New Significance Thresholds as a Function of Sample Size and the Prior Null Probability

Tom Engsted *

Posted Date: 22 November 2024

doi: 10.20944/preprints202411.1731.v1

Keywords: Hypothesis testing; prior null probability; Bayesian Information Criterion; significance threshold; sample size.

 doi:10.20944/preprints202411.1731.v1

Article

New Significance Thresholds as a Function of Sample Size and the Prior Null Probability

Tom Engsted

Department of Economics and Business Economics, Aarhus University, Fuglesangs alle 4, DK-8210 Aarhus V.; tengsted@econ.au.dk

Abstract: In economics, declaring an empirical result usually involves rejection of a point null hypothesis in favor of a composite alternative, and most researchers seem to (implicitly) attach a low prior probability to the null. However, this view is fundamentally at odds with the underlying statistical requirement that point nulls have a high prior probability assessment. Acknowledging this requirement implies that significance at the 5% level is a very low hurdle rate. I suggest to use the Bayesian Information Criterion (BIC) to set new significance thresholds as a function of sample size.

Keywords: hypothesis testing; prior null probability; Bayesian Information criterion; significance threshold; sample size

1. Introduction

It seems to be a common view among economists that point null hypotheses should be attached a low prior probability weight. Economists typically develop models of relationships between variables, and an important ingredient in the empirical evaluation of the models is the rejection of a point null hypothesis of *no* relationship (e.g., rejection of H_0: $b = 0$ in the regression $Y = a + bX + u$) in favor of a composite alternative (H_1: $b \neq 0$). Naturally, the economist who has developed the model has a strong (prior) belief in the model, so the economist explicitly – or, usually, implicitly – operates with a low prior probability assessment of the null. As Abadie (2020, p. 193) states: "in empirical contexts that are common in economics ... there are rarely reasons to put substantial prior probability on a point null".[1]

However, although Abadie (2020) is probably right when stating that economists tend to put low prior probability weight on their null hypotheses, this state of affairs is fundamentally at odds with the underlying statistical rationale behind testing a point null against a composite alternative in classical hypothesis testing. In such a testing setup, the null hypothesis constitutes the "established theory" (Berger and Sellke, 1987, p. 115) that requires strong evidence against it in the data to be rejected, corresponding to the choice of a low significance level (like 5%) such that the probability of rejecting a true H_0 (i.e., the Type I error probability) is low. Thus, almost by definition – from a statistical point of view – a point null should be attached a *high* prior probability weight.

If the model stated in the null hypothesis cannot reasonably be attached a relatively high prior probability, one should not test the model using a classical 'point null against composite alternative' testing setup. On the other hand, if the model *is* given a high prior probability weight, significance at the conventional 5% level is a very low hurdle rate for declaring an empirical finding. In any case, the interpretation usually given by empirical researchers to the traditional hypothesis testing setup – and to the results of applying this setup – needs to be adjusted.

In a widely cited proposal, Benjamin et al. (2018) suggest to redefine statistical significance such that 0.5% becomes the new default threshold instead of the conventionel 5%. One problem with this proposal is that it only applies to experimental settings where the sample size can be adjusted to

[1] Abadie (2020) argues that when the null has a low prior probability, a non-rejection is more informative than a rejection. Intuitively, if we put a low prior probability weight on H_0, we expect H_0 to be rejected, so a non-rejection is more surprising and – hence – more informative than a rejection.

 doi:10.20944/preprints202411.1731.v1

maintain a high level of statistical power. For non-experimental data where the sample size is fixed – the typical case in economics – the proposal is less relevant. In a (very) small sample with low test power, the 0.5% threshold may not be suitable. Similarly, it is well-known that for a fixed and arbitrarily low significance level, a very small and completely unimportant deviation from the null value becomes statistically significant in a sufficiently large sample. Thus, ideally the significance threshold α should depend on the sample size n. A simple and widely used model selection criterion can be invoked to make such a connection between α and n: the Bayesian Information Criterion (BIC) of Schwarz (1978). This criterion is appealing if the researcher has some – but not very precise – prior knowledge or idea of the range of variation for the parameter in question; a situation often encountered in economic modeling.

In the rest of this paper I elaborate on these points. First, I argue that classical testing of point null hypotheses should not be done unless one puts a high prior probability weight on the null. Second, I show how to use the BIC to relate the significance threshold to sample size as a function of the prior null probability. This threshold correctly corresponds to the way significance levels and p-values are often informally (but incorrectly) interpreted, namely as a probability assessment of the null, given the data.

