What a p-value can and *cannot* tell you

The p-value can tell you:

- the probability that a null hypothesis generated the data.

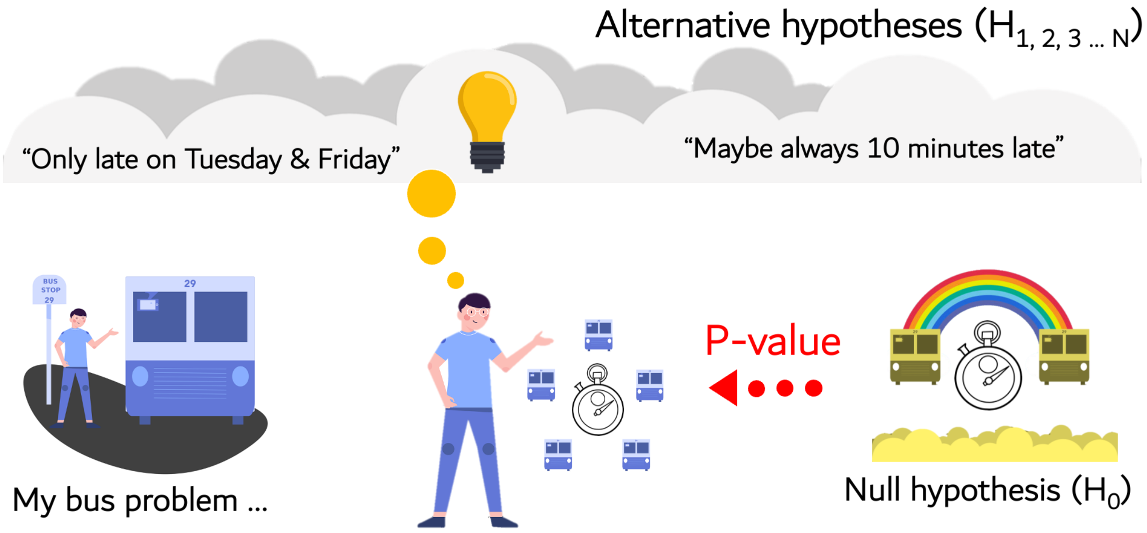
The p-value CANNOT tell you:

- the effect size.
- the quality of the alternative hypotheses.
- that there is no effect.
- the sampling plan.
- details about data pre-processing.

by Garston Liang, Quentin F. Gronau, & Alexander Thorpe

1 For the love of p-values

Love stories stand the test of time. Romeo and Juliet, Paris of Troy and Helen of Sparta, Emperor Hadrian and Antinous; surely, these are but a drop in a vast and ancient ocean. Each story contains their own crescendo, tragedy, and pathos, and the dynamics are as enthralling as the relationships are complex. So where in this vastness, might we find science's love affair with the p-value?


Alternative hypotheses ($H_{1, 2, 3 \ldots N}$)
"Only late on Tuesday & Friday"
"Maybe always 10 minutes late"
BUS STOP 29
29
My bus problem ...
P-value
Null hypothesis (H_0)

The humble p-value is a familiar travelling companion. Long is the road of scientific enquiry, and so familiarity fosters the comforts of convention. Yet, the p-value is often misunderstood, despite - or perhaps because of - its constant companionship with the measurement, consumption, and interpretation of scientific writing. Here, we revisit an old friend. In this primer, we explore what a p-value is, what a p-value tells us, and what a p-value cannot. Together, we hope to lift the rose-coloured spectacles of convention and to look honestly upon the p-value for what it truly is.

2 What does a p-value tells us?

The p-value helps us draw statistical conclusions about the state of the world by highlighting the probability or, more often, improbability of a single 'baseline' hypothesis. Formally, the p-value is the probability of obtaining the data or more extreme data, assuming that the null hypothesis is true.

To make the abstract more concrete, consider my bus dilemma in the diagram. Having recently moved, I want to know whether my local bus leaves on time. Most days it is late and others days it is on time. Besides the emotional frustration, I want statistical verification of the local bus company's claim that their buses run on schedule. Suppose each day for a month, I timed the bus past the scheduled arrival time (i.e., middle of the diagram). From this data, we can calculate a variety of summary values, such as an average waiting time. For demonstration purposes, lets suppose the average wait time was 20 minutes.

