Krämer, Walter

Working Paper

The cult of statistical significance

CESifo Working Paper, No. 3246

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

The Cult of Statistical Significance

Walter Krämer

CESIFO WORKING PAPER NO. 3246
CATEGORY 5: ECONOMICS OF EDUCATION
NOVEMBER 2010

CESifo *Working Paper No. 3246*

The Cult of Statistical Significance

Abstract

This article takes issue with a recent book by Ziliak and McCloskey (2008) of the same title. Ziliak and McCloskey argue that statistical significance testing is a barrier rather than a booster for empirical research in many fields and should therefore be abandoned altogether. The present article argues that this is good advice in some research areas but not in others. Taking all issues which have appeared so far of the German Economic Review and a recent epidemiological meta-analysis as examples, it shows that there has indeed been a lot of misleading work in the context of significance testing, and that at the same time many promising avenues for fruitfully employing statistical significance tests, disregarded by Ziliak and McCloskey, have not been used.

JEL-Code: A00.

Keywords: significance testing.

Walter Krämer
Faculty of Statistics
Technical University Dortmund
44221 Dortmund
Germany
walterk@statistik.tu-dortmund.de

Version November 2010
Research supported by DFG under SFB 823. On purpose, the title is the same as that of a recent book by Ziliak and McClowskey (2008). I am grateful to Michael Bücker and Ronja Walter for excellent research assistance, and to Stephen Ziliak for comments on a book review in Statistical Papers on which this article is based.

1. Introduction and summary

A significant statistical test only means: if the null hypothesis were true – a big if – then the tail probability of the observed event would be less than a pre- chosen level of significance. This rather modest claim is even further compromised by its extreme dependence on the size and on the generation of the sample and by the common practice of disguised multiple testing. i.e. doing lots of tests and reporting only the most “significant” results, and the ensuing understatement of the true probability of an error of the first kind (“data mining”). This is what every decent statistician knows, or at least should have been taught in any introductory mathematical statistics course.

Additional, rather popular empirical improprieties are HARKing (“Hypothesizing after the Results are Known), collective as opposed to individual data mining, and what I call an Error of the Third Kind, by which I mean mistaking a rejected null as proof that the alternative is true. All of these deficiencies figure prominently in a critical literature of long standing that is summarized in section 2 below. But the critique which is the subject of the present paper is much more fundamental. It dates back at least to Tyler (1931), Sterling (1959) and Rozeboom (1960) and concludes that even it significance testing were properly done according to the rules of the game, it would still be fundamentally flawed as an approach to empirical research in many fields due to implied disregard of what really counts in many applications, the size, as opposed to the mere existence, of an effect. Ziliak and McCloskey (2008) have created the neologism “oomph” for this; they argue “that ’oomph’, the difference a treatment makes, dominates precision” (p. xvii), and that a rather disproportionate amount of attention is devoted to the latter, taking away scarce resources from more promising avenues of research.

There are therefore three types of misleading test-based inference: (a) there is no effect, but due to technical deficiencies, “significance” nevertheless obtains, (b) there is a large effect (much “oomph”), but due to variability, it is not “significant” and therefore discarded (c) there is only a small effect (no “oomph”) , but due to precision it is highly "significant" and therefore taken seriously. Taken together, these sources of error have led McCloskey (2002, p. 44) to conclude: "The progress of economic science has been seriously damaged [by the common practice of significance testing]. You can't believe anything that comes out of [it]. Not a word. It is all nonsense, which future generations of economists are going to have to do all over again. Most of what appears in the best journals of economics is unscientific rubbish. I find this unspeakably sad. All my friends, my dear, dear friends in economics, have been wasting their time....They are vigorous, difficult, demanding activities, like hard chess problems. But they are worthless as science." Or even more bluntly, in her book with Ziliak (2008): "If null-hypothesis significance testing is as idiotic as we and other critics have so long believed, how on earth has it survived?" (p. 240)).

One can hardly imagine a more devastating critique of this aspect of empirical work in economics (or in any other field).

The purpose of the present article is to put the Ziliak-McCloskey view into perspective, by supporting it for some types of tests but not for others. The first class of tests, where much is going wrong indeed, is sometimes referred to as "presumptive" or "confirmatory" testing (see e.g. Tang et al. 1993). It is from here that Ziliak and McCloskey (2008) and other critics draw most of their examples. Confirmatory testing means that there is a particular alternative one has in mind, with the aim or wish of establishing this as true. Section 2 summarizes various illegal ways in which this goal is often achieved in applications, plus related aberrations when interpreting confirmatory tests. Section 3 exemplifies such type (a) mistakes using a recent example from epidemiology, while Section 4 considers type (b) and (c) mistakes. This is done by checking all empirical articles ever from the German Economic Review, the Journal of the Verein für Socialpolitik, which is distributed to about 4000 member four times a year. Both sections confirm Ziliak and McCloskey (2008) insofar as lots of useless and misleading inferences are unearthed. But they also show that confirmatory testing only makes sense, no matter whether one is after mere significance or "oomph", if the underlying model is reasonably correct, and that it is the common failure to test for this which is the real threat to meaningful statistical results.

