Statistical significance in choice modelling: a commentary on the use of confidence intervals, *t*-ratios, *p*-values and *star* measures

Stephane Hess* Andrew Daly† Ricardo Daziano‡ Michiel Bliemer§

September 6, 2020

Abstract

This note offers a commentary on the use of notions of statistical significance in choice modelling. We argue that, as in many other areas of science, there is an over-reliance on 95% confidence levels. We also note a lack of precision in the reporting of measures of uncertainty in many studies, especially when using *p*-values and even more so with *star* measures. We suggest that these should only ever be presented alongside standard errors or *t*-ratios. The note also highlights the distinction between looking at the precision of estimates and statistical significance, and stresses the importance of considering behavioural or policy significance in addition to statistical significance.

Keywords: choice modelling; standard errors; *t*-ratios; *p*-values; confidence intervals

1 Introduction

There is growing criticism of the over-reliance on statistical significance, and its misuse, in several branches of science, calling for care in interpretation (see e.g. Wasserstein and Lazar, 2016), the use of much stricter tests (Benjamin et al., 2017), or even a complete abandonment of the notion of statistical significance (Amrhein et al., 2018; Wasserstein et al., 2019). Against this background, the timing seems apt to offer a commentary on this issue within the field of choice modelling.

This note focusses on how measures of significance and uncertainty are computed in choice modelling, how they are used in taking decisions on model specification, and how they are reported. In doing so, we question the strict adherence to especially the 95% level of confidence, misconceptions as to what this implies, and lack of precision in some reporting practices. We also stress the need for more precise language in describing the outcome of significance test and further call for analysts to not focus on statistical significance alone, but also consider the precision of estimates as well as the significance of a finding from a behavioural or policy perspective. While the points raised in this note will be obvious to many choice modellers with a strong econometric background, this seems to not be the case in the more applied community, and there are frequent examples of papers using questionable approaches or indeed reviewers asking authors to do so.

While a Bayesian perspective is offered in the concluding points, our discussion primarily centres on classical estimation and inference. Some basic understanding of maximum likelihood estimation is thus helpful for the remainder of this paper. By means of background, the frequentist view holds that, for each single parameter β, there exists, subject to the chosen model specification, a unique true value β^*, but that our data is incomplete and we thus have sampling error. The use of maximum likelihood estimation of a choice model yields an estimate $\hat{\beta}$ for parameter β, with a standard error σ_β. As the sample size N increases, the maximum likelihood estimate (MLE)

*s.hess@leeds.ac.uk; Institute for Transport Studies and Choice Modelling Centre, University of Leeds
†andrew@alogit.com; Institute for Transport Studies and Choice Modelling Centre, University of Leeds
‡daziano@cornell.edu; School of Civil and Environmental Engineering, Cornell University
§michiel.bliemer@sydney.edu.au; Institute of Transport and Logistics Studies, The University of Sydney

$\hat{\beta}$ converges to a normal distribution around the true values β^*, i.e. $\sqrt{N}\left(\hat{\beta}-\beta^*\right) \rightarrow \mathcal{N}(0,\sigma_\beta)$, which is what is understood as asymptotic normality.

In the context of the present note, a number of general points can be made. First, the above properties relate to sampling error rather than errors in the specification of the model or in the data. For example, if an analyst uses a linear specification when the true specification is non-linear, or if the data is affected by hypothetical or strategic bias, then the true values β^* for our model are themselves biased. The standard error for $\hat{\beta}$ simply reflects the uncertainty in our estimated values, rather than being related to bias of the true values. Second, given finite sample sizes, the key property of MLEs is one of asymptotic normality, rather than normality, and this thus applies only in a narrow region around the estimate. Third, as discussed in detail by Daly et al. (2012), error measures can similarly be computed for functions of different model parameters, such as differences or ratios, using the Delta method, with the same underlying maximum likelihood estimate properties. Our discussions in terms of confidence intervals and hypothesis testing thus apply in the same way to functions of parameters (e.g. willingness-to-pay) as to individual estimates. Finally, robust standard errors (i.e. those calculated using the sandwich matrix) are used in many applications, and it should be noted that a) the same considerations in terms of asymptotic normality apply, b) robust standard errors can be used as inputs to the Delta method too, with the same properties, and c) the observations in the present note in terms of reporting significance apply in the same way when using robust standard errors.

