypothesis testing was popularized early in the 20th century, early forms were used in the 1700s. The first use is credited to John Arbuthnot (1710),[32] followed by Pierre-Simon Laplace (1770s), in analyzing the human sex ratio at birth; see § Human sex ratio.

Modern origins and early controversy

Modern significance testing is largely the product of Karl Pearson (*p*-value, Pearson's chi-squared test), William Sealy Gosset (Student's t-distribution), and Ronald Fisher ("null hypothesis", analysis of variance, "significance test"), while hypothesis testing was developed by Jerzy Neyman and Egon Pearson (son of Karl). Ronald Fisher began his life in statistics as a Bayesian (Zabell 1992), but Fisher soon grew disenchanted with the subjectivity involved (namely use of the principle of indifference when determining prior probabilities), and sought to provide a more "objective" approach to inductive inference.[33]

Fisher was an agricultural statistician who emphasized rigorous experimental design and methods to extract a result from few samples assuming Gaussian distributions. Neyman (who teamed with the younger Pearson) emphasized mathematical rigor and methods to obtain more results from many samples and a wider range of distributions. Modern hypothesis testing is an inconsistent hybrid of the Fisher vs Neyman/Pearson formulation, methods and terminology developed in the early 20th century.

Fisher popularized the "significance test". He required a null-hypothesis (corresponding to a population frequency distribution) and a sample. His (now familiar) calculations determined whether to reject the null-hypothesis or not. Significance testing did not utilize an alternative hypothesis so there was no concept of a Type II error.

The *p*-value was devised as an informal, but objective, index meant to help a researcher determine (based on other knowledge) whether to modify future experiments or strengthen one's faith in the null hypothesis.[34] Hypothesis testing (and Type I/II errors) was devised by Neyman and Pearson as a more objective alternative to Fisher's *p*-value, also meant to determine researcher behaviour, but without requiring any inductive inference by the researcher.[35][36]

Neyman & Pearson considered a different problem (which they called "hypothesis testing"). They initially considered two simple hypotheses (both with frequency distributions). They calculated two probabilities and typically selected the hypothesis associated with the higher probability (the hypothesis more likely to have generated the sample). Their method always selected a hypothesis. It also allowed the calculation of both types of error probabilities.

Fisher and Neyman/Pearson clashed bitterly. Neyman/Pearson considered their formulation to be an improved generalization of significance testing.(The defining paper[35] was abstract. Mathematicians have generalized and refined the theory for decades.[37]) Fisher thought that it was not applicable to scientific research because often, during the course of the experiment, it is discovered that the initial assumptions about the null hypothesis are questionable due to unexpected sources of error. He believed that the use of rigid reject/accept decisions based on models formulated before data is collected was incompatible with this common scenario faced by scientists and attempts to apply this method to scientific research would lead to mass confusion.[38]

The dispute between Fisher and Neyman–Pearson was waged on philosophical grounds, characterized by a philosopher as a dispute over the proper role of models in statistical inference.[39]

Events intervened: Neyman accepted a position in the western hemisphere, breaking his partnership with Pearson and separating disputants (who had occupied the same building) by much of the planetary diameter. World War II provided an intermission in the debate. The dispute between Fisher and Neyman terminated (unresolved after 27 years) with Fisher's death in 1962. Neyman wrote a well-regarded eulogy.[40] Some of Neyman's later publications reported *p*-values and significance levels.[41]

The modern version of hypothesis testing is a hybrid of the two approaches that resulted from confusion by writers of statistical textbooks (as predicted by Fisher) beginning in the 1940s.[42] (But signal detection, for example, still uses the Neyman/Pearson formulation.) Great conceptual differences and many caveats in addition to those mentioned above were ignored. Neyman and Pearson provided the stronger terminology, the more rigorous mathematics and the more consistent philosophy, but the subject taught today in introductory statistics has more similarities with Fisher's method than theirs.[43] This history explains the inconsistent terminology (example: the null hypothesis is never accepted, but there is a region of acceptance).

Sometime around 1940,[42] in an apparent effort to provide researchers with a "non-controversial"[44] way to have their cake and eat it too, the authors of statistical text books began anonymously combining these two strategies by using the *p*-value in place of the test statistic (or data) to test against the Neyman–Pearson "significance level".[42] Thus, researchers were encouraged to infer the strength of their data against some null hypothesis using *p*-values, while also thinking they are retaining the post-data collection objectivity provided by hypothesis testing. It then became customary for the null hypothesis, which was originally some realistic research hypothesis, to be used almost solely as a strawman "nil" hypothesis (one where a treatment has no effect, regardless of the context).[45]

A comparison between Fisherian, frequentist (Neyman–Pearson)

#	Fisher's null hypothesis testing	Neyman–Pearson decision theory
1	Set up a statistical null hypothesis. The null need not be a nil hypothesis (i.e., zero difference).	Set up two statistical hypotheses, H1 and H2, and decide about α, β, and sample size before the experiment, based on subjective cost-benefit considerations. These define a rejection region for each hypothesis.
2	Report the exact level of significance (e.g. p = 0.051 or p = 0.049). Do not use a conventional 5% level, and do not talk about accepting or rejecting hypotheses. If the result is "not significant", draw no conclusions and make no decisions, but suspend judgement until further data is available.	If the data falls into the rejection region of H1, accept H2; otherwise accept H1. Note that accepting a hypothesis does not mean that you believe in it, but only that you act as if it were true.
3	Use this procedure only if little is known about the problem at hand, and only to draw provisional conclusions in the context of an attempt to understand the experimental situation.	The usefulness of the procedure is limited among others to situations where you have a disjunction of hypotheses (e.g. either μ1 = 8 or μ2 = 10 is true) and where you can make meaningful cost-benefit trade-offs for choosing alpha and beta.

