TRL THE FUTURE OF TRANSPORT

The use of hypothesis testing in transport research

Jonathan Kent & Caroline Wallbank

Death of the p-value?

In 2015, the academic journal Basic and Applied Social Psychology (BASP) announced that the null hypothesis significance testing procedure was 'invalid', and therefore implemented a ban on the use of it in future submissions. Their reasoning behind the decision was that the criteria for a statistically significant result, namely obtaining a p-value < 0.05, is a "bar (that) is too easy to pass and sometimes serves as an excuse for lower quality research'"[1].

Throughout the literature (in transport research and in other domains) 'p-values' are often used to determine if a finding is meaningful or not, but should they still be used, or are BASP right to suggest that they are now dead in the water? What are p-values? How should these be interpreted and what are the common misconceptions? What alternatives are there and when should these be used instead?

This paper answers these questions and comments on whether p-values are still relevant in transport research.

[1] https://www.tandfonline.com/doi/full/10.1080/01973533.2015.1012991

Hypothesis testing is a statistical method to evaluate whether a proposition is true or false. How useful is the p-value in transport research?

What is hypothesis testing?

Before turning our examination to p-values, we first need to explain what hypothesis testing is and how this relates. Hypothesis testing is a well-used statistical method to evaluate whether a proposition is true or false, based on the level of empirical evidence for or against it. It begins with the underlying assumption that a null hypothesis, H_0, is true. The purpose of the test is to determine whether there is sufficient evidence to reject this null hypothesis in favour of an alternative hypothesis, H_1, based on the value of an appropriate test statistic, T.

The orange box on the next paage contains two examples from the field of transport research: we've specified what the null and alternative hypotheses and test statistic would be in each case. Both examples are taken from recent TRL research, one from on-road trials of 60mph speed limits through roadworks on the Strategic Road Network, the other from the MOVE_UK project which contributed to the progression towards automated driving through connected systems validation and analysis of big data.

The first example is a one-tailed test, where the difference between H_0 and H_1 is specified in a given direction (i.e. flow is higher at 60mph). The second example is a two-tailed test, where no direction is specified. The test is simply looking for any difference in performance between the Traffic Sign Recognition systems, without specifying which system is expected to be better.

The hypothesis test starts by assuming that H_0 is true, then takes some observed data and calculates the test statistic, which is assumed to follow a particular distribution.This distribution depends on the context of the problem. From this, a p-value can be calculated, which represents the probability, under the null hypothesis, of the test statistic being at least as extreme as the value that has been observed. For instance, in the first example, if the observed average flow is 5 vehicles per minute higher at 60 mph than at 50 mph, then T = 5 and the p-value is the probability that $T \geq 5$, based on the assumption under the null hypothesis that there is no difference in average flow.

Example 1

Does traffic flow improve when the speed limit through a set of roadworks is increased from 50 mph to 60 mph?

H_0: The average vehicle flow per minute through the section of roadworks is the same whether the speed limit is 50 mph or 60 mph

H_1: The average vehicle flow is higher when the speed limit is 60 mph, compared to when it is 50 mph

T = The average flow at 60 mph, minus the average flow at 50 mph

Example 2

Does traffic flow improve when the speed limit through a set of roadworks is increased from 50 mph to 60 mph?

Is there a difference between the sign detection rates of two Traffic Sign Recognition systems, A and B?

H_0: The sign detection rates of systems A and B are equal

H_1: The sign detection rates of systems A and B are not equal

T = The difference in sign detection rates between systems A and B

What does the p-value actually mean?

The p-value(s) associated with the test being performed are interpreted based on whether they are above or below a threshold, α. In most cases, tests are carried out at the 5% significance level (or, equivalently, the 95% confidence level), meaning that $\alpha = 0.05$. If the p-value is below α, the null hypothesis can be rejected in favour of the alternative, whereas if it is above α, then it cannot be.

However, while the notion of a p-value sounds straightforward, some care should be taken when interpreting it and evaluating what it means. In particular, a common misconception is to say that if a p-value is above the threshold, α, then the alternative hypothesis can be rejected as being false and the null hypothesis can be accepted as being true. However, this is not the case. The only conclusion that can be drawn is that there is insufficient evidence to reject the null hypothesis in favour of the alternative.

While the notion of a p-value sounds straightforward, care should be taken with interpretation to avoid common misconceptions.

How to use and report p-values

The p-value is an important part of a hypothesis test, as it provides a quantitative measure which can be used to determine whether or not to reject the null hypothesis. However, in many cases, drawing meaningful conclusions from the test is not as simple as interpreting this p-value. If the sample size is large then very small differences between groups can be classified as significant (and we would reject the null hypothesis that the two groups are the same), but these differences may not be meaningful in the context of the problem.