The remainder of this note is organised as follows. We look separately at confidence intervals (Section 2), hypothesis testing for parameters or functions thereof (Section 3) and model comparisons (Section 4). We then provide an empirical example in Section 5 before Section 6 offers some conclusions.

2 Confidence intervals

Standard errors relate to parameter uncertainty and a natural use of them thus comes in computing confidence intervals, either around the estimated values of individual parameters, or the computed value of functions of multiple parameters. A $C\%$ confidence interval gives us a range of values around the estimated value $\hat{\beta}$ where we can have a $C\%$ confidence that this contains the true value β^*. Relying on the property of asymptotic normality of the MLEs, a $C\%$ confidence interval for β could thus be obtained as $\hat{\beta} \pm z^C \sigma_\beta$, where z^C is the upper $\frac{1-C}{2}$ critical value for a $\mathcal{N}(0,1)$ distribution, e.g. using $\hat{\beta} \pm 1.96\sigma_\beta$ for a 95% confidence interval.

The use of standard errors in the computation of confidence intervals highlights their role in relation to the *precision* of parameter estimates, which has a different focus from that of significance testing. The smaller a standard error is relative to a parameter estimate, the more precise that estimate is. It is insightful in this context to express the width of one side (i.e. half) of a confidence interval as a percentage of the MLE, i.e. $\hat{\beta}$, where, for a $C\%$ confidence interval, this is given by $\frac{z^C \sigma_\beta}{|\hat{\beta}|}$. An example illustration of the confidence intervals for various relative measures of σ and $\hat{\beta}$ is given in Figure 1. Except for the highest value of $\frac{\sigma}{\hat{\beta}}$, all the other values would mean that the 95% confidence interval contains strictly positive values only, often seen to imply

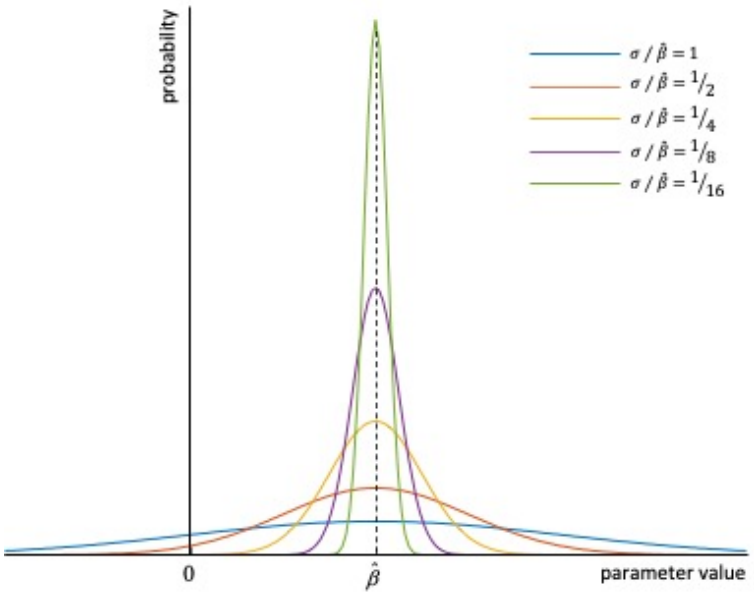

Figure 1: Parameter precision ($\frac{\hat{\beta}}{\sigma}$) and asymptotic distributions

"statistical significance". However, the distinction between significance and precision is clear - in practice, we would want much narrower confidence intervals, such that the entire range (or most of the range) represents an effect that is meaningful for real world application.

We now turn to an issue that has seen past discussion in the choice modelling literature (e.g. Armstrong et al., 2001), but remains poorly understood and largely ignored, namely the distinction between normality and asymptotic normality. Just as for other uses of standard errors that we discuss later on, the computation of confidence intervals relies on the property of normality, while maximum likelihood estimation assures only asymptotic normality, which applies close to the optimum.

To be specific, if the estimated parameters were distributed exactly normal, this would imply that the log likelihood function was quadratic in the parameters. A shift away from the optimum $\hat{\beta}$ for parameter β by a value $\Delta\beta$ would reduce the log-likelihood by $\Delta LL = \frac{\Delta\beta^2}{2\sigma^2}$. However, the 'asymptotic' qualifier means that the property applies exactly only for very large data sets. For finite data sets, it is an approximation that applies only in the neighbourhood of the optimum. Indeed, as is shown by Armstrong et al. (2001), for example, the log-likelihood function is not exactly quadratic, so that reductions in likelihood do not follow this formula exactly at all. What happens as we move further away from the optimum is not defined by maximum likelihood theory. The parameter estimates are not distributed exactly normal, and any theoretical insights/claims based on an assumption of normality are misguided. The problem is that what defines *"close"* is unclear and depends on the model. There is thus no guarantee that the shape of a $C\%$ confidence interval (especially for large C) is normal, that the bounds are at or close to $\hat{\beta} \pm z^C \sigma_\beta$ or that the bounds are symmetric around the MLE, $\hat{\beta}$.

