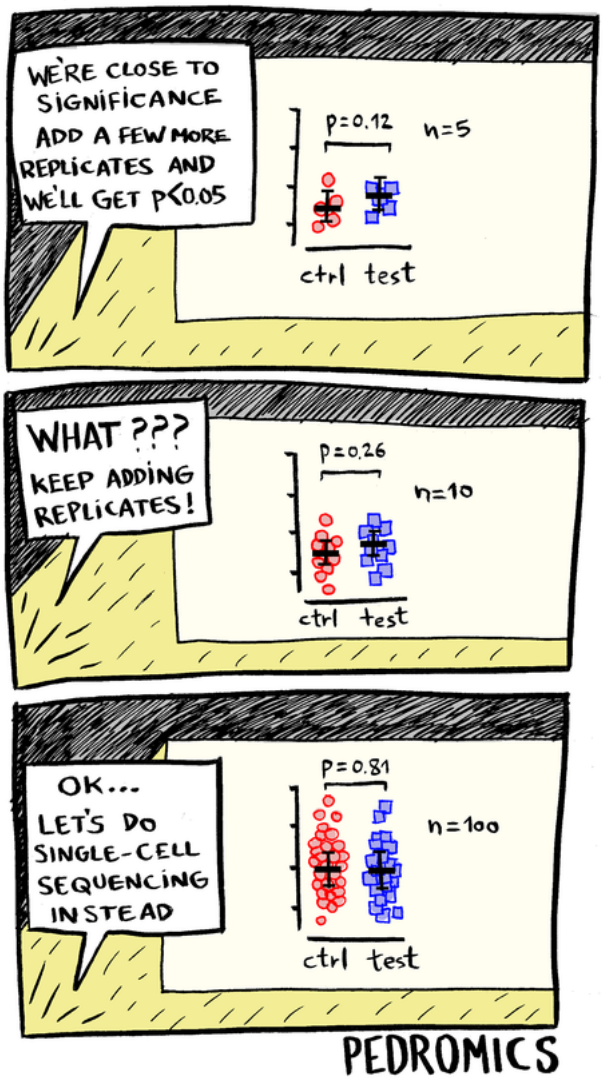
Hypothesis Test 02

EN5423 | Spring 2024

w05_hypothesis_02.pdf
(Week 5)

Contents



1 Tests for Normality

• When we are planning a study that requires a parametric test, which assumes your data are normally distributed, you need to first check if your data actually fits this normal distribution assumption.

• If it does not, we might need to use a non-parametric test, like a permutation test, instead. This is crucial because the normality test helps you decide which statistical test is appropriate for your data.

• The null hypothesis (H_0) for normality tests is that your data are normally distributed. If you reject H_0, it means your data probably does not follow a normal distribution. ***However, not being able to reject H_0 does not necessarily confirm that your data is normally distributed***, especially if you are working with a ***small sample size***. It simply suggests that there is not enough evidence to say it is not normal based on the data you have.

• For normality testing, two common tests are introduced: **the Probability Plot Correlation Coefficient (PPCC) test** and **the Shapiro-Wilk test**. Both tests involve comparing your data to a theoretical normal distribution using a probability plot, which graphs the quantiles of your data against the quantiles of a normal distribution.

• If your data lines up well with a straight line in this plot, it indicates your data may be normal. The PPCC and Shapiro-Wilk tests both yield statistics that are compared to the value 1 (samples from a normal distribution will have a correlation coefficient and R^2 very close to 1); the closer to 1, the more likely your data is normally distributed. Deviations from 1 suggest non-normality.

• To perform a test of H_0 the data are normally distributed) versus H_A (the data are not normally distributed), the statistics are analyzed to see if they are significantly less than 1.

• For example, to illustrate these tests, consider data on water yield from wells, comparing wells in unfractured versus fractured rock (**Table 1** and **Figure 1**). The probability plot for wells in unfractured rock shows a correlation coefficient (r*) of 0.805, indicating a weaker alignment with normality, while wells in fractured rock show r* = 0.943, suggesting closer adherence to a normal distribution.

• For example, with a significance level of 5% ($\alpha = 0.05$), a p-value less than 0.05 would lead you to reject H_0, indicating non-normal distribution.

• The most common test for normality is the Shapiro-Wilk test, as its power to detect non-normality is as good or better than other tests (Thode, 2002). A table of quantiles for this test statistic is available for n <50 (Conover, 1999). Shapiro and Francia (1972) modified the Shapiro-Wilk test for all sample sizes and statistical software usually performs this form of the test, including the , `scipy.stats.shapiro(data` in Python.

