P-value or Confidence Intervals What should I do? (an elemental approach)

Macarena Hernández Jiménez
Master in Pharmacological Research
Communication and Data Analysis
November 2023

INDEX

1. INTRODUCTION

Good statistical practice is an essential component of good scientific practice, including standards of Good Laboratory Practices (GLP) and Good Clinical Practices (GCP). Such practice emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean. The development of any study should be carefully designed and reported, in order to guarantee the quality of methods and processes of statistical collection, production and dissemination of results, as well as the coordination, cooperation, and statistical innovation. With this approach, the dissemination of reliable results should be ensured through all the scientific community.

In this context, the design and reporting of results obtained in preclinical and clinical studies is something crucial. However, the methodology of stablished statistical parameters, such us the p-value, is under discussion nowadays.

In the 1920s, Ronald Fisher introduced the p-value as an informal way to judge whether evidence was significant in the old-fashioned sense: worthy of a second look. At that moment, he did not mean it to be a definitive test but to run an experiment, and then see if the results were consistent with what random chance might produce. Researchers would first set up a "null hypothesis" that they wanted to disprove, such as there being no correlation or no difference between two groups. Next, they would play the devil's advocate and, assuming that this null hypothesis was in fact true, calculate the chances of getting results at least as extreme as what was actually observed. This probability was the p-value.

With this approach, Fisher intended it to be just one part of a fluid, non-numerical process that blended data and background knowledge to lead to scientific conclusions. This new approach turned soon in a way to make evidence-based decision-making as rigorous and objective as possible. As a result, it is an abundance of confusion about what the p-value means, and many authors (i.e. J. Neyman or E. Pearson) introduced an alternative framework for data analysis that included statistical power, false positives, false negatives and many other concepts, letting out the p-value.

2. WHAT IS P- VALUE?

A p-value, or probability value, is a statistical measure that helps researchers to assess the evidence against a null hypothesis. The null hypothesis typically represents a default or no-effect assumption, and the p-value indicates the probability of obtaining the observed data (or more extreme) if the null hypothesis was true.

Therefore, facing a real-life problem, in order to calculate p-values, we should **standardize** our value so that the mean value of distribution always equals zero (Figure 1). The benefit of standardization is that statisticians already generate a table that includes the area under each standardized value, so there is no need to calculate the area case by case, but standardize the data using the **z-score** to transform the data.

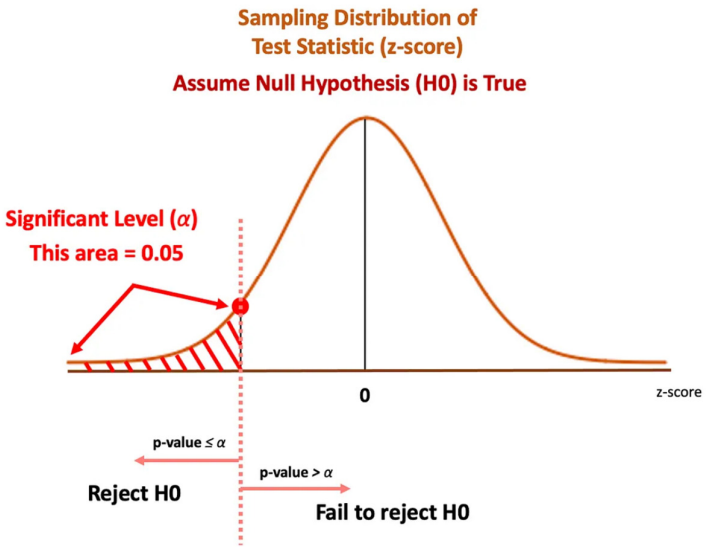

Figure 1. Representative figure of statistical parameters related to p-value. Taken from Chiang 2021.

It is therefore needed to know some important concepts:

Null Hypothesis (H0): the null hypothesis is a statement of no effect, no difference, or no association. It serves as the default assumption that researchers seek to test against. For instance, in a drug trial, the null hypothesis might state that the new drug has no effect, and any observed differences are due to chance. Importantly, H0 can only be rejected but never accepted. Therefore, the results obtained in a specific study can support the H0 rejection and H1 acceptance, but never the H0 acceptance.

Alternative Hypothesis (H1 or Ha): the alternative hypothesis represents what researchers hope to demonstrate. It suggests that there is a significant effect, difference, or association in the data. Using the drug trial example, the alternative hypothesis might state that the new drug has a significant effect compared to a placebo.

P-value Calculation: the p-value is a probability that quantifies the likelihood of observing the data (or more extreme) under the assumption that the null hypothesis is true. A low p-value (typically below a predetermined threshold, often 0.05) suggests that the observed results are unlikely to occur by random chance alone.