If an analyst wishes to compute confidence intervals without relying on any assumptions about normality, asymptotic or otherwise, then an alternative approach is to rely on bootstrapping. The Bootstrap operates by sampling N observations from the original sample of N observations, with replacement, repeated a number of times as chosen by the analyst, say leading to S samples. In the context of data with multiple observations per individual, it makes most sense to sample at the level of individuals rather than observations. The concept on which the Bootstrap is based is that, if the original data is a representative sample from the population being studied, then the Bootstrap samples also resemble samples that might be drawn if the sampling were done again. For that reason, they give the sampling variance that may be expected. Individual models are then estimated yielding S sets of parameter values, and these can be used to produce an empirical confidence interval for each parameter, e.g. what range of values will contain 95% of the estimates values. The covariance of the parameters over the Bootstrap samples can be used as the covariance matrix of the parameter estimates, and the square root of its diagonal can be used instead of the classical or robust standard errors - using these for a theoretical rather than empirical confidence interval then of course again requires an assumption of normality.

3 Hypothesis tests for parameters and functions of parameters

The second focus of our note is on statistical tests applied to individual model parameters or functions of parameters. In particular, this relates to testing the hypothesis that a given parameter or function of parameters is equal to a specific value. The key focus is on a test where the null hypothesis is that a parameter is equal to zero[1], i.e. $H_0 : \beta = 0$. What is commonly described as a test of significance is thus a test to see whether the null hypothesis of a parameter being equal to zero can be rejected.

Hypothesis testing of this type is conveniently carried out by the use of a t-ratio[2], which compares the estimate $\hat{\beta}$ against a value q, where $t_{q,\hat{\beta}} = \frac{\hat{\beta}-q}{\sigma_\beta}$. We then have the null hypothesis that $H_0 : \beta = q$ and the alternative $H_1 : \beta \neq q$. The absolute value of the test-value $t_{q,\hat{\beta}}$ is compared to a critical value from a $\mathcal{N}(0,1)$ distribution to indicate the level of confidence C at which H_0 can be rejected.[3].

The key use of t-ratios is of course with $q = 0$, i.e. giving $t_{0,\hat{\beta}}$ which is used to test whether a parameter is "significantly different from 0". It is in the context of such significance tests that the language used in choice modelling (and other disciplines) is often imprecise or in fact incorrect from a statistics perspective. It should first be noted that hypothesis testing is only concerned with whether the null hypothesis can be rejected, not whether the alternative hypothesis is true. In other words, it establishes whether our finding could be due to chance and that the true value is in fact zero. The statistical significance of the test relates to the probability of H_0 being rejected

[1]Other variants exist, for example looking at whether the difference between two parameters (potentially from different models/datasets) is zero, or whether multipliers or nesting parameters are equal to 1. The same logic applies though the Delta method may be needed to compute standard errors for functions of parameters.

[2]A t-ratio is also known as t-statistic or t-value. The term t-test should not be used when reporting t-ratios.

[3]With the sample sizes usually used in choice modelling, the distinction between the Normal distribution and the t distribution can be ignored.

when it is in fact true, i.e. a type I error. It is thus incorrect to talk about a parameter being 95% significant. The significance level in this case is 5%, and we can reject H_0 with a 95% level of confidence, not significance. The significance level is commmonly expressed using p values, where, with the level of confidence being C, $p = 1 - C$ gives the probability of observing this (or a more extreme) outcome (i.e. a value of $\hat{\beta}$ if the null hypothesis of $H_0 : \beta = q$ is true). The relationship with the earlier discussion on confidence intervals is clear. Using a negative estimate of $\hat{\beta}$ as our example, if we can reject $H_0 : \beta = 0$ with a confidence level of C%, then this means that 0 forms the upper bounds of the C% confidence interval around the estimated value $\hat{\beta}_k$. This also highlights that t-ratios are again reliant on normality, while MLE only have the property of asymptotic normality.