Such tests, often called “exploratory” or “specification” tests (see Krämer and Sonnberger 1986), are the topic of the final section 5. Specification tests are not aimed at any specific alternative, so a rejection of the null only tells the investigator that he or she should look out for a better model, without establishing whichever type of “effect” there is supposed to exist. They are also more in line with the Popperian paradigm of scientific progress, where the null hypothesis corresponds to established beliefs, to be abandoned only in the presence of compelling evidence. In a sense, therefore, the Ziliak-McCloskey argument is turned on its head: in order to extract meaningful information (“oomph”) from the economy or whatever field of application via formal statistical models, one has to do a lot of significance testing first.

2. Confirmatory testing and errors of the third kind

A significance level of $\alpha = 5\%$ for a statistical test implies that, even when the null hypothesis were true, the procedure would still reject it in roughly 5 out of 100 experiments. In the context of a specific alternative, usually some kind of “effect”, this means that even without any effect being present, the test will nevertheless claim one in roughly 5 out of 100 trials. This is the well known error of the first kind.

A first objection to the routine use of statistical significance testing concerns the ease with which a significant test often leads to what I have termed above an error of the third kind: to assume that a significant test implies that the alternative is true: “The sin comes in believing a causal hypothesis is true because your study came up with a positive result” (Sander Greenland from UCLA, as quoted in Taubes, 1995, p. 169).

This error of the third kind, or some variant such as “the null hypothesis is wrong with 95 % probability” occurs even among professional statisticians. Haller and Krauss (2002) have asked 30 statistics instructors, 44 statistics students and 39 practicing researchers from six psychology departments in Germany about the meaning of a significant two-sample t-test (significance level = 1%). The test was supposed to detect a possible treatment effect based on a control group and a treatment group. The subjects were asked to comment upon the following six statements (all of which are false). They were told in advance that several or perhaps none of the statements were correct.

> 1) You have absolutely disproved the null hypothesis (that is, that there is no difference between the population means). ❍ true / false ❍
>
> 2) You have found the probability of the null hypothesis being true. ❍ true / false ❍
>
> 3) You have absolutely proved your experimental hypothesis (that there is a difference between the population means). ❍ true / false ❍
>
> 4) You can deduce the probability of the experimental hypothesis being true. ❍ true / false ❍
>
> 5) You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision. ❍ true / false ❍
>
> 6) You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions. ❍ true / false ❍

All of the statistics students, 90% of the practicing psychologists and 80% of the methodology instructors marked at least one of the above faulty statements as correct. And what is more, even lots of statistics textbooks do. Examples from the German market include Wyss (1991, p. 547) or Schuchard-Fischer et al. (1982), who on p. 83 of their best-selling textbook explicitly advise their readers that a rejection of the null at 5% implies a probability of 95% that the alternative is correct. For details, see Krämer and Gigerenzer (2005) or Krämer (2008, chapter 8).

A second, even more popular mistake is to claim some nominal significance level α when in reality the reported test statistic is the most significant one among n trials, each conducted at the level α . The true significance level is then simply the probability that the maximum of n test statistics is larger than some critical value and increases rapidly with n. Krämer and Runde (1992) have used this device to establish what they call the "Krämer-Runde-seven-modulo 1 effect." This means in words, that on days of the month Nr. 1, 8, 15, 22, and 29 the German stock price index DAX performs significantly better than average (t=3.161). Or in technical terms, the null hypothesis that stocks perform the same on these days as on others could be rejected, given the available data, at a level of 5%. What Krämer and Runde also did were tests of many other hypotheses: There is no six-modulo-2-effect, there is no six-modulo-3-effect, there is no seven-modulo- 2-effect, eight-modulo-3-effect, and so on, ad nauseam. Given a particular data set and one hundred such hypotheses, all of them true, one is still bound to find about 5 "significant" effects, i.e. rejections of the null. And it is well known (see e.g. McCloskey 1985 or Ziliak and McCloskey 2008) that many other authors procede along similar lines, without reporting the unsuccessful trials. See also Krämer (2010, chapter 15).