Interpretation of p-value: if the p-value is less than the chosen significance level (commonly 0.05), researchers may reject the null hypothesis in favor of the alternative hypothesis. This decision is based on the idea that the observed results are sufficiently unlikely to have occurred by random chance alone.

Significance Level (Alpha): The significance level, often denoted as alpha (α), is the predetermined threshold used to decide whether to reject the H0. Common choices include 0.05, 0.01, or 0.10. A p-value less than or equal to the significance level leads to the rejection of the H0.

P-values are widely used in hypothesis testing across various scientific disciplines, but it's important to interpret them with caution and consider them alongside other statistical measures and the broader context of the study. There has been ongoing discussion about the limitations of p-values, leading researchers to explore alternative statistical approaches and reporting practices.

2.1 PROBLEMS WITH p-VALUE

As mentioned previously in this document, the p-value is a commonly used statistical measure that indicates the evidence against a null hypothesis. However, it's important to note that a significant p-value does not definite prove that the H1 is true, or that the observed effect is practically significant. It only suggests that the observed data are inconsistent with the H0. Conversely, a non-significant p-value does not prove that the H0 is true; it may be due to factors like a small sample size or insufficient statistical power.

Therefore, there are several limitations associated with p-values that researchers and statisticians have recognized so far. Some of these issues include:

1. **Misinterpretation of Results:**

    - **False Positives**: a significant p-value does not prove that the H0 is false or that the effect is practically significant. It only suggests that the observed results are unlikely to have occurred by random chance.
    - **False Negatives**: a non-significant p-value does not prove the H0 is true; it may be due to a lack of power or other factors.

2. **Dependence on Sample Size:** larger sample sizes can lead to smaller p-values even when the effect size is not practically significant. A small p-value with a large sample size might indicate statistical significance but may not be practically relevant. Therefore, researchers should consider not only statistical significance but also practical or clinical significance when interpreting results. A statistically significant result may not necessarily be meaningful in real-world terms.

3. **Multiple Comparisons Problem:** conducting multiple statistical tests increases the likelihood of obtaining at least one significant result by chance (Type I error). Adjustments, such as Bonferroni correction, are often needed to account for this issue.

4. **Publication Bias:** studies with significant results are more likely to be published than those with non-significant results. This can lead to an overestimation of the true effect size in the literature.

5. **Interpretation Issues:** the p-value provides no information about the size or importance of an observed effect (effect size). A small p-value may be associated with a large effect size, but it could also result from a large sample size with a small effect.

6. **Assumption Dependence:** p-values are based on assumptions about the distribution of the data and the statistical model. Violations of these assumptions can lead to inaccurate results.

7. **Cutoff Rigidity:** the common significance level of 0.05 is somewhat arbitrary. Choosing a different cutoff can lead to different conclusions. There is a growing call for using other statistical measures (e.g., confidence intervals) alongside p-values for a more comprehensive assessment.

8. **Lack of Reproducibility:** the use of p-values as a binary decision maker (significant or not) can contribute to challenges in the reproducibility of scientific findings. Replicating studies solely based on p-values may not guarantee similar results.

9. **Large variability**: the p-value is itself a random variable. That means that, in a random sample, we can have experiments with the same underlying truth, but some of them are not significant and others have several orders of magnitude of significance.

P-values have always had critics. In their almost nine decades of existence, they have been likened to mosquitoes (annoying and impossible to swat away), the emperor's new clothes (fraught with obvious problems that everyone ignores) and the tool of a "sterile intellectual rake" who ravishes science but leaves it with no progeny.

3. WHAT ARE CONFIDENCE INTERVALS?

Confidence intervals (CIs) are a statistical concept used to quantify the uncertainty or precision associated with a sample estimate of a population parameter. Instead of providing a single point estimate, a confidence interval provides a range of values within which the true population parameter is likely to fall, along with a level of confidence.

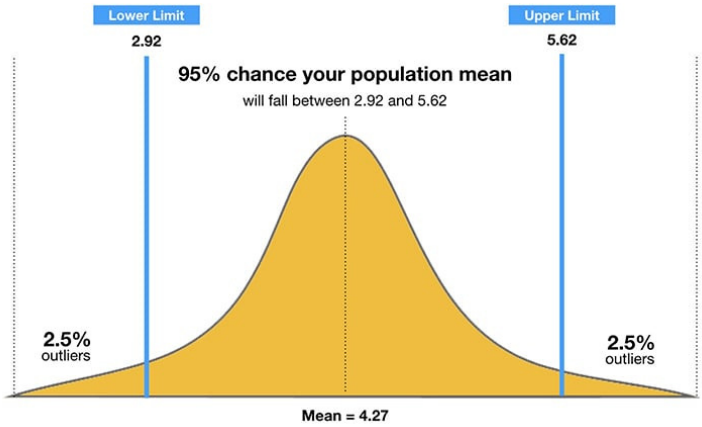

Figure 2. Representative figure of 95% CIs. Taken from McLeod 2023.