Early choices of null hypothesis

Paul Meehl has argued that the epistemological importance of the choice of null hypothesis has gone largely unacknowledged. When the null hypothesis is predicted by theory, a more precise experiment will be a more severe test of the underlying theory. When the null hypothesis defaults to "no difference" or "no effect", a more precise experiment is a less severe test of the theory that motivated performing the experiment.[46] An examination of the origins of the latter practice may therefore be useful:

1778: Pierre Laplace compares the birthrates of boys and girls in multiple European cities. He states: "it is natural to conclude that these possibilities are very nearly in the same ratio". Thus Laplace's null hypothesis that the birthrates of boys and girls should be equal given "conventional wisdom".[23]

1900: Karl Pearson develops the chi squared test to determine "whether a given form of frequency curve will effectively describe the samples drawn from a given population." Thus the null hypothesis is that a population is described by some distribution predicted by theory. He uses as an example the numbers of five and sixes in the Weldon dice throw data.[47]

1904: Karl Pearson develops the concept of "contingency" in order to determine whether outcomes are independent of a given categorical factor. Here the null hypothesis is by default that two things are unrelated (e.g. scar formation and death rates from smallpox).[48] The null hypothesis in this case is no longer predicted by theory or conventional wisdom, but is instead the principle of indifference that led Fisher and others to dismiss the use of "inverse probabilities".[49]

Null hypothesis statistical significance testing

An example of Neyman–Pearson hypothesis testing can be made by a change to the radioactive suitcase example. If the "suitcase" is actually a shielded container for the transportation of radioactive material, then a test might be used to select among three hypotheses: no radioactive source present, one present, two (all) present. The test could be required for safety, with actions required in each case. The Neyman–Pearson lemma of hypothesis testing says that a good criterion for the selection of hypotheses is the ratio of their probabilities (a likelihood ratio). A simple method of solution is to select the hypothesis with the highest probability for the Geiger counts observed. The typical result matches intuition: few counts imply no source, many counts imply two sources and intermediate counts imply one source. Notice also that usually there are problems for proving a negative. Null hypotheses should be at least falsifiable.

Neyman–Pearson theory can accommodate both prior probabilities and the costs of actions resulting from decisions.[50] The former allows each test to consider the results of earlier tests (unlike Fisher's significance tests). The latter allows the consideration of economic issues (for example) as well as probabilities. A likelihood ratio remains a good criterion for selecting among hypotheses.

The two forms of hypothesis testing are based on different problem formulations. The original test is analogous to a true/false question; the Neyman–Pearson test is more like multiple choice. In the view of Tukey[51] the former produces a conclusion on the basis of only strong evidence while the latter produces a decision on the basis of available evidence. While the two tests seem quite different both mathematically

and philosophically, later developments lead to the opposite claim. Consider many tiny radioactive sources. The hypotheses become 0,1,2,3... grains of radioactive sand. There is little distinction between none or some radiation (Fisher) and 0 grains of radioactive sand versus all of the alternatives (Neyman–Pearson). The major Neyman–Pearson paper of 1933[35] also considered composite hypotheses (ones whose distribution includes an unknown parameter). An example proved the optimality of the (Student's) *t*-test, "there can be no better test for the hypothesis under consideration" (p 321). Neyman–Pearson theory was proving the optimality of Fisherian methods from its inception.

Fisher's significance testing has proven a popular flexible statistical tool in application with little mathematical growth potential. Neyman–Pearson hypothesis testing is claimed as a pillar of mathematical statistics,[52] creating a new paradigm for the field. It also stimulated new applications in statistical process control, detection theory, decision theory and game theory. Both formulations have been successful, but the successes have been of a different character.

The dispute over formulations is unresolved. Science primarily uses Fisher's (slightly modified) formulation as taught in introductory statistics. Statisticians study Neyman–Pearson theory in graduate school. Mathematicians are proud of uniting the formulations. Philosophers consider them separately. Learned opinions deem the formulations variously competitive (Fisher vs Neyman), incompatible[33] or complementary.[37] The dispute has become more complex since Bayesian inference has achieved respectability.

The terminology is inconsistent. Hypothesis testing can mean any mixture of two formulations that both changed with time. Any discussion of significance testing vs hypothesis testing is doubly vulnerable to confusion.

Fisher thought that hypothesis testing was a useful strategy for performing industrial quality control, however, he strongly disagreed that hypothesis testing could be useful for scientists.[34] Hypothesis testing provides a means of finding test statistics used in significance testing.[37] The concept of power is useful in explaining the consequences of adjusting the significance level and is heavily used in sample size determination. The two methods remain philosophically distinct.[39] They usually (but *not always*) produce the same mathematical answer. The preferred answer is context dependent.[37] While the existing merger of Fisher and Neyman–Pearson theories has been heavily criticized, modifying the merger to achieve Bayesian goals has been considered.[53]

Criticism

Criticism of statistical hypothesis testing fills volumes.[54][55][56][57][58][59] Much of the criticism can be summarized by the following issues:

- The interpretation of a *p*-value is dependent upon stopping rule and definition of multiple comparison. The former often changes during the course of a study and the latter is unavoidably ambiguous. (i.e. "p values depend on both the (data) observed and on the other possible (data) that might have been observed but weren't").[60]
- Confusion resulting (in part) from combining the methods of Fisher and Neyman–Pearson which are conceptually distinct.[51]
- Emphasis on statistical significance to the exclusion of estimation and confirmation by repeated experiments.[61]
- Rigidly requiring statistical significance as a criterion for publication, resulting in publication bias.[62] Most of the criticism is indirect. Rather than being wrong, statistical hypothesis testing is misunderstood, overused and misused.
- When used to detect whether a difference exists between groups, a paradox arises. As improvements are made to experimental design (e.g. increased precision of measurement and sample size), the test becomes more lenient. Unless one accepts the absurd assumption that all sources of noise in the data cancel out completely, the chance of finding statistical significance in either direction approaches 100%.[63] However, this absurd assumption that the mean difference between two groups cannot be zero implies that the data cannot be independent and identically distributed (i.i.d.) because the expected difference between any two subgroups of i.i.d. random variates is zero; therefore, the i.i.d. assumption is also absurd.
- Layers of philosophical concerns. The probability of statistical significance is a function of decisions made by experimenters/analysts.[11] If the decisions are based on convention they are termed arbitrary or mindless[44] while those not so based may be termed subjective. To minimize type II errors, large samples are recommended. In psychology practically all null hypotheses are claimed to be false for sufficiently large samples so "...it is usually nonsensical to perform an experiment with the *sole* aim of rejecting the null hypothesis.".[64] "Statistically significant findings are often misleading" in psychology.[65] Statistical significance does not imply practical significance and correlation does not imply causation. Casting doubt on the null hypothesis is thus far from directly supporting the research hypothesis.
- "[I]t does not tell us what we want to know".[66] Lists of dozens of complaints are available.[58][67][68]

Critics and supporters are largely in factual agreement regarding the characteristics of null hypothesis significance testing (NHST): While it can provide critical information, it is *inadequate as the sole tool for statistical analysis*. *Successfully rejecting the null hypothesis may offer no support for the research hypothesis.* The continuing controversy concerns the selection of the best statistical practices for the near-term future given the (often poor) existing practices. Critics would prefer to ban NHST completely, forcing a complete departure from those practices, while supporters suggest a less absolute change.

