

What is a P-value?

I have found that many students are unsure about the interpretation of P-values and other concepts related to tests of significance. These ideas are used repeatedly in various applications so it is important that they be understood. I will explain the concepts in general terms first, then their application in the problem of assessing normality.

We wish to test a null hypothesis against an alternative hypothesis using a dataset. The two hypotheses specify two statistical models for the process that produced the data. The alternative hypothesis is what we expect to be true if the null hypothesis is false. We cannot prove that the alternative hypothesis is true but we may be able to demonstrate that the alternative is much more plausible than the null hypothesis given the data. This demonstration is usually expressed in terms of a probability (a P-value) quantifying the strength of the evidence against the null hypothesis in favor of the alternative.

We ask whether the data appear to be consistent with the null hypothesis or whether it is unlikely that we would obtain data of this kind if the null hypothesis were true, assuming that at least one of the two hypotheses is true. We address this question by calculating the value of a test statistic, i.e., a particular real-valued function of the data. To decide whether the value of the test statistic is consistent with the null hypothesis, we need to know what sampling variability to expect in our test statistic if the null hypothesis is true. In other words, we need to know the null distribution, the distribution of the test statistic when the null hypothesis is true. In many applications, the test statistic is defined so that its null distribution is a “named” distribution for which tables are widely accessible; e.g., the standard normal distribution, the Binomial distribution with $n = 100$ and $p = 1/2$, the t distribution with 4 degrees of freedom, the chi-square distribution with 23 degrees of freedom, the F distribution with 2 and 20 degrees of freedom.

Now, given the value of the test statistic (a number), and the null distribution of the test statistic (a theoretical distribution usually represented by a probability density), we want to see whether the test statistic is in the middle of the distribution (consistent with the null hypothesis) or out in a tail of the distribution (making the alternative hypothesis seem more plausible). Sometimes we will want to consider the right-hand tail, sometimes the left-hand tail, and sometimes both tails, depending on how the test statistic and alternative hypothesis are defined. Suppose that large positive values of the test statistic seem more plausible under the alternative hypothesis than under the null hypothesis. Then we want a measure of how far out our test statistic is in the right-hand tail of the null distribution. The P-value provides a measure of this distance. The P-value (in this situation) is the probability to the right of our test statistic calculated using the null distribution. The further out the test statistic is in the tail, the smaller the P-value, and the stronger the evidence against the null hypothesis in favor of the alternative.

The P-value can be interpreted in terms of a hypothetical repetition of the study. Suppose the null hypothesis is true and a new dataset is obtained independently of the first dataset but using the same sampling procedure. If the new dataset is used to calculate a new value of the test statistic (same formula but new data), what is the probability that the new value will be further out in the tail (assuming a one-tailed test) than the original value? This probability is the P-value.

The P-value is often incorrectly interpreted as the probability that the null hypothesis is true. Try not to make this mistake. In a frequentist interpretation of probability, there is nothing random about whether the hypothesis is true, the randomness is in the process generating the data. One can interpret “the probability that the null hypothesis is true” using subjective probability, a measure of one’s belief that the null hypothesis is true. One can

then calculate this subjective probability by specifying a prior probability (subjective belief before looking at the data) that the null hypothesis is true, and then use the data and the model to update one's subjective probability. This is called the Bayesian approach because Bayes' Theorem is used to update subjective probabilities to reflect new information.

When reporting a P-value to persons unfamiliar with statistics, it is often necessary to use descriptive language to indicate the strength of the evidence. I tend to use the following sort of language. Obviously the cut-offs are somewhat arbitrary and another person might use different language.

$P > 0.10$	No evidence against the null hypothesis. The data appear to be consistent with the null hypothesis.
$0.05 < P < 0.10$	Weak evidence against the null hypothesis in favor of the alternative
$0.01 < P < 0.05$	Moderate evidence against the null hypothesis in favor of the alternative.
$0.001 < P < 0.01$	Strong evidence against the null hypothesis in favor of the alternative.
$P < 0.001$	Very strong evidence against the null hypothesis in favor of the alternative.