We next look at three key questions that analysts face in applied work.

Should one or two-sided tests be used?

For confidence intervals, it clearly makes sense to work with a two-sided confidence interval, looking at the distribution to either side of the estimated value. In significance testing, analysts commonly rely on a critical value of 1.96 to imply a 95% confidence level. The source of this is in the fact that 2.5% of the probability of a standard Normal distribution lies below -1.96, and 2.5% lies above 1.96, hence the earlier point about a 95% confidence interval being given by $\hat{\beta} \pm 1.96\sigma_\beta$. We argue that a one-sided test for significance is often more appropriate. For example, if a model contains a parameter whose sign is known a priori, such as a (negative) cost coefficient, then a test whether the parameter has a value different from zero would naturally be one-sided. To be specific, let us assume that for a cost coefficient β_c, we have $\frac{\hat{\beta}_c}{\sigma_{\beta_c}} = -1.8$, i.e. less than the 'critical' value of 1.96. If an analyst considers that this parameter is therefore not significantly different from zero, so that the corresponding variable should be considered for removal from the model, then this implies that we would also consider estimates that are below $\hat{\beta}_c - 1.96\sigma_{\beta_c}$ to be unacceptable, i.e. those that are very negative in this case. Most often, this is not what we want and we would reject only values that have the wrong sign: the appropriate critical t-ratio value for a 95% one-sided significance test is then approximately 1.64 and a t-ratio of 1.8 would thus imply that the estimate is in fact 'significantly different' from zero. A value of $t_{0,\hat{\beta}_c} = -1.96$ would imply that there is a 2.5% probability (under repeated sampling) for the cost coefficient being positive, not a 5% chance. We mention this point because these values often seem to be misused in practice, and in fact serve as the default in several existing software packages[4].

How should significance be reported?

The next consideration for an analyst is how significance levels should be reported, where we raise three key issues, namely interpretation, numerical precision and further use of results.

Remember that estimation gives values for $\hat{\beta}$ and σ_β, i.e. estimates and standard errors. If an analyst includes these two outputs in reporting of the results, then this ensures that a reader is able to further process the results. The same applies if an analyst reports $\hat{\beta}$ and $t_{0,\hat{\beta}}$,

[4] Apollo (Hess and Palma, 2019) uses one-sided tests by default but gives the user the option to request two-sided tests instead.

i.e. the estimates and the t-ratios for parameters, as a reader can calculate $\sigma_\beta = \frac{\hat{\beta}}{t_{0,\hat{\beta}}}$. From this perspective, reporting either standard errors or t-ratios raises no further issues in theory. A practical consideration arises in terms of the numerical precision that is used in presenting the results. As standard errors are often an order of magnitude smaller than parameter estimates, the use of too low a number of digits will mean valuable information is lost in reporting results. For t-ratios, the use of one or two digits after the decimal point is generally sufficient, but for standard errors, analysts should ensure to report at least two significant digits (i.e. not counting leading zeros).

The situation with reporting becomes more complicated if an analyst reports p-values instead of t-ratios and standard errors, as is common practice in some disciplines, such as health, while being rare in others, such as transport. First, given the above points about one-sided *vs* two-sided tests, if an analyst reports p-values only (i.e. without standard errors and/or t-ratios), then, in order to avoid misinterpretation, information must be provided on whether these related to a one-sided or a two-sided test. Second, an issue with precision and further use of the results arises. Let us assume that we have $\mid t_{0,\hat{\beta}} \mid > 10$, in which case, whether using a one-sided or a two-sided test, an analyst is likely going to report $p = 0$ or $p < 10^{-5}$. Of course, the test clearly rejects $H_0 : \beta = 0$, but reporting p-values alone will prevent a reader from using the results to compute confidence intervals or to gain any insights into differences in uncertainty for two parameters which both have $p < 10^{-5}$.

The issue is further compounded when an analyst relies on *star* measures instead of p-values, e.g. using * for 90% confidence, ** for 95% confidence and *** for 99% confidence. This again requires reporting whether one or two-sided tests were used, but issues with precision and further use of the results are even more severe than with p-values, given the arbitrary division of p-values into three categories. The further use of results for e.g. confidence intervals is impossible when significance levels are reported only with *star* measures, even in those cases where with p-values, this would be possible still (i.e. if the reported p is shown with enough digits).