Controversy over significance testing, and its effects on publication bias in particular, has produced several results. The American Psychological Association has strengthened its statistical reporting requirements after review,[69] medical journal publishers have recognized the obligation to publish some results that are not statistically significant to combat publication bias[70] and a journal (*Journal of Articles in*

Support of the Null Hypothesis) has been created to publish such results exclusively.[71] Textbooks have added some cautions[72] and increased coverage of the tools necessary to estimate the size of the sample required to produce significant results. Major organizations have not abandoned use of significance tests although some have discussed doing so.[69]

Alternatives

A unifying position of critics is that statistics should not lead to an accept-reject conclusion or decision, but to an estimated value with an interval estimate; this data-analysis philosophy is broadly referred to as estimation statistics. Estimation statistics can be accomplished with either frequentist [1] (https://www.ncbi.nlm.nih.gov/pubmed/31217592) or Bayesian methods.[73]

One strong critic of significance testing suggested a list of reporting alternatives:[74] effect sizes for importance, prediction intervals for confidence, replications and extensions for replicability, meta-analyses for generality. None of these suggested alternatives produces a conclusion/decision. Lehmann said that hypothesis testing theory can be presented in terms of conclusions/decisions, probabilities, or confidence intervals. "The distinction between the ... approaches is largely one of reporting and interpretation."[75]

On one "alternative" there is no disagreement: Fisher himself said,[26] "In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result." Cohen, an influential critic of significance testing, concurred,[66] "... don't look for a magic alternative to NHST *[null hypothesis significance testing]* ... It doesn't exist." "... given the problems of statistical induction, we must finally rely, as have the older sciences, on replication." The "alternative" to significance testing is repeated testing. The easiest way to decrease statistical uncertainty is by obtaining more data, whether by increased sample size or by repeated tests. Nickerson claimed to have never seen the publication of a literally replicated experiment in psychology.[67] An indirect approach to replication is meta-analysis.

Bayesian inference is one proposed alternative to significance testing. (Nickerson cited 10 sources suggesting it, including Rozeboom (1960)).[67] For example, Bayesian parameter estimation can provide rich information about the data from which researchers can draw inferences, while using uncertain priors that exert only minimal influence on the results when enough data is available. Psychologist John K. Kruschke has suggested Bayesian estimation as an alternative for the *t*-test.[76] Alternatively two competing models/hypothesis can be compared using Bayes factors.[77] Bayesian methods could be criticized for requiring information that is seldom available in the cases where significance testing is most heavily used. Neither the prior probabilities nor the probability distribution of the test statistic under the alternative hypothesis are often available in the social sciences.[67]

Advocates of a Bayesian approach sometimes claim that the goal of a researcher is most often to objectively assess the probability that a hypothesis is true based on the data they have collected.[78][79] Neither Fisher's significance testing, nor Neyman–Pearson hypothesis testing can provide this information, and do not claim to. The probability a hypothesis is true can only be derived from use of Bayes' Theorem, which was unsatisfactory to both the Fisher and Neyman–Pearson camps due to the explicit use of subjectivity in the form of the prior probability.[35][80] Fisher's strategy is to sidestep this with the *p*-value (an objective *index* based on the data alone) followed by *inductive inference*, while Neyman–Pearson devised their approach of *inductive behaviour*.

Philosophy

Hypothesis testing and philosophy intersect. Inferential statistics, which includes hypothesis testing, is applied probability. Both probability and its application are intertwined with philosophy. Philosopher David Hume wrote, "All knowledge degenerates into probability." Competing practical definitions of probability reflect philosophical differences. The most common application of hypothesis testing is in the scientific interpretation of experimental data, which is naturally studied by the philosophy of science.

Fisher and Neyman opposed the subjectivity of probability. Their views contributed to the objective definitions. The core of their historical disagreement was philosophical.

Many of the philosophical criticisms of hypothesis testing are discussed by statisticians in other contexts, particularly correlation does not imply causation and the design of experiments. Hypothesis testing is of continuing interest to philosophers.[39][81]

Education

Statistics is increasingly being taught in schools with hypothesis testing being one of the elements taught.[82][83] Many conclusions reported in the popular press (political opinion polls to medical studies) are based on statistics. Some writers have stated that statistical analysis of this kind allows for thinking clearly about problems involving mass data, as well as the effective reporting of trends and inferences from said data, but caution that writers for a broad public should have a solid understanding of the field in order to use the terms and concepts correctly.[84][85][84][85] An introductory college statistics class places much emphasis on hypothesis testing – perhaps half of the course. Such fields as literature and divinity now include findings based on statistical analysis (see the Bible Analyzer). An introductory statistics class teaches hypothesis testing as a cookbook process. Hypothesis testing is also taught at the postgraduate level. Statisticians learn how to create good statistical test procedures (like *z*, Student's *t*, *F* and chi-squared). Statistical hypothesis testing is considered a mature area within statistics,[75] but a limited amount of development continues.

An academic study states that the cookbook method of teaching introductory statistics leaves no time for history, philosophy or controversy. Hypothesis testing has been taught as received unified method. Surveys showed that graduates of the class were filled with philosophical misconceptions (on all aspects of statistical inference) that persisted among instructors.[86] While the problem was addressed more than a decade ago,[87] and calls for educational reform continue,[88] students still graduate from statistics classes holding fundamental misconceptions about hypothesis testing.[89] Ideas for improving the teaching of hypothesis testing include encouraging students to search for statistical errors in published papers, teaching the history of statistics and emphasizing the controversy in a generally dry subject.[90]

See also

- Statistics
- Behrens–Fisher problem
- Bootstrapping (statistics)
- Checking if a coin is fair
- Comparing means test decision tree
- Complete spatial randomness
- Counternull
- Falsifiability
- Fisher's method for combining independent tests of significance
- Granger causality
- Look-elsewhere effect
- Modifiable areal unit problem
- Multivariate hypothesis testing
- Omnibus test
- Dichotomous thinking
- Almost sure hypothesis testing