Before acquiring this data, however, we might have assumed that the average wait time should be *0* minutes, at least according to the bus company's claim. This baseline assumption is also known as the null hypothesis (signified as H_0). Shown on the right side of the diagram, the null hypothesis is a hypothetical and, at a minimum, possible state of the world. It serves as a baseline and generally represents a state of ignorance before any data is observed.

Once a null hypothesis is proposed, the p-value is a statistic that quantifies how probable, or conversely, how extreme our dataset would be if the null hypothesis generated our data. In the example, the p-value would present how likely was it to obtain a mean wait time of 20 minutes, if the bus runs on schedule. A low p-value means that these data were unlikely given the null hypothesis was true and so the real world must be quite different from the hypothetical null world. In other words, we reject the null as a viable state of the world. Statistically, we would conclude that the bus does not run on time ($p < 0.05$) and propose new alternative hypotheses about the buses actual schedule, shown above the clouds in the diagram.

3 What a p-value cannot tell you

The p-value cannot tell you about whether the null hypothesis was a good baseline. Null hypotheses can be derived from many different sources ranging from expected outcomes from chance (e.g., coin tosses should converge to 50% heads) to informed specific values based on previous observations (e.g., my next-door neighbour says the buses are 2 hours late).

Similarly, the p-value cannot tell you about the quality of alternative hypotheses. The logic of the above process, called Null Hypothesis Significance Testing, is such that alternative explanations are proposed only after the null hypothesis is rejected. However, the p-value does not make contact with the new explanation and so cannot tell you whether the new hypothesis offers a better account of the data. Along a similar vein, a p-value cannot tell you that there is 'no difference' in the data. If the null hypothesis is retained, this is not evidence that the null is true, only that the data *could* have been generated from the null hypothesis. Note, for a Bayesian framework that does account for both alternative hypotheses and evidence for the null, see the Bayes Factor Primer by Gronau et al. in this series.

A p-value can be calculated for any hypothesis but assessing the quality of the null and alternative hypotheses is a task for the reader.

3.1 P-value is not an effect size

The p-value is not a measure of effect size. The fact that an effect is statistically significant can be divorced from what makes an effect clinically important.

In the bus example, I could have collected a large sample of data (i.e., recording times everyday of the year) and statistically determined that the bus was late everyday by 1 minute. Although, this effect would be statistically different from the 0-minute null hypothesis, who would care? In some scenarios, small effects may be important whereas in others, only large effects are worthwhile pursuing. Clinical judgement of the effect size is therefore highly informative alongside, but distinct from, the p-value.

3.2 P-value cannot tell you about the sampling plan

Before the p-values are calculated, experimenters must decide on how to obtain data from their population. Many decisions are made at this point, such as the ideal sample size which affects statistical power, or whether to stop data collection based on time, i.e., next three months, or until a minimum number of subjects are obtained, e.g., at least 20 patients. Although these decisions are made prior to analysis, they affect the sampling distribution of the data, and consequently the p-value. Interrogating these methodological details would be prudent.

3.3 P-value cannot tell you about data processing

At the data-level, the p-value is silent with respect to how the data were processed. Once in custody, a dataset can be interrogated and shaped through transformations, median splits, or outlier removals to name just a few techniques. While each of these steps may well be necessary for data analysis, it is important that the reader is able to assess the complete chain of events because the p-value bears no trace of the original data prior to processing.

4 Further reading

Below, are three resources that discuss issues around p-values in greater detail. These include:

- Effect sizes in greater detail and why the p-value is not enough:
  Sullivan, G. M., & Feinn, R. (2012). Using effect size—or why the P value is not enough. Journal of graduate medical education, 4, 279-282.

- The problem with interpreting effects strictly by significance (< 0.05):
  The p -value fallacy and how to avoid it. *Canadian Journal of Experimental Psychology / Revue canadienne de psychologie expérimentale*, 57, 189–202

- Common problems with p-values and a Bayesian alternative approach:
  Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic bulletin & review*, 14, 779-804