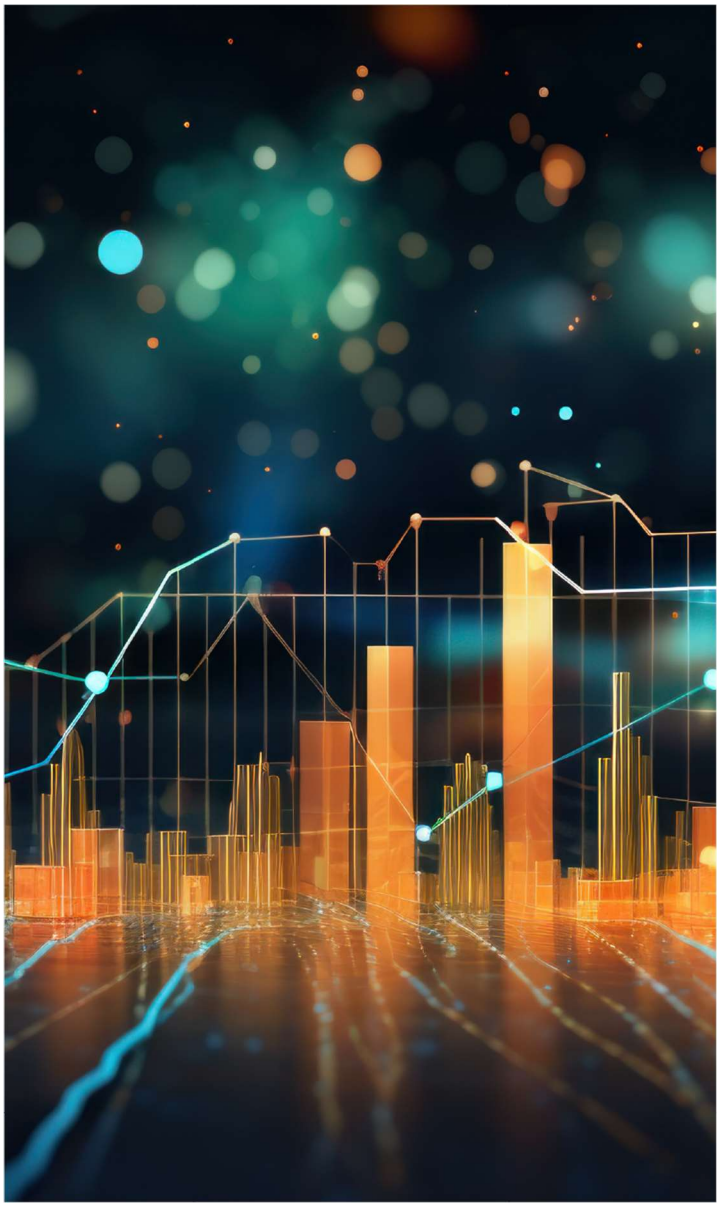
Biotech Essential Statistics: Understanding the pitfalls of P-vales and hypothesis tests.

P-values and hypothesis tests are ubiquitous in clinical research, but what do they really tell us about our data, and are they sufficient to guide informed decisions on treatment effect?

In the second of **Phastar's Biotech Essential Statistics webinars, Professor Jennifer Visser-Rogers, Vice President of Statistical Research and Consultancy,** explained the basis of p values, hypothesis tests, and statistical significance, before looking at their potential pitfalls and dispelling some common fallacies.

P-values: A potted history



R.A. Fisher first popularized the concept of a p-value in his 1926 book, Statistical Methods for Research Workers. **It is a method of measuring the strength of evidence against a null hypothesis, and is the probability, under the assumption of no effect, of obtaining a result equal to or more extreme than what was actually observed.** In essence, it enables analysts to untangle whether an observed effect has occurred by chance. A low p-value is accepted as evidence against the null hypothesis.

Later on, Neyman and Pearson proposed the idea of hypothesis tests, to provide a mechanism for making quantitative decisions and remove some of the subjectivity associated with Fisher's original idea. Making quantitative decisions introduces the notion of making the "wrong" decision and Neyman and Pearson argued that there were two types of errors that could be made in interpreting the results of an experiment. A type I error, or a false positive, occurs when a researcher rejects a true null hypothesis that is true in the population, and a type II error, or false negative, is failing to reject the null hypothesis when it should have been rejected. Under this updated framework, and with false positive results generally being seen as more dangerous to patients, a type I error rate (α) is set as a significance level to guide decision-making.

A p-value of <0.05 has been generally accepted as a convenient cut-off point for statistical significance. It means that there was a less than 5% chance the results obtained occurred under the null hypothesis of no effect, and a 95% chance they occurred because the alternative hypothesis was true

Magic number limitations

There are, however, a number of limitations to this approach, not least the arbitrary nature of the commonly used threshold.

While there is stronger evidence against the null hypothesis as the p-value becomes smaller, an α = 0.05 is just a convention that evolved from Fisher's and Neyman and Pearson's work and has no objective basis. Results that fall just a few decimal points on either side of the cut off, for example, are broadly similar, yet can be interpreted as vastly different, with 0.46 being statistically significant and 0.52 not being statistically significant.

The use of **0.05 as a "magic number"** that provides a **"passport to publication"** can also result in researchers' cherry-picking promising findings, also known as data dredging, or p-hacking, and leads to publication bias as a result of scientific journals' reluctance to publish negative results.

Many other fallacies exist, and it's important to be aware of them when analyzing and interpreting results.

Firstly, p-values do not measure the strength of evidence against the null, not the probability that the null is true. There is a reason why researchers use the terminology "reject" or "fail to reject" the null hypothesis. Rejecting the null hypothesis says that the results were not compatible with the null. Failing to reject the null doesn't mean that we "accept" the null as true. Rather, it simply means there is insufficient evidence in this study to reject it, but it doesn't commit us. There are numerous reasons why a trial could fail to reject a null hypothesis; it could be under-powered, for example. What's more, while it is true that effect size does influence the p-value, and larger effect sizes have smaller p-values, p-values do not, in themselves, tell us anything about the effect size. They are also influenced, for example, by sample size and measurement precision. With a big enough sample size, even the tiniest of differences could be statistically significant.

Clinical relevance

It is also crucially important to recognize that statistical significance does not imply clinical importance, or the impact on clinical practice. Large studies, for instance, can detect clinically unimportant, yet statistically significant findings, and vice versa.

This adds to the complexity of setting sample sizes, as well as the interpretation of results. To ascertain clinical importance, analysis needs to include an evaluation of benefit and risk. Whilst not without their own complexities, measures such as the **number needed to treat (NNT)** and the **number needed to harm (NNH)** can be useful tools here.

NNT is defined as the inverse of absolute risk reduction and enables statisticians to calculate the number of patients clinicians would need to treat in order to meet a particular endpoint. In a heart failure trial, for example, that endpoint may be first hospitalization or cardiovascular death. Similar calculations inform the NNH, which can be used to assess safety.

Protect the data

Summing up, Prof Visser-Rogers said there was a lot more to consider than just p-values and hypothesis tests when building, conducting, and interpreting clinical trials.

> Continuous data is seen as the gold standard of data collection and we do all that we can to collect the data in the best possible way," she said. "After all of this work… why would we then boil down our analyses and make our final decisions on the basis of a completely arbitrary pair of dichotomous categories?

Learn more. Watch the full webinar here