References

1. Stuart A., Ord K., Arnold S. (1999), *Kendall's Advanced Theory of Statistics: Volume 2A—Classical Inference & the Linear Model* (Arnold) §20.2.
2. Rice, John A. (2007). *Mathematical Statistics and Data Analysis* (3rd ed.). Thomson Brooks/Cole. §9.3.
3. Burnham, K. P.; Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A practical information-theoretic approach* (https://archive.org/details/modelselectionmu0000burn) (2nd ed.). Springer-Verlag. ISBN 978-0-387-95364-9.
4. Lehmann, E. L.; Romano, Joseph P. (2005). *Testing Statistical Hypotheses* (3E ed.). New York: Springer. ISBN 978-0-387-98864-1.
5. Triola, Mario (2001). *Elementary statistics* (https://archive.org/details/elementarystatis00trio/page/388) (8 ed.). Boston: Addison-Wesley. p. 388 (https://archive.org/details/elementarystatis00trio/page/388). ISBN 978-0-201-61477-0.
6. Hinkelmann, Klaus and Kempthorne, Oscar (2008). *Design and Analysis of Experiments*. I and II (Second ed.). Wiley. ISBN 978-0-470-38551-7.
7. Montgomery, Douglas (2009). *Design and analysis of experiments*. Hoboken, N.J.: Wiley. ISBN 978-0-470-12866-4.
8. R. A. Fisher (1925).*Statistical Methods for Research Workers*, Edinburgh: Oliver and Boyd, 1925, p.43.
9. Nuzzo, Regina (2014). "Scientific method: Statistical errors" (https://doi.org/10.1038%2F506150a). *Nature*. **506** (7487): 150–152. Bibcode:2014Natur.506..150N (https://ui.adsabs.harvard.edu/abs/2014Natur.506..150N). doi:10.1038/506150a (https://doi.org/10.1038%2F506150a). PMID 24522584 (https://pubmed.ncbi.nlm.nih.gov/24522584).
10. Siegrist, Kyle. "Hypothesis Testing - Introduction" (http://www.randomservices.org/random/hypothesis/Introduction.html). *www.randomservices.org*. Retrieved March 8, 2018.
11. Bakan, David (1966). "The test of significance in psychological research". *Psychological Bulletin*. **66** (6): 423–437. doi:10.1037/h0020412 (https://doi.org/10.1037%2Fh0020412). PMID 5974619 (https://pubmed.ncbi.nlm.nih.gov/5974619).
12. Richard J. Larsen; Donna Fox Stroup (1976). *Statistics in the Real World: a book of examples*. Macmillan. ISBN 978-0023677205.
13. Hubbard, R.; Parsa, A. R.; Luthy, M. R. (1997). "The Spread of Statistical Significance Testing in Psychology: The Case of the Journal of Applied Psychology". *Theory and Psychology*. **7** (4): 545–554. doi:10.1177/0959354397074006 (https://doi.org/10.1177%2F0959354397074006). S2CID 145576828 (https://api.semanticscholar.org/CorpusID:145576828).
14. Moore, David (2003). *Introduction to the Practice of Statistics*. New York: W.H. Freeman and Co. p. 426. ISBN 9780716796572.
15. Huff, Darrell (1993). *How to lie with statistics* (https://archive.org/details/howtoliewithstat00huff). New York: Norton. ISBN 978-0-393-31072-6.
16. Huff, Darrell (1991). *How to Lie with Statistics*. London: Penguin Books. ISBN 978-0-14-013629-6.
17. "Over the last fifty years, How to Lie with Statistics has sold more copies than any other statistical text." J. M. Steele. ""Darrell Huff and Fifty Years of *How to Lie with Statistics*" (http://www-stat.wharton.upenn.edu/~steele/Publications/PDF/TN148.pdf). *Statistical Science*, 20 (3), 2005, 205–209.
18. John Arbuthnot (1710). "An argument for Divine Providence, taken from the constant regularity observed in the births of both sexes" (http://www.york.ac.uk/depts/maths/histstat/arbuthnot.pdf) (PDF). *Philosophical Transactions of the Royal Society of London*. **27** (325–336): 186–190. doi:10.1098/rstl.1710.0011 (https://doi.org/10.1098%2Frstl.1710.0011). S2CID 186209819 (https://api.semanticscholar.org/CorpusID:186209819).
19. Brian, Éric; Jaisson, Marie (2007). "Physico-Theology and Mathematics (1710–1794)". *The Descent of Human Sex Ratio at Birth* (https://archive.org/details/descenthumansexr00bria). Springer Science & Business Media. pp. 1 (https://archive.org/details/descenthumansexr00bria/page/n17)–25. ISBN 978-1-4020-6036-6.
20. Conover, W.J. (1999), "Chapter 3.4: The Sign Test", *Practical Nonparametric Statistics* (Third ed.), Wiley, pp. 157–176, ISBN 978-0-471-16068-7