A key motivation for including p-values is clearly to allow readers to quickly see significance levels without having to compare t-ratios to critical values. The same applies to *star* measures, which are a somewhat more qualitative reporting tool as they simply split parameter estimates into four groups (i.e. $< 90\%$ confidence, 90% to 95%, 95% to 99%, and $> 99\%$). The use of these measures, accompanied by information on whether one-sided or two-sided tests are used, can thus be seen as providing additional information to readers. It is fine to report these measures alongside but not instead of standard errors or t-ratios, but this of course goes against the motivation often given for the use of *star* measures that it makes tables less cluttered and more readable.

How much attention should be paid to significance levels?

Looking next at tests for individual parameters (or functions of parameters), a case can be made that 95% confidence is actually a low level. With the large datasets used in many studies on revealed preference (RP) data, or with good experimental designs for stated choice (SC) data, t-ratios in excess of 10 or 20 are often the rule rather than the exception, and t-ratios of 1.96 (for a typical two-sided test) are thus low. Let us also consider real-world implications, and imagine a situation where a parameter estimate has a value of $t_{0,\hat{\beta}} = 4$, well into the territory of rejecting

$H_0 : \beta = 0$ at the 99% confidence level. From the discussions in Section 2, we then have that the width of a 95% confidence to either side of $\hat{\beta}_k$, expressed as a proportion of the MLE, is $\frac{1.96\sigma_\beta}{|\hat{\beta}|} = \frac{1.96}{4} = 0.49$. This would imply that a 95% confidence interval runs roughly from 50% of the estimated value to 150% of the estimate, which is not much use if we consider using the information to justify a multi-billion pound investment.

While the above has argued for not overstating findings with high levels of significance, there are conversely also cases where lower levels of significance should not be seen as a reason for simply removing parameters from a model. Indeed, a parameter can be behaviourally meaningful and have a non-trivial impact on behaviour in a model even if it doesn't pass the 95% threshold. Or the inclusion of a parameter may be justified on the basis of it being an important policy-test input, providing the value is plausible. Sample size effects also come into play. While good practice in model selection, as discussed above, implies expecting larger improvements in fit with larger samples (and accepting smaller improvements with smaller samples), little consideration is given to this point in the context of parameter significance. Of course, large samples are always desirable, but not always possible, for example when looking at niche applications with a limited pool of potential decision makers to include in the sample. In that case, less stringent significance levels may well deserve consideration and the a-priori expectations for the influence of a variable on choice become even more relevant.

The above two paragraphs suggest that analysts should look further than statistical significance alone. In a health context, it is common to talk about *"clinical significance"*, which measures whether a treatment has a genuine and noticeable effect on health outcomes. In choice modelling more broadly, analysts may wish to consider the *"behavioural significance"* of a parameter, i.e. whether it changes predictions (a point we return to in Section 4) and the *"policy significance"*, i.e. whether a finding has a significance impact on the outcome of any process using the results.

4 Model comparisons

A final topic of discussion relates to the comparison between different models, i.e. comparing goodness of fit for the purpose of model selection, where a variety of different tests exist. The majority of these tests incorporate penalties for additional parameters with the aim of deriving a parsimonious model.

The relationship between statistical testing of the model structure in terms of the impact of adding or removing parameters, and significance tests for individual parameters becomes clear in the context of standard errors. Changes to the value of parameters with larger t-ratios (i.e. smaller standard errors relative to their estimates) will lead to larger changes in model fit. The removal of a more *significant* parameter will thus lead to a bigger drop in fit. At the same time, it is well known that as the sample size N increases, the error decreases, with σ_β being inversely proportional to $\sqrt{N}$. As a result, it can be easily understood that bigger changes in log-likelihood are commonly observed when adding or removing parameters in models with larger samples, and this justifies why measures such as the Bayesian Information Criterion (BIC) impose stricter criteria for including additional parameters with larger N.