This is why it is important that effect sizes are considered alongside the p-value to give a fuller picture of what the data is showing. In the context of hypothesis testing, an effect size provides a means of assessing the magnitude of a difference. In the examples on page 4, this is a difference between the two samples.

There are a number of different effect size metrics; one of the most common is Cohen's d. For the examples on page 4, this metric is calculated as the difference between the means of the two samples, divided by their pooled standard deviation. Typically, if no other frame of reference is available, Cohen's d figures of 0.2 are considered to be 'small' effects, 0.5 are considered to be 'medium' effects and 0.8 are considered to be 'large'.

Other effect size metrics exist, such as Hedges' G or Glass' ∆, which use slightly different estimates of standard deviation. Another alternative is eta squared (or, similarly, partial eta squared), which looks at the proportion of variability in the data that can be attributed to the effect of interest.

An effect size can add a considerable amount of extra insight to a p-value: it considers the sample size and variability within the data, and is evaluated in terms of what is meaningful within the context. In addition, the level of uncertainty in each effect size can be determined by looking at the corresponding standard error.

Example

The Driver2020 project is a world-first randomised controlled trial which will measure the effectiveness of five different interventions designed to reduce novice driver crash risk in the first year of solo driving. The project, which completes in 2022, will enable the UK Government to make evidence-based decisions as to which interventions are considered for roll-out either as part of the licensing system or as part of voluntary uptake.

The study has been designed around the principles of hypothesis testing: testing to see if there is a significant difference in collision involvement between participants who receive an intervention and those who don't.

What are the advantages and disadvantages to p-values?

Advantages

1. **It gives a measure of whether or not a difference is statistically significant**

   In a typical hypothesis test, the observed data usually consists of sample(s) from a population(s). To take example 2 on page 4, there are two populations; the set of vehicles which have system A installed, and those which have system B installed. To answer the research question sign detection data would be collected from two samples of vehicles within these two populations. For example, if each population contains 1000 vehicles, we might select a sample of 50 from each. As a result, the dataset would contain some sampling error to account for any differences between each sample and its corresponding population.

   Therefore, even if the null hypothesis were true, there would still be some differences between the two samples, due to this sampling error. A hypothesis test evaluates whether the observed difference is due to this variability, or whether it is substantial enough to be attributable to a systematic difference between the two populations.

2. **No requirement for previous scientific theory**

   There is no stipulation for a hypothesis test to be underpinned by previous research in the same field, or by prior beliefs about whether a not a particular relationship might exist. This increases the range of fields in which the method can be applied, and allows hypothesis testing to be used in novel research areas. Having said that, if findings can be shown to corroborate with relevant literature, this will help to support the message that is being communicated.

   In instances when it is beneficial (or even necessary) to take prior research or beliefs into account, the alternatives to hypothesis testing described below, should be considered.

Disadvantages

1. Publication bias – only "significant" p-values get accepted in the literature

This is arguably the biggest challenge with using hypothesis testing. Many research projects are framed with a desired outcome in mind, and if this conclusion cannot be definitively drawn from statistically significant findings, then the potential for the research being accepted for publication may be limited. This has a number of drawbacks, of which we mention two here. Firstly, there are many occasions when results which are not statistically significant can still produce interesting findings. This might be because the results are different to what the reader would have expected, or because the data still supports a particular conclusion, but not to the extent that it can be deemed to be conclusive.

Secondly, and more importantly, there is a danger that researchers prioritise obtaining a significant result over following a rigorous approach. For example, it is important that the distributional assumptions underpinning the hypothesis and test statistic are checked against the dataset under consideration, before the test is carried out. If these assumptions do not hold true, then findings which are presented as statistically significant may be invalid.

2. The choice of α is arbitrary

The most common practice is to set $\alpha = 0.05$, although in some cases it can be appropriate to set a lower value, such as $\alpha = 0.01$, or a higher value, such as $\alpha = 0.10$, depending on the context of the problem under consideration. However, these thresholds are chosen arbitrarily, rather than having a theoretical basis for which differences should be classified as being "significant".

3. The null hypothesis is often unlikely to be true

There is a risk of this claim being made if the hypotheses underpinning the test are not defined carefully enough. Consider Example 1 on page 4: it can be argued that changing the speed limit through a set of roadworks is inevitably going to have an impact on driver behaviour, particularly in terms of the speed at which they drive. This will consequently have an impact on the flow of vehicles through the section. Therefore, the null hypothesis in this instance is probably highly unlikely to be true anyway.

What are the alternatives?