2. Materials and Methods

2.1. Testing a Point Null Hypothesis

The testing setup applied in the vast majority of empirical studies in economics and other social sciences is the classical formulation of a point (or simple) null hypothesis for a population parameter b, against a composite two-sided alternative where the parameter can take an infinite number of values different from the point null value, i.e., H_0: $b = b_0$; H_1: $b \neq b_0$. Often $b_0 = 0$, such that the researcher is testing the hypothesis of 'no effect' or 'no relationship'. For two reasons, this formulation of the test implies that H_0 from the outset should be attached a relatively high probability weight:

First, singling out the point value b_0 in H_0 compared to the continuum of values in H_1 naturally implies that this point value is considered particularly likely. If not, why single it out? Second, the conventional choice of a low significance level (like 5%) in the test reflects the (implicit) strong prior belief in H_0. As Lehmann and Romano (2008, p. 58) state in their classic text on testing statistical hypotheses: "If one firmly believes the hypothesis to be true, extremely convincing evidence will be required before one is willing to give up this belief, and the significance level will accordingly be set very low".[2]

In textbook descriptions of hypothesis testing (especially the Neyman-Pearson variant), the choice of a low significance level is often justified by reference to Type I error (i.e, rejecting a true H_0) having more serious consequences than Type II error (i.e., not rejecting a false H_0). Imbens (2021) emphasizes the importance of Type I and II error for concrete decision making problems. However, in economics and the other social sciences, meaningfully attaching costs to the two error types is in most cases impossible and is hardly ever done in practice (economic models have many different uses, so how should we meaningfully and uniquely assess the costs of falsely rejecting or not rejecting, say, the efficient markets hypothesis?). Instead, the choice of a low significance level implicitly reflects the view that H_0 is highly likely and, hence, should only be rejected if there is strong evidence against it in the data.

However, most researchers do not think carefully about the choice of significance level. The conventional level of 5% (on rare occasions, 1% or 10%) is the standard (implicit) choice in almost all empirical studies, as shown by, among others, Harvey (2017) and Andrews and Kasy (2019). At the same time, the 'point null against composite alternative' testing setup is applied in the vast majority

[2] This quote from Lehman and Romano (2008) also shows that classical (frequentist) hypothesis testing is not free from priors. There is an (implicit) prior view of the null built into the choice of significance level.

 doi:10.20944/preprints202411.1731.v1

of studies. This is puzzling because, as noted above, most researchers hold the view that the null should *not* have a high prior probability weight. Thus, it seems that the statistical testing setup in most applications does not conform with the underlying view of the null and alternative hypotheses that researchers typically have.

It is sometimes argued that in economics – and the social sciences in general – point null hypotheses are rarely exactly true and, hence, for that reason point nulls should be attached a low prior probability weight. However, H_0: $b = 0$ should not be interpreted as an exact zero effect, but rather as a 'negligible' effect. Berger and Delampady (1987) show that, unless the sample size is very large, a point null will often be a good approximation to a small interval null. Thus, testing a point null is not necessarily meaningless if an *exact* zero effect is highly unlikely, but it requires that a non-zero effect in a small interval around zero is considered highly likely. If not, a classical test of a point null is not suitable for the analysis.[3]

3. Results

3.1. Setting Significance Thresholds Using BIC

Given that a classical test of a point null hypothesis is deemed suitable for the analysis, which – as I have argued above – requires a high prior probability assessment of the null, the question arises as to what significance or p-value threshold should be used. In his discussion of statistical hypothesis tests, Imbens (2021) argues that in some disciplines the importance of such tests have been overemphasized, but he also argues that in some situations there is a limited but important role for such tests, and for the question of when to abandon the null in favor of the alternative, he states: "I do think that small p-values are *necessary* for such a conclusion. More specifically, in cases where researchers test null hypotheses on which we place substantial prior probability, it is difficult to see how one could induce anyone to abandon that belief without having a very small p-value. Reporting such a p-value would seem a reasonable way to summarize evidence." (Imbens, 2021, p. 158).

Imbens' statement leaves the question of what constitutes a "very small p-value". How small should it be for the null to be rejected? For cases where the prior null probability is well above 0.50, Benjamin et al. (2018) – where "et al." includes Imbens – propose to set a new universal significance threshold at 0.005 (i.e., 0.5%), which is substantially stricter than the conventional threshold of 0.05 (5%). The proposal is based on a desire to control the false discovery rate (the proportion of true nulls among all rejected nulls) at 5%,[4] and with a view on what threshold is needed to be in accordance with conventionel Bayes factor classifications of 'substantial to strong evidence against the null'. In addition, the proposal is explicitly intended for experimental studies where the sample size can be increased so as to maintain a reasonably high level of statistical power.