21. Sprent, P. (1989), *Applied Nonparametric Statistical Methods* (Second ed.), Chapman & Hall, ISBN 978-0-412-44980-2
22. Stigler, Stephen M. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900*. Harvard University Press. pp. 225–226 (https://archive.org/details/historyofstatist00stig/page/225). ISBN 978-0-67440341-3.
23. Laplace, P. (1778). "Mémoire sur les probabilités" (http://cerebro.xu.edu/math/Sources/Laplace/memoir_probabilities.pdf) (PDF). *Mémoires de l'Académie Royale des Sciences de Paris*. **9**: 227–332.
24. Laplace, P. (1778). "Mémoire sur les probabilités (XIX, XX)" (http://gallica.bnf.fr/ark:/12148/bpt6k77597p/f386). *Oeuvres complètes de Laplace*. *Mémoires de l'Académie Royale des Sciences de Paris*. **9**. pp. 429–438.
25. Stigler, Stephen M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900* (https://archive.org/details/historyofstatist00stig/page/134). Cambridge, Mass: Belknap Press of Harvard University Press. p. 134 (https://archive.org/details/historyofstatist00stig/page/134). ISBN 978-0-674-40340-6.
26. Fisher, Sir Ronald A. (1956) [1935]. "Mathematics of a Lady Tasting Tea" (https://books.google.com/books?id=oKZwtLQTmNAC&q=%22mathematics+of+a+lady+tasting+tea%22&pg=PA1512). In James Roy Newman (ed.). *The World of Mathematics, volume 3* [*Design of Experiments*]. Courier Dover Publications. ISBN 978-0-486-41151-4. Originally from Fisher's book *Design of Experiments*.
27. Box, Joan Fisher (1978). *R.A. Fisher, The Life of a Scientist*. New York: Wiley. p. 134. ISBN 978-0-471-09300-8.
28. C. S. Peirce (August 1878). "Illustrations of the Logic of Science VI: Deduction, Induction, and Hypothesis" (https://en.wikisource.org/w/index.php?oldid=3592335). *Popular Science Monthly*. **13**. Retrieved March 30, 2012.
29. Jaynes, E. T. (2007). *Probability theory : the logic of science* (5. print. ed.). Cambridge [u.a.]: Cambridge Univ. Press. ISBN 978-0-521-59271-0.
30. Schervish, M (1996) *Theory of Statistics*, p. 218. Springer ISBN 0-387-94546-6
31. Kaye, David H.; Freedman, David A. (2011). "Reference Guide on Statistics" (http://www.nap.edu/openbook.php?record_id=13163&page=211). *Reference Manual on Scientific Evidence* (3rd ed.). Eagan, MN Washington, D.C: West National Academies Press. p. 259. ISBN 978-0-309-21421-6.
32. Bellhouse, P. (2001), "John Arbuthnot", *in Statisticians of the Centuries by C.C. Heyde and E. Seneta*, Springer, pp. 39–42, ISBN 978-0-387-95329-8
33. Raymond Hubbard, M. J. Bayarri, *P Values are not Error Probabilities (http://ftp.isds.duke.edu/WorkingPapers/03-26.pdf) Archived (https://web.archive.org/web/20130904000350/http://ftp.isds.duke.edu/WorkingPapers/03-26.pdf) September 4, 2013, at the Wayback Machine*. A working paper that explains the difference between Fisher's evidential *p*-value and the Neyman–Pearson Type I error rate α.
34. Fisher, R (1955). "Statistical Methods and Scientific Induction" (http://www.phil.vt.edu/dmayo/PhilStatistics/Triad/Fisher%201955.pdf) (PDF). *Journal of the Royal Statistical Society, Series B*. **17** (1): 69–78.
35. Neyman, J; Pearson, E. S. (January 1, 1933). "On the Problem of the most Efficient Tests of Statistical Hypotheses" (https://doi.org/10.1098%2Frsta.1933.0009). *Philosophical Transactions of the Royal Society A*. **231** (694–706): 289–337. Bibcode:1933RSPTA.231..289N (https://ui.adsabs.harvard.edu/abs/1933RSPTA.231..289N). doi:10.1098/rsta.1933.0009 (https://doi.org/10.1098%2Frsta.1933.0009).
36. Goodman, S N (June 15, 1999). "Toward evidence-based medical statistics. 1: The P Value Fallacy". *Ann Intern Med*. **130** (12): 995–1004. doi:10.7326/0003-4819-130-12-199906150-00008 (https://doi.org/10.7326%2F0003-4819-130-12-199906150-00008). PMID 10383371 (https://pubmed.ncbi.nlm.nih.gov/10383371). S2CID 7534212 (https://api.semanticscholar.org/CorpusID:7534212).
37. Lehmann, E. L. (December 1993). "The Fisher, Neyman–Pearson Theories of Testing Hypotheses: One Theory or Two?". *Journal of the American Statistical Association*. **88** (424): 1242–1249. doi:10.1080/01621459.1993.10476404 (https://doi.org/10.1080%2F01621459.1993.10476404).
38. Fisher, R N (1958). "The Nature of Probability" (http://www.york.ac.uk/depts/maths/histstat/fisher272.pdf) (PDF). *Centennial Review*. **2**: 261–274."We are quite in danger of sending highly trained and highly intelligent young men out into the world with tables of erroneous numbers under their arms, and with a dense fog in the place where their brains ought to be. In this century, of course, they will be working on guided missiles and advising the medical profession on the control of disease, and there is no limit to the extent to which they could impede every sort of national effort."
39. Lenhard, Johannes (2006). "Models and Statistical Inference: The Controversy between Fisher and Neyman–Pearson". *Br. J. Philos. Sci*. **57**: 69–91. doi:10.1093/bjps/axi152 (https://doi.org/10.1093%2Fbjps%2Faxi152). S2CID 14136146 (https://api.semanticscholar.org/CorpusID:14136146).
40. Neyman, Jerzy (1967). "RA Fisher (1890—1962): An Appreciation". *Science*. **156** (3781): 1456–1460. Bibcode:1967Sci...156.1456N (https://ui.adsabs.harvard.edu/abs/1967Sci...156.1456N). doi:10.1126/science.156.3781.1456 (https://doi.org/10.1126%2Fscience.156.3781.1456). PMID 17741062 (https://pubmed.ncbi.nlm.nih.gov/17741062). S2CID 44708120 (https://api.semanticscholar.org/CorpusID:44708120).
41. Losavich, J. L.; Neyman, J.; Scott, E. L.; Wells, M. A. (1971). "Hypothetical explanations of the negative apparent effects of cloud seeding in the Whitetop Experiment" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC389491). *Proceedings of the National Academy of Sciences of the United States of America*. **68** (11): 2643–2646. Bibcode:1971PNAS...68.2643L (https://ui.adsabs.harvard.edu/abs/1971PNAS...68.2643L). doi:10.1073/pnas.68.11.2643 (https://doi.org/10.1073%2Fpnas.68.11.2643). PMC 389491 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC389491). PMID 16591951 (https://pubmed.ncbi.nlm.nih.gov/16591951).
42. Halpin, P F; Stam, HJ (Winter 2006). "Inductive Inference or Inductive Behavior: Fisher and Neyman: Pearson Approaches to Statistical Testing in Psychological Research (1940–1960)". *The American Journal of Psychology*. **119** (4): 625–653. doi:10.2307/20445367 (https://doi.org/10.2307%2F20445367). JSTOR 20445367 (https://www.jstor.org/stable/20445367). PMID 17286092 (https://pubmed.ncbi.nlm.nih.gov/17286092).