Some important concepts related to confidence intervals:

Point Estimate: is a single value that serves as the best guess or approximation of the population parameter based on the sample data. For example, the sample mean, or sample proportion can be used as point estimates for the population mean or population proportion.

Interval Estimate (Confidence Interval): is a range of values derived from the sample data, within which the true population parameter is estimated to lie with a certain level of confidence. This level of confidence is often expressed as a percentage, commonly 95% or 90%.

Construction of Confidence Intervals: the construction of a CI involves using the sample data and statistical methods to determine the lower and upper bounds of the interval. The formula for constructing a confidence interval depends on the distribution of the data and the parameter being estimated.

Interpretation: a 95% confidence interval, for example, implies that if we were to take many random samples from the same population and calculate a confidence interval for each sample, approximately 95% of those intervals would contain the true population parameter.

Width of Confidence Intervals: this parameter is influenced by several factors, including the variability of the data and the chosen level of confidence. A wider interval indicates greater uncertainty, while a narrower interval suggests more precision.

For a better understanding let´s see the following example: suppose a researcher wants to estimate the average height of a certain population. After collecting a sample of data, they calculate a 95% confidence interval for the population mean height. The result might be something like "the true average height of the population is estimated to be between X and Y with 95% confidence."

In summary, confidence intervals provide a more stable and informative way to express the precision of a sample estimate by giving a range of plausible values for the population parameter. In fact, CIs are widely used in various fields, including statistics, epidemiology, and social sciences, to convey the uncertainty associated with parameter estimates.

3.1 PROBLEMS WITH CONFIDENCE INTERVALS

While confidence intervals are a valuable statistical tool for conveying the uncertainty associated with sample estimates, there are some considerations and potential issues to be aware of:

1. **Interpretation Complexity:** interpreting confidence intervals can be challenging for individuals who are not familiar with statistical concepts. People may mistakenly think that there is a 95% probability that the true parameter falls within the interval, but this is not the correct interpretation.

2. **Misleading Precision:** a narrow CI may give the impression of high precision, but this can be misleading if the sample size is small or if there is substantial variability in the data.

3. **Assumption Dependence:** the construction of CI relies on certain assumptions about the distribution of the data and the statistical model. If these assumptions are violated, the confidence intervals may not be accurate.

4. **Coverage Probability vs. Actual Coverage:** the 95% confidence level means that, on average, 95% of confidence intervals from repeated sampling will contain the true population parameter. However, in any specific case, a confidence interval either contains the true parameter or it does not; there is no guarantee.

5. **Sensitivity to Outliers:** extreme values or outliers in the data can have a disproportionate impact on the width and position of CI, especially in small sample sizes.

6. **Choice of Confidence Level:** the choice of the confidence level (e.g., 95%, 90%) is arbitrary. While 95% is commonly used, different confidence levels will result in different intervals.

7. **Non-Normality Issues:** in some cases, CI based on normal distribution assumptions may not perform well for non-normally distributed data. Alternatives, such as bootstrapping methods, can be considered in such situations.

8. **Multiple Testing Adjustments:** when conducting multiple statistical tests, there is a risk of obtaining at least one "statistically significant" result purely by chance. Adjustments may be needed to account for this multiplicity, similar to the issues with p-values.

9. **Dependence on Sample Size:** larger sample sizes generally result in narrower CI, which may convey a false sense of precision. It's essential to consider the clinical or practical significance of the interval, not just its statistical properties.

10. **Non-Symmetrical Intervals:** in some situations, confidence intervals may not be symmetric around the point estimate, especially if the underlying distribution is skewed.

It's important for researchers and practitioners to be aware of these issues and to interpret confidence intervals in the context of their study design, assumptions, and the practical significance of the results. Additionally, using alternative methods, such as bootstrapping or Bayesian approaches, may be considered in situations where the traditional assumptions of normality or large sample sizes are not met.

4. WHAT SHOULD I CHOSE, CONFIDENCE INTERVALS OR p-VALUE?

The choice between using p-values or CIs depends on the specific goals of the analysis and the information that is expected to be collected. Both p-values and confidence intervals provide different types of information and can be valuable in different contexts.

Here are some considerations to help with the decision:

4.1 When to Use p-Values?

- ✓ **Hypothesis Testing:** if the primary goal of the investigation is to assess whether there is a significant effect, difference, or association, p-values are commonly used in hypothesis testing. Therefore, it is typically set a significance level (e.g., 0.05) and compare the p-value to this threshold.

- ✓ **Binary Decision:** p-values are often associated with binary decisions, which mean reject or fail to reject the H0. If statistical significance is a key aspect of one specific analysis, p-values can help to make this decision.