The proposal of Benjamin et al. (2018) has received a lot of attention and it may lead to a significant change in empirical practice across various research fields. However, a drawback of the proposal is that it does not apply to non-experimental settings where the sample size of the observational data is fixed and power may be low, which is a situation often encountered in economics. In such cases it becomes important to take the sample size into account when setting the significance threshold. A general feature of classical hypothesis tests at conventional significance levels is that in small samples (with low test power) economically sizeable effects may be statistically insignificant, while in large samples (with high power) economically unimportant effects become statistically significant. For many years, econometricians (e.g., Leamer, 1978) have advocated letting the significance level α (or cut-off level for the p-value) be a decreasing function of the sample size n, although classical testing theory contains no formal rules for how to optimally relate α to n. However, if one is willing to make use of

---

[3] Similarly, in Bayesian hypothesis testing of a point null – interpreted as a small interval null – the prior probability of the null, $P(H_0)$, is typically set at 0.50 or higher (Berger and Sellke, 1987).

[4] For a detailed discussion of the false discovery rate in empirical economics, see Engsted (2024).

 doi:10.20944/preprints202411.1731.v1

insights from Bayesian hypothesis testing, a simple and appealing solution is readily available, which will now be described.[5]

To motivate the proposed solution, note that researchers often misinterpret classical test statistics and p-values as giving probabilistic evidence for or against the null (and alternative) hypothesis, given the data, cf. Wasserstein and Lazar (2016): The lower (higher) the p-value, the lower (higher) the probability that the null is true. This interpretation is incorrect because the p-value gives the probability of obtaining the observed data or more 'extreme' data (denoted $D+$), given that the null is true, i.e., p-value = $P(D+ \mid H_0)$. It is no wonder that people tend to make this mistake because having the probabilistic odds for H_0 versus H_1, given the observed data (denoted D), is intuitively more understandable and meaningful than the probability of the observed data or more extreme data (that were not observed!), $D+$, given that H_0 is true.

In Bayesian hypothesis testing and model comparison, the aim is to obtain $P(H_0 \mid D)$ and $P(H_1 \mid D)$, the two objects that researchers are really looking for. Fortunately, a simple technique exists for transforming the classical significance level and p-value into these conditional null and alternative probabilities. A further appealing feature of the approach is that it directly allows (in fact, requires) an explicit specification of the prior null probability, $P(H_0)$, that we saw in Section 2 is an important (implicit) ingredient in classical hypothesis testing.

In Bayesian analysis, the posterior odds ratio for the two hypotheses is given as $\frac{P(H_0|D)}{P(H_1|D)} = BF \cdot \frac{P(H_0)}{P(H_1)}$, where the 'Bayes factor' $BF = \frac{P(D|H_o)}{P(D|H_1)}$ is the ratio of data likelihoods under H_0 and H_1, respectively, and $\frac{P(H_0)}{P(H_1)}$ is the prior odds ratio. Probabilities add up to one, so from this expression it follows that

$$P(H_0 \mid D) = \frac{P(H_0) \cdot BF}{1 + [P(H_0) \cdot (BF - 1)]}. \tag{1}$$

Now, for the usual setup of a point null against a two-sided alternative – and if we use the Bayesian Information Criterion (BIC) of Schwarz (1978) to compute the Bayes factor – we obtain (cf. Raftery, 1995; Kass and Raftery, 1995): $BF = \exp(\frac{1}{2}\text{BIC})$, where $\text{BIC} \approx \log(n) - t^2$. Here, n is the sample size, log is the natural logarithm, and t is the ususal t-statistic for the model parameter in question.[6] The underlying assumption for using BIC to compute BF is that under H_1 the uncertainty associated with the population model parameter can be approximated by a normal distribution centered around the maximum likelihood estimate and with a relatively large variance, i.e., $b \sim N(\widehat{b}, \sigma^2)$, where σ^2 is equal to $n\sigma^2_{\widehat{b}}$. This assumption reflects quite well how researchers often have some – vague – prior knowledge or idea of the range of variation for the parameter.

Next, we need to specify the prior null probability $P(H_0)$. If we translate the requirement in classical testing that a point null should have a high prior probability assessment (cf. the discussion in Section 2) into $P(H_0) = 0.95$, corresponding to the conventional 5% significance level, then a natural procedure would be to reject H_0 if the posterior probability $P(H_0 \mid D)$ is less than 0.50. In other words, we start out having strong faith in the null hypothesis (95%), but if the information in the data leads to a downward revision of this probability such that the null becomes less likely than the alternative, then we reject the null in favor of the alternative.

The above considerations thus lead to $P(H_0) = 0.95$, and inserting $BF = \exp\left(\frac{1}{2}(\log(n) - t^2)\right)$ into Equation (1) then gives, after solving for t,

$$t = \sqrt{\log(n) - 2\log\left(\frac{0.05}{0.95}\right)}, \tag{2}$$

---

[5] The idea of using insights from Bayesian hypothesis testing to relate α to n in classical hypothesis testing was introduced in the sociological literature by Raftery (1995).