Equivalence testing

An equivalence test (also known as a non-inferiority test when the test is one-tailed) is a variation of a hypothesis test, based on a predetermined judgment of how large a difference should be in order to be considered 'scientifically relevant' within the context of the problem being considered. This threshold, δ, is typically defined in advance based on previous research in the same field. The null hypothesis is that a difference between two populations is greater than δ, with the alternative hypothesis being that the difference is less than δ.

Equivalence testing can be performed either on its own, or as a supplement to a standard hypothesis test, to determine from a statistical point of view whether the absence of a difference can be confirmed to be true. It helps to guard against the danger of the null hypothesis in a standard test being accepted as true, simply because there is insufficient evidence to reject it.

One advantage of this approach is that the choice of δ is based on previous research, rather than being arbitrarily chosen like α is in the context of hypothesis testing. However, determining a suitable value for δ can be hard if this previous research is not available.

Bayesian hypothesis testing

Traditional hypothesis testing is known as a frequentist approach, in that the outcome of the test is based solely on the observed data that is collected. In contrast, Bayesian hypothesis testing includes information on prior beliefs about the relative probabilities of the null and alternative hypotheses being true.

A Bayesian hypothesis test begins with a null hypothesis, H_0, and an alternative hypothesis, H_1, as in the standard approach. However, it also requires a prior odds ratio, that is, the prior belief about how likely H_1 is to be true, relative to H_0. For example, a prior odds ratio of 0.5 means that prior to any data being observed, the researcher believes that H_1 is half as likely to be true as H_0 (i.e. H_0 is twice as likely to be true as H_1). Alternatively, a ratio of 3 means that they believe H_1 is three times more likely to be true than HE_0.

Several other statistical methods can be applied instead of, or in addition to, hypothesis testing.

Once data has been observed, a *posterior odds* ratio is then calculated. This represents the likelihood of H_1 being true, relative to H_0, given the data that has been seen. This is then divided by the prior odds ratio to give a Bayes factor. This factor represents how much more plausible H_1 is, relative to H_0, than it was before the data was observed. Large values of the Bayes factor (e.g. > 10) indicate strong evidence for the alternative hypothesis, H_1, whereas small values (e.g. $< 1/10$) indicate strong evidence for the null hypothesis, H_0.

An advantage of this approach is that it provides a measure of the magnitude of evidence in favour of H_0 or H_1. A standard hypothesis test simply provides a 'yes-or-no' answer as to whether or not H_0 should be rejected in favour of H_1. This means that it can only be used for this purpose and cannot be used to determine whether or not H_0 can be accepted. By contrast, the Bayesian approach can assess the feasibility of both hypotheses in one test, without having to set an arbitrary threshold like α. It can also be extended to test more than two hypotheses by comparing each pair of them in turn.

However, this approach is less suitable if the test needs to provide a definitive outcome, or if there is lack of prior research or knowledge from which a prior odds ratio can be determined.

Conclusion

To respond to BASP's claim, are p-values 'invalid' and no longer fit for purpose? The answer from this paper is 'no'. When used and interpreted properly, and when alternatives are considered where these might be more suitable, hypothesis testing is still a valid statistical method to use in many circumstances, including in transport research.

TRL's experienced statisticians understand that it is not as simple as looking at a p-value to see if it is less than 0.05: hypotheses need to be carefully defined so that the outcome of the test is meaningful; assumptions underlying the test need to be checked and verified; p-values should be clearly explained so that their interpretation is clear, and further supporting evidence, such as effect sizes, should be looked at before conclusions are drawn.

On top of this, there are other approaches which should be considered, such as those explained above, particularly when prior knowledge or supporting research is available. While standard hypothesis testing is appropriate in a lot of contexts, it should not be seen as the only possible method.

P-values are still valid and fit for purpose. However, care should be taken to use and interpret them properly, and alternative approaches should also be considered.

THE FUTURE
OF TRANSPORT

Hypothesis testing is a well-used statistical method to evaluate whether a proposition is true or false. A fundamental part of the testing procedure is the calculation and interpretation of a p-value, which represents the probability of a set of data being observed, under the assumption that the proposition is true. This null hypothesis is then rejected if the p-value is less than a certain threshold, usually 0.05.

In recent years, some members of the scientific community have called into question the validity of the hypothesis testing approach, because it places so much emphasis on whether or not a value is above or below an arbitrary threshold. We think that hypothesis testing is still a valid method, but it is important that, as well as the p-value, additional information such as effect sizes is taken into account when interpreting results. In addition, there are alternative approaches, such as equivalence testing or Bayesian hypothesis testing, which should be considered in certain circumstances.

ACA012
ISBN: 978-1-913246-86-0

enquiries@trl.co.uk
www.trl.co.uk

TRL Crowthorne House, Nine Mile Ride,
Wokingham, Berks, RG40 3GA