• Tests for normality not related to probability plots include the Kolmogorov and chi-square tests, described in more detail by Thode (2002). Both tests have lower power than the Shapiro-Wilk test to detect non-normality when data are continuous (Thode, 2002).

• In Python, you would use use libraries such as SciPy or StatsModels. For the Shapiro-Wilk test, `scipy.stats.shapiro(data)` gives you a W statistic and a p-value, helping you assess normality. Python does not have a direct PPCC test function in standard libraries, but you can perform similar assessments of normality through plotting and statistical analysis available in these libraries.

• In summary, assessing normality is a fundamental step in choosing the right statistical tests for your data. Through tests like the PPCC and Shapiro-Wilk, and tools available in Python, you can make informed decisions on how to proceed with your analysis.

Table 1. Unit well yields (y_i) from Virginia, in gallons per minute per foot (Wright, 1985). [-, no data]

Wells in unfractured rock (y_i)	Wells in fractured rock (y_i)
0.001	0.020
0.003	0.031
0.007	0.086
0.020	0.13
0.030	0.16
0.040	0.16
0.041	0.18
0.077	0.30
0.10	0.40
0.454	0.44
0.49	0.51
1.02	0.72
-	0.95

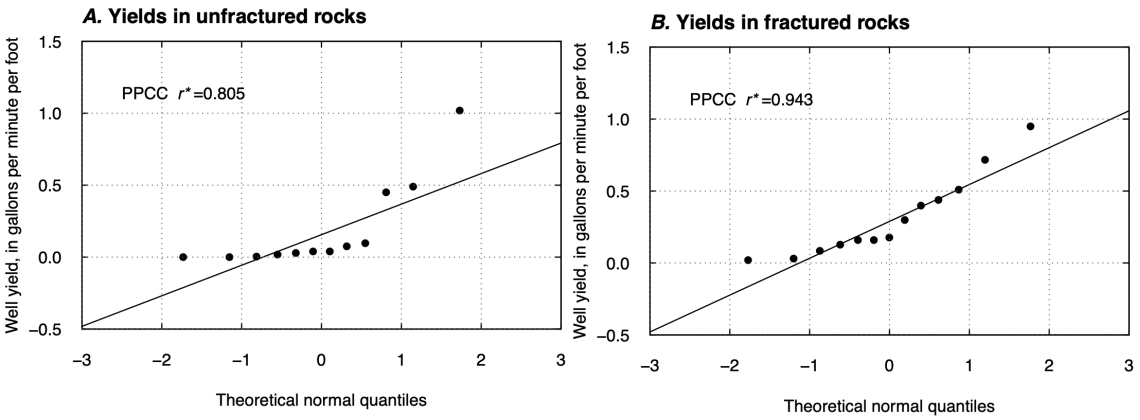


Figure 1. Probability plots for yields of wells in (A) unfractured and (B) fractured rock, with probability plot correlation coefficient (PPCC) correlation coefficient (r). Data from Wright (1985).

2 Other Hypothesis Tests

Many other hypothesis tests exist for questions beyond central tendency and distribution; table of previous lecture chapter lists some other tests and the chapters in which they are introduced. Additional tests include tests for proportions, tests of independence, tests related to variance or

spread, and tests related to skew and kurtosis, among many others. There is an extensive discussion of tests for equality of variances in later lectures. You may consult specialized texts such as Sheskin (2011) for details about many more hypothesis tests.

3 Considerations and Criticisms About Hypothesis Tests

Though widely used, hypothesis tests and *p*-values are subject to misinterpretation and much criticism. Over the decades many have tried to discourage the use of *p*-values in published research, yet they remain widely reported because they are simply an expression of strength of evidence shown by the data. In some instances, if authors do not report a *p*-value with their results, reviewers will ask for a *p*-value as a measure of the statistical significance of the results.

3.1 Considerations and Criticisms About Hypothesis Tests

• The *p*-values obtained from hypothesis testing serve as indicators of statistical significance, not as measures of relevance in environemntal research. It is crucial for researchers to evaluate both the statistical significance and the practical implications of their findings in environmental research.

• McCuen (2016) highlights the challenge of determining scientific significance due to the lack of widely accepted criteria, which often leads to an overreliance on statistical decision-making.

• However, it is essential to make an effort to evaluate the practical importance of findings. For instance, a study might reveal a statistically significant variation in calcium levels in stream water between two locations, but this difference may not necessarily impact human or aquatic life.

