Topic #7: P-value

In statistical hypothesis testing, the p-value is the probability of obtaining a result at least as extreme as that obtained, assuming the truth of the null hypothesis that the finding was the result of chance alone. The fact that p-values are based on this assumption is crucial to their correct interpretation.

More technically, the p-value of an observed value tobserved of some random variable T used as a test statistic is the probability that, given that the null hypothesis is true, T will assume a value as or more unfavorable to the null hypothesis as the observed value tobserved. "More unfavorable to the null hypothesis" can in some cases mean greater than, in some cases less than, and in some cases further away from a specified center.

Interpretation

Generally, one rejects the null hypothesis if the p-value is smaller than or equal to the significance level, often represented by the Greek letter ? (alpha). If the level is 0.05, then the results are only 5% likely to be as extraordinary as just seen, given that the null hypothesis is true.

In the above example, the calculated p-value exceeds 0.05, and thus the null hypothesis - that the observed result of 14 heads out of 20 flips can be ascribed to chance alone - is not rejected. Such a finding is often stated as being "not statistically significant at the 5 % level".

However, had a single extra head been obtained, the resulting p-value would be 0.02. This time the null hypothesis - that the observed result of 15 heads out of 20 flips can be ascribed to chance alone - is rejected. Such a finding would be described as being "statistically significant at the 5 % level".

There is often an alternative hypothesis, but the construction of the test does not allow for 'supporting' a specific alternative.

Critics of p-values point out that the criterion used to decide "statistical significance" is based on the somewhat arbitrary choice of level (often set at 0.05). A proposed replacement for the p-value is p-rep, which is the probability that an effect can be replicated.

Frequent misunderstandings

There are several common misunderstandings about p-values.

1. The p-value is not the probability that the null hypothesis is true, (claimed to justify the "rule" of considering as significant p-values closer to 0 (zero)). In fact, frequentist statistics does not, and cannot, attach probabilities to hypotheses. Comparison of Bayesian and classical approaches shows that a p-value can be very close to zero while the posterior probability of the null is very close to unity. This is the Jeffreys-Lindley paradox.
2. The p-value is not the probability that a finding is "merely a fluke" (again, justifying the "rule" of considering small p-values as "significant"). As the calculation of a p-value is based on the assumption that a finding is the product of chance alone, it patently cannot simultaneously be use to gauge the probability of that assumption being true.
3. The p-value is not the probability of falsely rejecting the null hypothesis. This error is a version of the so-called prosecutor's fallacy.
4. The p-value is not the probability that a replicating experiment would not yield the same conclusion.
5. 1 - (p-value) is not the probability of the alternative hypothesis being true (see (1)).
6. The significance level of the test is not determined by the p-value.The significance level of a test is decided upon before any data are collected, and does not depend on the p-value or any other statistic calculated after the test has been performed.