43. Gigerenzer, Gerd; Zeno Swijtink; Theodore Porter; Lorraine Daston; John Beatty; Lorenz Kruger (1989). "Part 3: The Inference Experts". *The Empire of Chance: How Probability Changed Science and Everyday Life*. Cambridge University Press. pp. 70–122. ISBN 978-0-521-39838-1.
44. Gigerenzer, G (November 2004). "Mindless statistics". *The Journal of Socio-Economics*. **33** (5): 587–606. doi:10.1016/j.socec.2004.09.033 (https://doi.org/10.1016%2Fj.socec.2004.09.033).
45. Loftus, G R (1991). "On the Tyranny of Hypothesis Testing in the Social Sciences" (https://www.ics.uci.edu/~sternh/courses/210/loftus91_tyranny.pdf) (PDF). *Contemporary Psychology*. **36** (2): 102–105. doi:10.1037/029395 (https://doi.org/10.1037%2F029395).
46. Meehl, P (1990). "Appraising and Amending Theories: The Strategy of Lakatosian Defense and Two Principles That Warrant It" (http://rhowell.ba.ttu.edu/meehl1.pdf) (PDF). *Psychological Inquiry*. **1** (2): 108–141. doi:10.1207/s15327965pli0102_1 (https://doi.org/10.1207%2Fs15327965pli0102_1).
47. Pearson, K (1900). "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling" (http://www.economics.soton.ac.uk/staff/aldrich/1900.pdf) (PDF). *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*. **5** (50): 157–175. doi:10.1080/14786440009463897 (https://doi.org/10.1080%2F14786440009463897).
48. Pearson, K (1904). "On the Theory of Contingency and Its Relation to Association and Normal Correlation" (https://archive.org/details/cu31924003064833). *Drapers' Company Research Memoirs Biometric Series*. **1**: 1–35.
49. Zabell, S (1989). "R. A. Fisher on the History of Inverse Probability" (https://doi.org/10.1214%2Fss%2F1177012488). *Statistical Science*. **4** (3): 247–256. doi:10.1214/ss/1177012488 (https://doi.org/10.1214%2Fss%2F1177012488). JSTOR 2245634 (https://www.jstor.org/stable/2245634).
50. Ash, Robert (1970). *Basic probability theory*. New York: Wiley. ISBN 978-0471034506.Section 8.2
51. Tukey, John W. (1960). "Conclusions vs decisions". *Technometrics*. **26** (4): 423–433. doi:10.1080/00401706.1960.10489909 (https://doi.org/10.1080%2F00401706.1960.10489909). "Until we go through the accounts of testing hypotheses, separating [Neyman–Pearson] decision elements from [Fisher] conclusion elements, the intimate mixture of disparate elements will be a continual source of confusion." ... "There is a place for both "doing one's best" and "saying only what is certain," but it is important to know, in each instance, both which one is being done, and which one ought to be done."
52. Stigler, Stephen M. (August 1996). "The History of Statistics in 1933" (https://doi.org/10.1214%2Fss%2F1032280216). *Statistical Science*. **11** (3): 244–252. doi:10.1214/ss/1032280216 (https://doi.org/10.1214%2Fss%2F1032280216). JSTOR 2246117 (https://www.jstor.org/stable/2246117).
53. Berger, James O. (2003). "Could Fisher, Jeffreys and Neyman Have Agreed on Testing?" (https://doi.org/10.1214%2Fss%2F1056397485). *Statistical Science*. **18** (1): 1–32. doi:10.1214/ss/1056397485 (https://doi.org/10.1214%2Fss%2F1056397485).
54. Morrison, Denton; Henkel, Ramon, eds. (2006) [1970]. *The Significance Test Controversy*. AldineTransaction. ISBN 978-0-202-30879-1.
55. Oakes, Michael (1986). *Statistical Inference: A Commentary for the Social and Behavioural Sciences*. Chichester New York: Wiley. ISBN 978-0471104438.
56. Chow, Siu L. (1997). *Statistical Significance: Rationale, Validity and Utility*. ISBN 978-0-7619-5205-3.
57. Harlow, Lisa Lavoie; Stanley A. Mulaik; James H. Steiger, eds. (1997). *What If There Were No Significance Tests?*. Lawrence Erlbaum Associates. ISBN 978-0-8058-2634-0.
58. Kline, Rex (2004). *Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research*. Washington, D.C.: American Psychological Association. ISBN 9781591471189.
59. McCloskey, Deirdre N.; Stephen T. Ziliak (2008). *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. University of Michigan Press. ISBN 978-0-472-05007-9.
60. Cornfield, Jerome (1976). "Recent Methodological Contributions to Clinical Trials" (http://www.epidemiology.ch/history/PDF%20bg/Cornfield%20J%201976%20recent%20methodological%20contributions.pdf) (PDF). *American Journal of Epidemiology*. **104** (4): 408–421. doi:10.1093/oxfordjournals.aje.a112313 (https://doi.org/10.1093%2Foxfordjournals.aje.a112313). PMID 788503 (https://pubmed.ncbi.nlm.nih.gov/788503).
61. Yates, Frank (1951). "The Influence of Statistical Methods for Research Workers on the Development of the Science of Statistics". *Journal of the American Statistical Association*. **46** (253): 19–34. doi:10.1080/01621459.1951.10500764 (https://doi.org/10.1080%2F01621459.1951.10500764). "The emphasis given to formal tests of significance throughout [R.A. Fisher's] Statistical Methods ... has caused scientific research workers to pay undue attention to the results of the tests of significance they perform on their data, particularly data derived from experiments, and too little to the estimates of the magnitude of the effects they are investigating." ... "The emphasis on tests of significance and the consideration of the results of each experiment in isolation, have had the unfortunate consequence that scientific workers have often regarded the execution of a test of significance on an experiment as the ultimate objective."
62. Begg, Colin B.; Berlin, Jesse A. (1988). "Publication bias: a problem in interpreting medical data". *Journal of the Royal Statistical Society, Series A*. **151** (3): 419–463. doi:10.2307/2982993 (https://doi.org/10.2307%2F2982993). JSTOR 2982993 (https://www.jstor.org/stable/2982993).