• Wasserstein and Lazar (2016) have pointed out the frequent misinterpretations of the *p*-value, such as viewing it as ***the probability that the null hypothesis is correct***, as the chance of a particular event occurring, or as the odds of a specific outcome happening.

• In reality, the *p*-value represents ***the likelihood of observing a test statistic that is as extreme as, or more extreme than, the actual observation, assuming the null hypothesis (H_0) is true***. Essentially, it measures the risk of mistakenly identifying a pattern in the data when none exists (a **type I** error). They explain that a smaller *p*-value suggests a higher degree of statistical inconsistency with the null hypothesis, given that the assumptions for calculating the *p*-value are valid. Yet, they also caution that there is not a universally accepted threshold that definitively indicates statistical inconsistency.

• When presenting the outcomes of hypothesis tests, it is important to include detailed information ***such as 1) the sample size, because the sensitivity of the results to sample size might necessitate a specialized test or an adjustment to the existing test***. Authors should clearly state ***2) the null and alternative hypotheses to clarify what the test aims to examine***; **3) specify the significance level (e.g., $\alpha = 0.05$) used for the test; and report the exact *p*-value**, ***not merely stating that $p < 0.05$***. This approach allows readers to make informed evaluations of the findings' significance.

3.2 Criticisms

• *p*-values have been subjected to severe criticism over many years (Rozeboom, 1960; Nuzzo, 2014), yet they continue to be a staple in statistical analysis. Critics often point out several limitations:

1. *p*-values cannot quantify the size or significance of an effect.
2. *p*-values do not offer a range of possible values like confidence or prediction intervals do.
3. There is a tendency to overlook publishing results that do not reject the null hypothesis. Such results are often wrongly perceived as lacking scientific value, even though they can provide meaningful insights into the studied phenomena.
4. The practice of '*p*-hacking', where researchers engage in multiple hypothesis testing, outlier removal, or additional data collection to secure significant results, is problematic.

• It is important to recognize that *p*-values were never intended to measure effect size. For this purpose, different methods exist, such as the Hodges-Lehmann estimator, to assess the size of an effect. These estimators and *p*-values should be viewed as complementary, each with their own limitations. Merely estimating an average effect without considering its statistical significance can be misleading, as it might only reflect random variation.

• The criticism regarding the dismissal of non-statistically significant findings as scientifically irrelevant is well-founded. In reality, identifying no change can be enlightening, particularly in environmental studies, where it might indicate that certain interventions did not produce the expected outcomes. The scientific community is increasingly recognizing the value of publishing such findings to enrich future research.

• While the manipulation of data or statistical analyses to achieve significant results, known as '*p*-hacking', is a genuine issue, eliminating *p*-values is not the answer. It is critical to remember that with a significance level (α) of 0.05, one might falsely reject the null hypothesis purely by chance in 1 out of every 20 tests.

• Misrepresentation of data, whether intentional or due to lack of knowledge, is a broader problem that transcends the misuse of *p*-values. We do not endorse practices like removing outliers just to achieve certain results, and neither does the field of statistics. The solution lies in improving statistical education among environmental scientists and other researchers.

• Moreover, the relationship between hypothesis tests and confidence intervals is often misunderstood. Some advocate for abandoning hypothesis tests in favor of confidence intervals, not realizing that these concepts are interconnected. A *p*-value essentially reflects the confidence level for the broadest confidence interval that indicates a non-zero effect size, making the debate somewhat paradoxical.

3.3 Discussion

Responding to widespread critiques and concerns, the American Statistical Association (ASA) made an unprecedented move by articulating its stance on statistical practices, specifically addressing the use of p-values (Wasserstein and Lazar, 2016). Rather than advocating for their abandonment, Wasserstein and Lazar sought to clarify the consensus on p-values' application and interpretation. They emphasized the following principles:

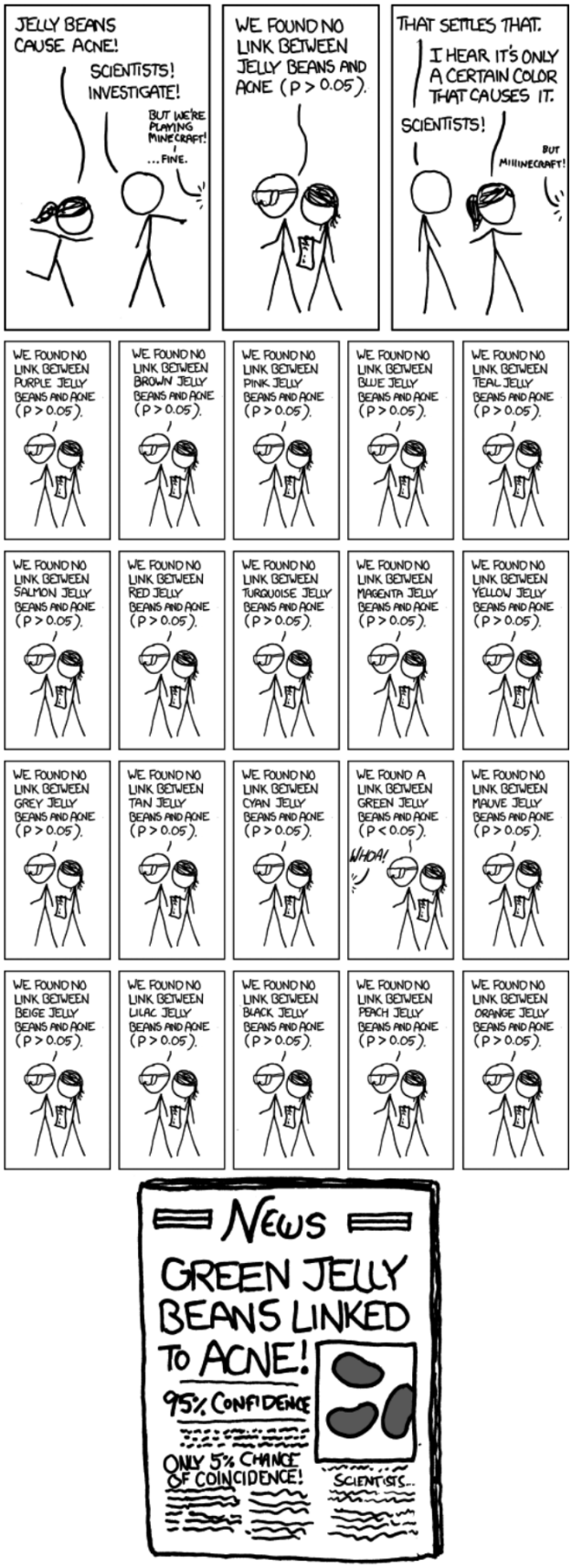


Figure 2. Cartoon showing what can easily happen when running multiple hypothesis tests, *p*-hacking or *p*-fishing. Figure from xkcd.com (Munroe, 2016), used under creative commons attribution-noncommercial license.

1. ***p*-values signal the compatibility of data with a given statistical model.**
2. ***p*-values do not quantify the likelihood of the hypothesis being correct or the data resulting purely from chance.**
3. **Decisions in science, business, or policy should not hinge solely on whether a *p*-value crosses a predetermined threshold.**
4. **Comprehensive inference demands thorough reporting and openness.**
5. **A *p*-value, or its statistical significance, does not gauge the magnitude or relevance of an effect.**
6. **On its own, a *p*-value offers a limited view of the evidence concerning a model or hypothesis.**

• Echoing these principles, we have explored how *p*-values function within environmental research. Statistical tests often guide decisions on necessary actions.

• For instance, examining the impact of a certain chemical on fish reproduction might yield a *p*-value of 0.20 in a one-sided test. Although traditionally, an alpha (α) of 0.05 is used, a *p*-value of 0.20 still suggests a notable adverse effect of the chemical, albeit with some uncertainty. Here, the precautionary principle could justify measures to mitigate the chemical's presence, given its probable negative impact.

• Moreover, *p*-values can obscure insights when multiple hypothesis tests across various locations yield non-significant results individually, yet collectively suggest a consistent trend.

• Omitting the magnitude, direction, or *p*-value due to non-significance deprives readers of valuable insights. Presenting complete test outcomes, even when not statistically significant, enriches the analysis, as demonstrated by Hirsch and Ryberg (2012). Although formal tests exist for multi-site analyses, sharing comprehensive results, regardless of significance, is beneficial, keeping in mind the potential for spatial correlation among data points.

• The ASA's commentary extends beyond statistical methods to underscore that scientific credibility, including reproducibility, is multifaceted. Graphical representations can enhance the interpretation of statistical tests and highlight hydrological findings' significance, illustrating that the judicious use of *p*-values, alongside other analytical tools, enriches scientific inquiry.