In using this kind of language, one should keep in mind the difference between statistical significance and practical significance. In a large study one may obtain a small P-value even though the magnitude of the effect being tested is too small to be of importance (see the discussion of power below). It is a good idea to support a P-value with a confidence interval for the parameter being tested.

A P-value can also be reported more formally in terms of a fixed level α test. Here α is a number selected independently of the data, usually 0.05 or 0.01, more rarely 0.10. We reject the null hypothesis at level α if the P-value is smaller than α , otherwise we fail to reject the null hypothesis at level α . I am not fond of this kind of language because it suggests a more definite, clear-cut answer than is often available. There is essentially no difference between a P-value of 0.051 and 0.049. In some situations it may be necessary to proceed with some course of action based on our belief in whether the null or alternative hypothesis is true. More often, it seems better to report the P-value as a measure of evidence.

A fixed level α test can be calculated without first calculating a P-value. This is done by comparing the test statistic with a critical value of the null distribution corresponding to the level α . This is usually the easiest approach when doing hand calculations and using statistical tables, which provide percentiles for a relatively small set of probabilities. Most statistical software produces P-values which can be compared directly with α . There is no need to repeat the calculation by hand.

Fixed level α tests are needed for discussing the power of a test, a useful concept when planning a study. Suppose we are comparing a new medical treatment with a standard treatment, the control. The null hypothesis is that of no treatment effect (no difference between treatment and control). The alternative hypothesis is that the treatment effect (mean difference of treatment minus control using some outcome variable) is positive. We want to have good chance of reporting a small P-value assuming the alternative hypothesis

is true and the magnitude of the effect is large enough to be of practical importance. The power of a level α test is defined to be the probability that the null hypothesis will be rejected at level α (i.e., the P-value will be less than α) assuming the alternative hypothesis is true. The power generally depends on the variability of the data (lower variance, higher power), the sample size (higher n , higher power), and the magnitude of the effect (larger effect, higher power).

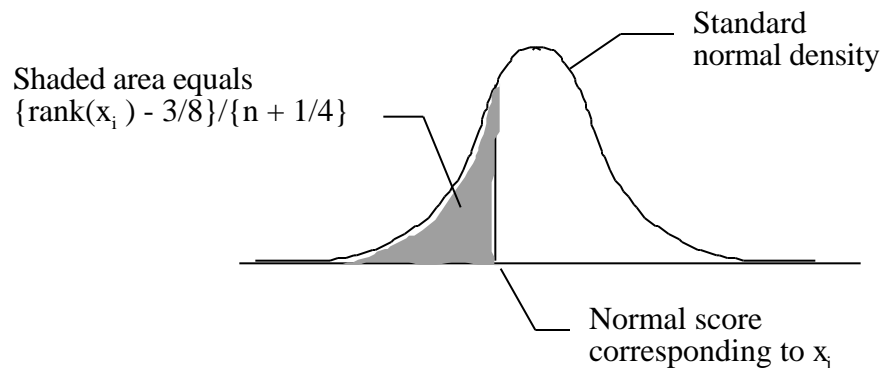
Assessing normality using the Ryan-Joiner test.

Null hypothesis: the data $\{x_1, \dots, x_n\}$ are a random sample of size n from a normal distribution.

Alternative hypothesis: the data are a random sample from some other distribution.

Test statistic: $r =$ the correlation between the data and the normal scores.

The normal scores are defined by the following graph.



Rationale: If the data are a sample from a normal distribution then the normal probability plot (plot of normal scores against the data) will be close to a straight line, and the correlation r will be close to 1. If the data are sampled from a non-normal distribution then the plot may show a marked deviation from a straight line, resulting in a smaller correlation r . Smaller values of r are therefore regarded as stronger evidence against the null hypothesis.

Null distribution of r : I do not know whether this distribution has a name. We might call it the Ryan-Joiner distribution, corresponding to the name of the test. The density will be skewed to the left, with most of the probability close to 1, as in the picture below.

P-value: The probability to the left of the observed correlation r calculated using the null distribution; i.e., the area under the density to the left of r . You do not need to know how to calculate this. Minitab does the calculation for you.

Interpretation: If you want to use simple descriptive language, you can use the table above. The strength of evidence is described directly in terms of the P-value.