63. Meehl, Paul E. (1967). "Theory-Testing in Psychology and Physics: A Methodological Paradox" (https://web.archive.org/web/20131203010657/http://mres.gmu.edu/pmwiki/uploads/Main/Meehl1967.pdf) (PDF). *Philosophy of Science*. **34** (2): 103–115. doi:10.1086/288135 (https://doi.org/10.1086%2F288135). S2CID 96422880 (https://api.semanticscholar.org/CorpusID:96422880). Archived from the original (http://mres.gmu.edu/pmwiki/uploads/Main/Meehl1967.pdf) (PDF) on December 3, 2013. Thirty years later, Meehl acknowledged statistical significance theory to be mathematically sound while continuing to question the default choice of null hypothesis, blaming instead the "social scientists' poor understanding of the logical relation between theory and fact" in "The Problem Is Epistemology, Not Statistics: Replace Significance Tests by Confidence Intervals and Quantify Accuracy of Risky Numerical Predictions" (Chapter 14 in Harlow (1997)).
64. Nunnally, Jum (1960). "The place of statistics in psychology". *Educational and Psychological Measurement*. **20** (4): 641–650. doi:10.1177/001316446002000401 (https://doi.org/10.1177%2F001316446002000401). S2CID 144813784 (https://api.semanticscholar.org/CorpusID:144813784).
65. Lykken, David T. (1991). "What's wrong with psychology, anyway?". *Thinking Clearly About Psychology*. **1**: 3–39.
66. Jacob Cohen (December 1994). "The Earth Is Round (p < .05)" (https://semanticscholar.org/paper/2cc7be3d5161e865807e13de7975c9d77fbd2815). *American Psychologist*. **49** (12): 997–1003. doi:10.1037/0003-066X.49.12.997 (https://doi.org/10.1037%2F0003-066X.49.12.997). S2CID 380942 (https://api.semanticscholar.org/CorpusID:380942). This paper lead to the review of statistical practices by the APA. Cohen was a member of the Task Force that did the review.
67. Nickerson, Raymond S. (2000). "Null Hypothesis Significance Tests: A Review of an Old and Continuing Controversy" (https://semanticscholar.org/paper/8c5e0e6f85b9dc15ecf23d43a49404925c4c41bf). *Psychological Methods*. **5** (2): 241–301. doi:10.1037/1082-989X.5.2.241 (https://doi.org/10.1037%2F1082-989X.5.2.241). PMID 10937333 (https://pubmed.ncbi.nlm.nih.gov/10937333). S2CID 28340967 (https://api.semanticscholar.org/CorpusID:28340967).
68. Branch, Mark (2014). "Malignant side effects of null hypothesis significance testing" (https://semanticscholar.org/paper/48f8711f3ca3535192ce695fa987847725374b0e). *Theory & Psychology*. **24** (2): 256–277. doi:10.1177/0959354314525282 (https://doi.org/10.1177%2F0959354314525282). S2CID 40712136 (https://api.semanticscholar.org/CorpusID:40712136).
69. Wilkinson, Leland (1999). "Statistical Methods in Psychology Journals; Guidelines and Explanations". *American Psychologist*. **54** (8): 594–604. doi:10.1037/0003-066X.54.8.594 (https://doi.org/10.1037%2F0003-066X.54.8.594). "Hypothesis tests. It is hard to imagine a situation in which a dichotomous accept-reject decision is better than reporting an actual p value or, better still, a confidence interval." (p 599). The committee used the cautionary term "forbearance" in describing its decision against a ban of hypothesis testing in psychology reporting. (p 603)
70. "ICMJE: Obligation to Publish Negative Studies" (https://web.archive.org/web/20120716211637/http://www.icmje.org/publishing_1negative.html). Archived from the original (http://www.icmje.org/publishing_1negative.html) on July 16, 2012. Retrieved September 3, 2012. "Editors should seriously consider for publication any carefully done study of an important question, relevant to their readers, whether the results for the primary or any additional outcome are statistically significant. Failure to submit or publish findings because of lack of statistical significance is an important cause of publication bias."
71. *Journal of Articles in Support of the Null Hypothesis* website: JASNH homepage (http://www.jasnh.com/). Volume 1 number 1 was published in 2002, and all articles are on psychology-related subjects.
72. Howell, David (2002). *Statistical Methods for Psychology* (https://archive.org/details/statisticalmetho0000howe/page/94) (5 ed.). Duxbury. p. 94 (https://archive.org/details/statisticalmetho0000howe/page/94). ISBN 978-0-534-37770-0.
73. Kruschke, J K (July 9, 2012). "Bayesian Estimation Supersedes the T Test" (http://www.indiana.edu/~kruschke/articles/Kruschke2012JEPG.pdf) (PDF). *Journal of Experimental Psychology: General*. **142** (2): 573–603. doi:10.1037/a0029146 (https://doi.org/10.1037%2Fa0029146). PMID 22774788 (https://pubmed.ncbi.nlm.nih.gov/22774788).
74. Armstrong, J. Scott (2007). "Significance tests harm progress in forecasting" (http://repository.upenn.edu/cgi/viewcontent.cgi?article=1104&context=marketing_papers). *International Journal of Forecasting*. **23** (2): 321–327. CiteSeerX 10.1.1.343.9516 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.343.9516). doi:10.1016/j.ijforecast.2007.03.004 (https://doi.org/10.1016%2Fj.ijforecast.2007.03.004).
75. E. L. Lehmann (1997). "Testing Statistical Hypotheses: The Story of a Book" (https://doi.org/10.1214%2Fss%2F1029963261). *Statistical Science*. **12** (1): 48–52. doi:10.1214/ss/1029963261 (https://doi.org/10.1214%2Fss%2F1029963261).
76. Kruschke, J K (July 9, 2012). "Bayesian Estimation Supersedes the T Test" (http://www.indiana.edu/~kruschke/articles/Kruschke2012JEPG.pdf) (PDF). *Journal of Experimental Psychology: General*. **142** (2): 573–603. doi:10.1037/a0029146 (https://doi.org/10.1037%2Fa0029146). PMID 22774788 (https://pubmed.ncbi.nlm.nih.gov/22774788).
77. Kass, R. E. (1993). "Bayes factors and model uncertainty" (http://www.stat.washington.edu/research/reports/1993/tr254.pdf) (PDF). Department of Statistics, University of Washington.
78. Rozeboom, William W (1960). "The fallacy of the null-hypothesis significance test" (http://stats.org.uk/statistical-inference/Rozeboom1960.pdf) (PDF). *Psychological Bulletin*. **57** (5): 416–428. CiteSeerX 10.1.1.398.9002 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.398.9002). doi:10.1037/h0042040 (https://doi.org/10.1037%2Fh0042040). PMID 13744252 (https://pubmed.ncbi.nlm.nih.gov/13744252). "...the proper application of statistics to scientific inference is irrevocably committed to extensive consideration of inverse [AKA Bayesian] probabilities..." It was acknowledged, with regret, that a priori probability distributions were available "only as a subjective feel, differing from one person to the next" "in the more immediate future, at least".
79. Berger, James (2006). "The Case for Objective Bayesian Analysis" (https://doi.org/10.1214%2F06-ba115). *Bayesian Analysis*. **1** (3): 385–402. doi:10.1214/06-ba115 (https://doi.org/10.1214%2F06-ba115). In listing the competing definitions of "objective" Bayesian analysis, "A major goal of statistics (indeed science) is to find a completely coherent objective Bayesian methodology for learning from data." The author expressed the view that this goal "is not attainable".