Let us take the case where an analyst compares two models, finding that the addition of a

single parameter leads to an improvement in log-likelihood by 4 units, i.e. a likelihood ratio test value of 8, which exceeds the χ^2_1 99% critical value. This would in many papers be accompanied by an endorsement that this improvement is highly significant. At the same time, analysts in many fields are interested in measures of prediction performance, i.e. how likely it is for the model to reproduce the observed outcomes. Using the above example of an improvement in fit by 4 units, let us assume that this is for a dataset with 3 alternatives, and an average probability of correct prediction of 60% in the base model. In a dataset with 1,000 observations, an improvement in log-likelihood by 4 units would mean that the average probability of correct prediction has risen to 60.2%, for the sake of simplicity assuming equal benefit across observations. With 5,000 observations, the improvement is even more negligible, at 60.05%. This calls for analysts to exercise restraint in describing improvements as highly significant on the basis of a likelihood improvement alone, and to consider also looking at goodness of fit measures where the penalty for additional parameters increases with the sample size (such as the BIC). In addition, there is merit in going further than model fit alone by also looking at the behavioural relevance of the additional parameters. This includes comparing different models in terms of how key metrics such as willingness-to-pay and elasticities are in line with generally accepted ranges for these measures. In this context, it is then also meaningful to contrast these measures across models developed independently not just in terms of their estimates, but also their statistical properties, in terms of confidence, for example.

5 Empirical example

As an illustration of the concepts discussed in this note, we now present a brief empirical example, using a stated choice dataset collected by Axhausen et al. (2008). A set of 388 people faced 9 choices each between two public transport routes, both using train. The alternatives are described by travel time (tt), travel cost (tc), headway (hw, time between successive trains) and the number of interchanges (ch). We estimate a simple Multinomial Logit (MNL) model, with the only additional complexity of an estimated income elasticity on the cost sensitivity. All models were estimated using Apollo (Hess and Palma, 2019). In addition to the standard estimation (and computation of classical and robust standard errors), we also conducted bootstrap estimation using 500 samples.

The results are shown in Table 1, where we note that all four marginal utility coefficients are negative, as expected, and that we obtain a negative income elasticity on the cost sensitivity. For all five parameters, the classical standard errors are substantially lower than the robust standard errors, which is in line with general findings. The bootstrap standard errors are very similar to the robust standard errors.

We first look at significance tests for the individual parameters. We see that for the four marginal utility coefficients, the $H_0 : \beta = 0$ hypothesis is rejected at high levels of confidence, easily exceeding the 99% level with either one-sided or two-sided tests. We report the p-value while most software would show these as zero. It is clear that, if we had only reported the p-values and the *star* measures, we would not be able to conduct any further analysis of the results, or to infer much in terms of difference in significance levels across parameters (not at all with the *star*

measures). For the income elasticity parameter, much lower t-ratios are obtained, albeit that we still reject $H_0 : \beta = 0$ at the 99% level using a one-sided test.

We next look at precision in terms of $\frac{\sigma}{\hat{\beta}}$ as well as confidence limits. It is here that the distinction with significance testing becomes clear. While all parameters would be judged to be significant at the 99% level, very substantial differences arise in $\frac{\sigma}{\hat{\beta}}$. Using a one-sided test, any parameter with a $\mid t_{0,\hat{\beta}} \mid > 2.33$ would reject $H_0 : \beta = 0$ at the 99% level, but $\frac{\sigma}{\hat{\beta}}$ could take any value below 0.43. Using the asymptotic confidence intervals with robust standard errors, and expressing the width relative to the estimated value, we can note that the typically used 95% confidence interval for λ_{inc} stretches 82.8% of the estimate to either side. For β_{ch}, which is similarly *"significant at the 99% level"*, the width of the 95% level to either side of the MLE is only 10.3%. This is a striking difference and highlights the earlier mentioned distinction between a *significant* and a *precise* estimate.

We finally return to the issue raised early on about asymptotic normality. The use of the formula $\hat{\beta} \pm z^C \sigma_\beta$ uses the more stringent property of normality, which may not be appropriate outside the immediate neighbourhood of the optimum. To investigate this issue, we contrast the asymptotic confidence intervals with those obtained using bootstrapping. Of course, the bootstrap and robust confidence intervals are wider than those obtained using the classical standard errors, in line with the difference in the standard errors. However, we also note that the bootstrap confidence intervals are not necessarily symmetric around the MLE. This is observed for all parameters for the 99% limits, where the difference is especially large for β_{tc}, where we also see asymmetry in the 95% confidence interval. For λ_{inc}, the confidence interval is right-skewed while for others, it is left-skewed. Overall, the bootstrap confidence intervals are also narrower, though these findings may of course be specific to the present application.