[6] See Raftery (1995) and Kass and Raftery (1995) for the case where the test statistic involves more than one parameter, e.g., a joint test on multiple parameters in a regression model.

 doi:10.20944/preprints202411.1731.v1

as the t-statistic threshold. This is the numerical threshold value that a t-statistic needs to obtain in order for $P(H_0 \mid D) < 0.50$ when we start out with $P(H_0) = 0.95$.[7] Table 1 shows the threshold values for the t-statistic and associated p-value, as a function of the sample size. As seen, the threshold becomes stricter as the sample size increases. With $n = 100{,}000$ or higher, a t-stat of at least 4.17 (p-value < 0.00003) is needed to declare a finding 'statistically significant'. But also for more moderate sample sizes, substantially stricter thresholds than the usual $t = 1.96$ (p-value < 0.05) are needed.

A caveat is in order: BIC provides only an approximation to the Bayes factor. The approximation is quite accurate for moderate to large samples (Raftery, 1995; Kass and Raftery, 1995), but in (very) small samples (like $n = 25$ in Table 1) the approximation may be too crude. Thus, the suggested procedure in this paper should be restricted to cases where the samples are at least moderately sized.

Table 1. Significance thresholds based on BIC

	Threshold	
	t-statistic	p-value
$n = 25$	3.02	0.0025
50	3.13	0.0017
100	3.24	0.0012
500	3.48	0.0005
100,000	4.17	0.00003
1,000,000	4.44	0.00001

4. Discussion

I have argued that the traditional approach in empirical economics of testing point null hypotheses with a 5% significance level requires that a high prior probability is attached to the null hypothesis; a requirement that is often not acknowledged by researchers. Given this requirement, significance at the 5% level is a much too low hurdle to pass in order to claim an empirical finding. By using the Bayesian Information Criterion (BIC) and insights from Bayesian hypothesis testing, intuitively meaningful new significance thresholds as a function of sample size are derived. These threshold values are much stricter than the conventional $t = 1.96$ (p-value < 0.05) threshold. BIC is a widely applied model selection tool in empirical economics, so also applying it to provide significance thresholds as a function of sample size should appeal to most researchers.

Elsewhere (Engsted and Schneider, 2024) I have argued that the dichotomous 'reject/don't reject' decision inherent in both classical and Bayesian hypothesis testing is fundamentally unsuited for the kind of models and data that we typically work with in empirical economics. However, *if* such a dichotomous decision is deemed relevant for a particular research project involving non-experimental data, I believe the procedure suggested in the present paper is to be preferred over procedures based on a fixed significance level, be it the conventional $\alpha = 0.05$ or a lower level.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Abadie, A. (2020): Statistical nonsignificance in empirical economics. *American Economic Review: Insights* 2, 193-208.
2. Andrews, I., and M. Kasy (2019): Identification of and correction for publication bias. *American Economic Review* 109, 2766–2794.
3. Benjamin, D.J., J.O. Berger + 70 coauthors (2018): Redefine statistical significance. *Nature Human Behaviour* 2, 6–10.

[7] Alternatively, we may start out with the neutral prior odds $P(H_0) = P(H_1) = 0.50$ and then compute the t-statistic threshold associated with $P(H_0 \mid D) < 0.05$. It turns out to give the exact same value as in Equation (2).

 doi:10.20944/preprints202411.1731.v1

4. Berger, J.O., and M. Delampady (1987): Testing precise hypotheses. *Statistical Science* 2, 317-352.
5. Berger, J.O., and T. Sellke (1987): Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association* 82, 112-139.
6. Engsted, T. (2024): What is the false discovery rate in empirical research? *Econ Journal Watch* 21, 92-112.
7. Engsted, T., and J.W. Schneider (2024): Non-experimental data, hypothesis testing, and the Likelihood Principle: A social science perspective. *Foundations and Trends in Econometrics* 13, 1-66.
8. Harvey, C.R. (2017): Presidential Address: The scientific outlook in financial economics. *Journal of Finance* 72, 1399-1440.
9. Imbens, G.W. (2021): Statistical significance, *p*-values, and the reporting of uncertainty. *Journal of Economic Perspectives* 35, 157-174.
10. Kass, R.E., and A.E. Raftery (1995): Bayes factors. *Journal of the American Statistical Association* 90, 773-795.
11. Leamer, E.E. (1978): *Specification Searches: Ad Hoc Inference with Non Experimental Data*. John Wiley & Sons.
12. Lehmann, E.L., and J.P. Romano (2008): *Testing Statistical Hypotheses* (Third edition). Springer.
13. Raftery, A.E. (1995): Bayesian model selection in social research. *Sociological Methodology* 25, 111-163.
14. Schwarz, G.E. (1978): Estimating the dimension of a model. *Annals of Statistics* 6, 461–464.
15. Wasserstein, R.L., and N.A. Lazar (2016): The ASA's statement on p-values: Context, process, and purpose. *The American Statistician* 70, 129-133.