80. Aldrich, J (2008). "R. A. Fisher on Bayes and Bayes' theorem" (https://web.archive.org/web/20140906190025/http://ba.stat.cmu.edu/journal/2008/vol03/issue01/aldrich.pdf) (PDF). *Bayesian Analysis*. **3** (1): 161–170. doi:10.1214/08-BA306 (https://doi.org/10.1214%2F08-BA306). Archived from the original (http://ba.stat.cmu.edu/journal/2008/vol03/issue01/aldrich.pdf) (PDF) on September 6, 2014.
81. Mayo, D. G.; Spanos, A. (2006). "Severe Testing as a Basic Concept in a Neyman–Pearson Philosophy of Induction". *The British Journal for the Philosophy of Science*. **57** (2): 323–357. CiteSeerX 10.1.1.130.8131 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.130.8131). doi:10.1093/bjps/axl003 (https://doi.org/10.1093%2Fbjps%2Faxl003). S2CID 7176653 (https://api.semanticscholar.org/CorpusID:7176653).
82. Mathematics > High School: Statistics & Probability > Introduction (http://www.corestandards.org/the-standards/mathematics/hs-statistics-and-probability/introduction/) Archived (https://archive.is/20120728122912/http://www.corestandards.org/the-standards/mathematics/hs-statistics-and-probability/introduction/) July 28, 2012, at Archive.today Common Core State Standards Initiative (relates to USA students)
83. College Board Tests > AP: Subjects > Statistics (http://www.collegeboard.com/student/testing/ap/sub_stats.html) The College Board (relates to USA students)
84. Huff, Darrell (1993). *How to lie with statistics* (https://archive.org/details/howtoliewithstat00huff/page/8). New York: Norton. p. 8 (https://archive.org/details/howtoliewithstat00huff/page/8). ISBN 978-0-393-31072-6.'Statistical methods and statistical terms are necessary in reporting the mass data of social and economic trends, business conditions, "opinion" polls, the census. But without writers who use the words with honesty and readers who know what they mean, the result can only be semantic nonsense.'
85. Snedecor, George W.; Cochran, William G. (1967). *Statistical Methods* (6 ed.). Ames, Iowa: Iowa State University Press. p. 3. "...the basic ideas in statistics assist us in thinking clearly about the problem, provide some guidance about the conditions that must be satisfied if sound inferences are to be made, and enable us to detect many inferences that have no good logical foundation."
86. Sotos, Ana Elisa Castro; Vanhoof, Stijn; Noortgate, Wim Van den; Onghena, Patrick (2007). "Students' Misconceptions of Statistical Inference: A Review of the Empirical Evidence from Research on Statistics Education" (https://lirias.kuleuven.be/bitstream/123456789/136347/1/CastroSotos.pdf) (PDF). *Educational Research Review*. **2** (2): 98–113. doi:10.1016/j.edurev.2007.04.001 (https://doi.org/10.1016%2Fj.edurev.2007.04.001).
87. Moore, David S. (1997). "New Pedagogy and New Content: The Case of Statistics" (http://www.stat.auckland.ac.nz/~iase/publications/isr/97.Moore.pdf) (PDF). *International Statistical Review*. **65** (2): 123–165. doi:10.2307/1403333 (https://doi.org/10.2307%2F1403333). JSTOR 1403333 (https://www.jstor.org/stable/1403333).
88. Hubbard, Raymond; Armstrong, J. Scott (2006). "Why We Don't Really Know What Statistical Significance Means: Implications for Educators" (https://web.archive.org/web/20060518054857/http://hops.wharton.upenn.edu/ideas/pdf/Armstrong/StatisticalSignificance.pdf) (PDF). *Journal of Marketing Education*. **28** (2): 114–120. doi:10.1177/0273475306288399 (https://doi.org/10.1177%2F0273475306288399). hdl:2092/413 (https://hdl.handle.net/2092%2F413). S2CID 34729227 (https://api.semanticscholar.org/CorpusID:34729227). Archived from the original on May 18, 2006. Preprint (http://escholarshare.drake.edu/bitstream/handle/2092/413/WhyWeDon't.pdf)
89. Sotos, Ana Elisa Castro; Vanhoof, Stijn; Noortgate, Wim Van den; Onghena, Patrick (2009). "How Confident Are Students in Their Misconceptions about Hypothesis Tests?" (https://doi.org/10.1080%2F10691898.2009.11889514). *Journal of Statistics Education*. **17** (2). doi:10.1080/10691898.2009.11889514 (https://doi.org/10.1080%2F10691898.2009.11889514).
90. Gigerenzer, G. (2004). "The Null Ritual What You Always Wanted to Know About Significant Testing but Were Afraid to Ask" (http://library.mpib-berlin.mpg.de/ft/gg/GG_Null_2004.pdf) (PDF). *The SAGE Handbook of Quantitative Methodology for the Social Sciences*. pp. 391–408. doi:10.4135/9781412986311 (https://doi.org/10.4135%2F9781412986311). ISBN 9780761923596.

Further reading

- Lehmann E.L. (1992) "Introduction to Neyman and Pearson (1933) On the Problem of the Most Efficient Tests of Statistical Hypotheses". In: *Breakthroughs in Statistics, Volume 1*, (Eds Kotz, S., Johnson, N.L.), Springer-Verlag. ISBN 0-387-94037-5 (followed by reprinting of the paper)
- Neyman, J.; Pearson, E.S. (1933). "On the Problem of the Most Efficient Tests of Statistical Hypotheses" (https://doi.org/10.1098%2Frsta.1933.0009). *Philosophical Transactions of the Royal Society A*. **231** (694–706): 289–337. Bibcode:1933RSPTA.231..289N (https://ui.adsabs.harvard.edu/abs/1933RSPTA.231..289N). doi:10.1098/rsta.1933.0009 (https://doi.org/10.1098%2Frsta.1933.0009).

External links

- "Statistical hypotheses, verification of" (https://www.encyclopediaofmath.org/index.php?title=Statistical_hypotheses,_verification_of), *Encyclopedia of Mathematics*, EMS Press, 2001 [1994]
- Wilson González, Georgina; Kay Sankaran (September 10, 1997). "Hypothesis Testing" (http://www.webapps.cee.vt.edu/ewr/environmental/teach/smprimer/hypotest/ht.html). *Environmental Sampling & Monitoring Primer*. Virginia Tech.
- Bayesian critique of classical hypothesis testing (http://www.cs.ucsd.edu/users/goguen/courses/275f00/stat.html)
- Critique of classical hypothesis testing highlighting long-standing qualms of statisticians (https://web.archive.org/web/20051124221846/http://www.npwrc.usgs.gov/resource/methods/statsig/stathyp.htm)
- Dallal GE (2007) The Little Handbook of Statistical Practice (http://www.tufts.edu/~gdallal/LHSP.HTM) (A good tutorial)

- References for arguments for and against hypothesis testing (http://core.ecu.edu/psyc/wuenschk/StatHelp/NHST-SHIT.htm)
- Statistical Tests Overview: (https://web.archive.org/web/20091029162244/http://www.wiwi.uni-muenster.de/ioeb/en/organisation/pfaff/stat_overview_table.html) How to choose the correct statistical test
- [2] (https://arxiv.org/abs/1401.2851) Statistical Analysis based Hypothesis Testing Method in Biological Knowledge Discovery; Md. Naseef-Ur-Rahman Chowdhury, Suvankar Paul, Kazi Zakia Sultana

Online calculators

- MBAStats confidence interval and hypothesis test calculators (http://www.mbastats.net)
- Some p-value and hypothesis test calculators (http://www.schramm.cc/link/Statistics-calculator.php).