- ✓ **Comparisons and Group Differences:** in scenarios where it is being compared groups or tested relationships, p-values can provide a concise summary of whether the observed differences are likely due to chance.

4.2 When to Use Confidence Intervals?

- ✓ **Estimation of Parameters:** if the primary interest is to estimate the value of a parameter related to a population (e.g., mean, proportion), CIs are more directly informative, being that they provide a range of plausible values for the parameter.

- ✓ **Effect Size and Precision:** CIs provide information about the magnitude of an effect or difference, and its precision. A narrower interval indicates greater precision, while a wider interval suggests more uncertainty.

- ✓ **Visual Representation:** CIs are often visually intuitive and can be plotted on graphs. They provide a clear representation of the range within which the true parameter is likely to fall.

- ✓ **Avoiding Dichotomous Thinking:** CIs encourage thinking beyond dichotomous decisions. Instead of a binary reject/fail-to-reject decision, they allow the researcher to consider a range of possibilities for the true parameter.

4.3 Consider Using Both

- ✓ **Comprehensive Reporting:** in many cases, it's beneficial to report both p-values and Cis. This provides a more complete picture of the findings, addressing both the significance of effects and the precision of the estimates.

- ✓ **Avoiding Overemphasis on p-Values:** recognizing the limitations of p-values alone, combining them with CIs can help mitigate some of the issues associated with relying solely on significance testing.

- ✓ **Context Matters:** the choice between p-values and CIs should align with the specific goals and questions of the specific analysis. Consider the needs of the audience and the overall objectives of the study.

In summary, the choice between p-values and CIs depends on the nature of your analysis and the information you want to communicate. Using both can provide a more comprehensive and nuanced interpretation of your results. Additionally, considering effect size and practical significance alongside statistical significance is crucial for a well-rounded interpretation.

5. DISCUSSION

While p-values have been a valuable tool in statistical analysis, it's crucial to interpret them cautiously and consider them alongside other statistical measures and scientific context. P-values measure the strength of statistical evidence in many scientific studies. They indicate the probability that a result at least as extreme as that observed would occur by chance and are a way of reporting the results of statistical tests. But they do not define the practical importance of the results. They depend upon a test statistic, a null hypothesis, and an alternative hypothesis. In this regard, when designing scientific studies, it is important to consider:

- Multiple tests and selection of subgroups, outcomes, or variables for analysis can yield misleading p-values. Fortunately, full reporting and statistical adjustment can help avoid these misleading values.
- Negative studies with low statistical power can lead to unjustified conclusions about the lack of effectiveness of medical interventions.

Nowadays, researchers are increasingly recognizing the importance of transparency, robustness, and a more nuanced approach to statistical inference. Some argue that relying solely on p-values is not sufficient, and researchers should consider other statistical approaches which provide a more direct interpretation of evidence for or against a hypothesis.

In this context, confidence intervals remain a widely used and valuable tool in statistics. They provide a simple way to measure how well your sample represents the population you are studying, and are especially important where the results are not statistically significant. In fact, CONSORT guidelines include the reporting of CIs as part of the requirements for reporting results in clinical trials. However, the p-value is still mandatory to communicate and report the results of most studies.

But, at least, researchers are admitting that they have a problem and need to realize the limits of conventional statistics. They should instead bring into their analysis elements of scientific judgement about the plausibility of a hypothesis and study limitations that are normally banished to the discussion section: results of identical or similar experiments, proposed mechanisms, clinical knowledge and so on.

In this context, scientists need to try to respond three questions after conducting a study:

- What is the evidence?
- What should I believe?
- What should I do?

One method cannot answer all these questions.

"*The numbers are where the scientific discussion should start, not end*" Goodman.

6. REFERENCES

- Chia-Yun Chiang. What is the p-value? A detailed explanation of p-value. *Towards Data Science*. January 2021.

- Kwakkenbos et al. Protocol for the development of a CONSORT extension for RCTs using cohorts and routinely collected health data. *BMC*. 2018; 3:9. DOI: 10.1186/s41073-018-0053-3

- Mcleod. Confidence Intervals Explained: Examples, Formula & Interpretation. *Simply Psychology Statistics website*. October 2023.

- Moher et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomized trials. *BMJ*. 2010; 340: c869.

- Nuzzo, R. Scientific method: Statistical errors. *Nature*; 506, 150–152 (2014). DOI: 10.1038/506150a

- OECD, Recommendation of the Council on Good Statistical Practice, OECD/LEGAL/0417.

- C.O.S. Sorzano and M. Parkinson. Statistical experiment design for animal research. 2016.

- Wareet al., P Values. *Medical Uses of Statistics*, 1992; 2nd Edition. eBook ISBN 9780429187445

- Wasserstein & Lazar. The ASA Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, 2016; 70:2, 129-133, DOI: 10.1080/00031305.2016.1154108