6 Conclusions

This note has looked at the issues of the interpretation and reporting of measures of statistical confidence in the context of choice model estimation. The misuse of the notion of "statistical significance" has been a long-standing concern for statisticians in particular and science more widely. The authoritative paper by Wasserstein et al. (2019) goes as far as concluding *"that it is time to stop using the term 'statistically significant' entirely. Nor should variants such as 'significantly different', 'p<0.05', and 'nonsignificant' survive, whether expressed in words, by asterisks in a table, or in some other way."* Less controversially, they say that *"[analysts should not] believe that an association or effect exists just because it was statistically significant [or] that an association or effect is absent just because it was not statistically significant."*

While we do not foresee a situation where choice modellers will abandon the notion completely, at least any time soon, we are of the opinion that a more nuanced approach is required. We believe that:

- Analysts should move away from the notion of 95% significance being a hard rule, recognising that this level is often easily obtained with large datasets, that lower thresholds may be acceptable for smaller datasets, and that a parameter can provide important behavioural

Table 1: Results on Swiss route choice data

					one-sided		two-sided			Confidence limits				$\frac{z^C \sigma_\beta}{\lvert\hat{\beta}\rvert}$			
	MLE	**errors**	σ	$t_{0,\hat{\beta}}$	p	*star*	p	*star*	$\frac{\sigma}{\hat{\beta}}$	**0.50%**	**2.50%**	**97.50%**	**99.50%**	**0.50%**	**2.50%**	**97.50%**	**99.50%**
β_{tt}	-0.0613	classical	0.0043	-14.37	$< 10^{-46}$	***	$< 10^{-46}$	***	0.0696	-0.0723	-0.0696	-0.0529	-0.0503	-17.95%	-13.64%	13.64%	17.95%
		robust	0.0066	-9.22	$< 10^{-19}$	***	$< 10^{-19}$	***	0.1084	-0.0784	-0.0743	-0.0482	-0.0441	-27.98%	-21.26%	21.26%	27.98%
		bootstrap	0.0067	-9.15	$< 10^{-19}$	***	$< 10^{-19}$	***	0.1078	-0.0792	-0.0748	-0.0494	-0.0472	-27.39%	-20.46%	20.47%	24.03%
β_{tc}	-0.1237	classical	0.0133	-9.27	$< 10^{-20}$	***	$< 10^{-19}$	***	0.1079	-0.1582	-0.1499	-0.0976	-0.0893	-27.83%	-21.14%	21.14%	27.83%
		robust	0.0222	-5.57	$< 10^{-7}$	***	$< 10^{-7}$	***	0.1796	-0.1810	-0.1673	-0.0802	-0.0664	-46.34%	-35.20%	35.20%	46.34%
		bootstrap	0.0225	-5.49	$< 10^{-7}$	***	$< 10^{-7}$	***	0.1775	-0.1885	-0.1747	-0.0893	-0.0807	-48.63%	-37.72%	29.59%	36.35%
β_{hw}	-0.0377	classical	0.0019	-20.29	$< 10^{-91}$	***	$< 10^{-90}$	***	0.0493	-0.0425	-0.0414	-0.0341	-0.0329	-12.72%	-9.66%	9.66%	12.72%
		robust	0.0023	-16.38	$< 10^{-59}$	***	$< 10^{-59}$	***	0.0610	-0.0437	-0.0422	-0.0332	-0.0318	-15.75%	-11.96%	11.96%	15.75%
		bootstrap	0.0023	-16.67	$< 10^{-61}$	***	$< 10^{-61}$	***	0.0596	-0.0438	-0.0425	-0.0335	-0.0323	-15.56%	-11.99%	11.59%	14.75%
β_{ch}	-1.1620	classical	0.0437	-26.58	$< 10^{-155}$	***	$< 10^{-154}$	***	0.0376	-1.2748	-1.2477	-1.0763	-1.0492	-9.71%	-7.37%	7.37%	9.71%
		robust	0.0611	-19.03	$< 10^{-80}$	***	$< 10^{-80}$	***	0.0525	-1.3195	-1.2817	-1.0424	-1.0045	-13.56%	-10.30%	10.30%	13.56%
		bootstrap	0.0617	-18.83	$< 10^{-78}$	***	$< 10^{-78}$	***	0.0528	-1.3528	-1.2935	-1.0485	-1.0222	-15.67%	-10.60%	10.34%	12.59%
λ_{inc}	-0.2572	classical	0.0584	-4.40	$< 10^{-5}$	***	$< 10^{-4}$	***	0.2270	-0.4079	-0.3717	-0.1428	-0.1066	-58.57%	-44.50%	44.50%	58.57%
		robust	0.1086	-2.37	0.0090	***	0.0179	**	0.4224	-0.5375	-0.4702	-0.0443	0.0231	-108.97%	-82.78%	82.78%	108.97%
		bootstrap	0.1093	-2.35	0.0093	***	0.0186	**	0.4226	-0.5140	-0.4718	-0.0490	0.0124	-98.73%	-82.41%	81.04%	104.79%

insights even if it has a lower level of significance. Failure to achieve 'significance' may simply indicate that not enough data has been collected.

- Analysts should look beyond statistical significance alone, and also consider behavioural and policy significance.
- Analysts should take care to report measures of parameter uncertainty in a numerically precise and well documented way. This means that if tests are presented, the analyst should clarify whether these are one-sided or two-sided, and excessive rounding should be avoided in the error measures.
- If analysts wish to highlight *significant* parameters using *star* measures, this should only be done if standard errors or t-ratios are presented alongside them. With p-values, analysts should similarly think about the degree of precision that is possible, and again present these measures not in isolation but alongside standard errors or t-ratios.
- A distinction should be made between parameter significance and precision, noting that two parameters that both pass a significance test above the 99% level can have vastly different confidence intervals. In practical work, precision may matter much more than statistical significance, and analysts may want to ensure that an effect is behaviourally meaningful and important across the entire width of say a 95% confidence interval.
- The property of asymptotic normality needs to be better understood, and analysts should recognise that computing confidence intervals in particular may be very inaccurate by assuming that normality holds several standard errors away from the MLE.

Finally and as an alternative to frequentist significance testing, analysts could consider the Bayesian approach to parameter uncertainty, where prior knowledge is updated in response to evidence provided by data, and statistical inference is based upon decision-making theory. Through the concept of posterior probabilities, Bayesian econometrics offers a probabilistic representation of not only parameters but also hypotheses. For instance, Bayes estimates are the whole posterior distribution of the parameters from which analysts can derive the exact probability of (credible) intervals containing the true parameter. Bayes' rule can also be used to derive measures of relative evidence across competing models, including posterior probabilities of hypotheses. For example, and unlike p-values, Bayes factors provide a data-supported measure of the odds in favour of the null hypothesis over the alternative. Even though Bayesian hypothesis testing is more intuitive and departs from the frequentist concepts, there is an open discussion among Bayesian practitioners about the flaws in null hypothesis significance testing, which goes beyond the scope of this note.

References

Amrhein, V., Greenland, S., McShane, B., 2018. Redefine statistical significance. Nature Human Behaviour 2, 6–10. doi:10.1038/s41562-017-0189-z.

Armstrong, P., Garrido, R.A., Ortúzar, J. de D., 2001. Confidence interval to bound the value of time. Transportation Research Part E 37, 143–161.

Axhausen, K.W., Hess, S., König, A., Abay, G., Bates, J., Bierlaire, M., 2008. State of the art estimates of the swiss value of travel time savings. Transport Policy 15, 173–185.

Benjamin, D., Berger, J., Johannesson, M., Nosek, B., Wagenmakers, E.J., Berk, R., Bollen, K., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C., Clyde, M., Cook, T., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., Johnson, V., 2017. Redefine statistical significance. Nature Human Behaviour 2. doi:10.1038/s41562-017-0189-z.

Daly, A., Hess, S., de Jong, G., 2012. Calculating errors for measures derived from choice modelling estimates. Transportation Research Part B 46, 333–341.

Hess, S., Palma, D., 2019. Apollo: A flexible, powerful and customisable freeware package for choice model estimation and application. Journal of Choice Modelling 32, 100170. URL: http://www.sciencedirect.com/science/article/pii/S1755534519300703, doi:https://doi.org/10.1016/j.jocm.2019.100170.

Wasserstein, R.L., Lazar, N.A., 2016. The asa statement on p-values: Context, process, and purpose. The American Statistician 70, 129–133. URL: https://doi.org/10.1080/00031305.2016.1154108, doi:10.1080/00031305.2016.1154108, arXiv:https://doi.org/10.1080/00031305.2016.1154108.

Wasserstein, R.L., Schirm, A.L., Lazar, N.A., 2019. Moving to a world beyond "p<0.05". The American Statistician 73, 1–19. URL: https://doi.org/10.1080/00031305.2019.1583913, doi:10.1080/00031305.2019.1583913, arXiv:https://doi.org/10.1080/00031305.2019.1583913.