This multiple testing problem has of course long been recognized in statistical research, see Krämer and Sonnberger (1986, chapter 6) for an overview of the early literature in econometrics and Tang et al. (1993) or Altmann et al. (2001) for some advice on how to cope with this problem in biometric applications. But is seems that the enormous theoretical work that has been done here has not yet made its way into routine empirical applications. And even if it had, it seems that the many restrictions that are attached to many multiple testing approaches would severely limit their impact on the problem we are discussing here.

In economics, this habit of reporting only the most “significant” results is sometimes referred to as "data mining" (Lovell, 1983)[2]. It is of course strictly illegal and rightly frowned upon, but has nevertheless been common practice in empirical economics ever since statistical test of significance have been introduced.

Not illegal, but equally misleading, is the related phenomenon known as “publication bias”: “There is some evidence that in fields where statistical tests of significance are commonly used, research which yields nonsignificant results is not published” (Sterling 1959, p. 30). “Such research being unknown to other investigators may be repeated independently until eventually by chance a significant result occurs.” Taken to the limit, this argument implies that a “significant“ effect will be found eventually almost surely, no matter what.

---

[2] Not to be confused with the serious business of the same name that is a modern subject of computer science

Denton (1985) calls this "collective data mining"; it happens even when no individual investigator engages in this practice and has long been to grossly distort the type 1 error probabilities wherever formal testing for significance is done. In psychology, this bias towards the alternative is known as the file drawer problem: negative results remain stuck in the file drawer. In medicine, Stern and Simes (1997) report that among 748 studies approved by the Royal Prince Alfred Hospital Ethics committee between 1979 and 1988, about 85% were eventually published if they reported significant results at levels 5% or less. Among studies which did not report significant results, the percentage of accepted papers was only 50 %. See also Beck-Bernholdt and Dubben (2004).

3. A particular application in detail: Childhood leukemia in the vicinity of nuclear power plants

Following Krämer and Arminger (2010), this section exemplifies the type (a) deficiencies explained in section 2 via a recent study of childhood leukemia in the vicinity of nuclear power plants - a meta analysis combining various previous investigations and some data collected independently - prepared by Greiser (2009) for the political party "Bündnis 90 / Grüne". Like many others, it purports to show that nuclear power plants induce a "statistically significantly elevated risk of leukemia for all age groups considered" (p. 3)[3] and starts with an error of the third kind: mistaking a rejected null hypothesis as proof that the alternative is true. "AKW erhöhen das Leukämierisiko (nuclear power plants increase risk of leukemia)" was the heading of a press release distributed by Bündnis 90 /Grüne in the fall of 2009, which strongly contributed to the fiercely held belief by many Germans that nuclear power is bad for you.

The present section exemplifies the arguments from section 2 by showing that not even statistical significance obtains. A first type (a) mistake is HARKing: "Hypothesizing after the results are known". As Kruskal (1969) puts it, "Almost any set of data […] will show anomalies of some kind when examined carefully, even if the underlying probabilistic structure is wholly random – that is, even if the observations stem from random variables that are independent and identically distributed. By looking carefully enough at random data, one can generally find some anomaly […] that gives statistical significance at customary levels although no real effect is present" (p. 247). A famous example is one of the very first applications of significance testing at all, the observation made by astronomers that the orbital planes of the planets are quite close together. In 1734, Daniel Bernoulli and his son John computed the probability that this is due to chance (given that orbital planes are determined randomly; in modern language, they computed the prob-value of a test; see Todhunter1949, sections 394-397). This probability however is only correct if the particular anomaly had been established beforehand, and is larger otherwise.

The same is true in the context of the leukemia vs nuclear power debate. In Germany, for instance, testing on a massive scale started only after an abnormal cluster of leukemia cases was found close to the Krümmel power generation plant. But there are dozens of additional illnesses which might be examined for clusters: Sudden death syndrome, other types of cancer, birth defects of all sorts, Alzheimer, Parkinson, high blood pressure and so on. According to Kruskal, the chances that some such malady can be "significantly" associated with nuclear power plants are almost one.

Then there is the publication bias, which is bound to particularly affect any meta-analyses which collect together previous work. At the time of this writing, there are well above 1000 nuclear installations worldwide available for testing. But the meta-analysis by Greiser is based on only 80 of these, located in the UK, France, Germany, the U.S. and Canada. What about the others? How many studies which did not find an excess of childhood leukemia have never made it into print? For instance, no excess incidence has so far been reported for nuclear sites in Japan, Taiwan, Sweden, Spain and Switzerland. Given the enormous media interest in occurrences of this kind, one can certainly be sure that any leukemia cluster close to

[3] English translation. The German original says: "Die Ergebnisse zeigen ein statistisch signifikant erhöhtes Erkrankungsrisiko an Leukämie für alle untersuchten Altersgruppen."

a nuclear facility in these countries would have made headlines there as well. Therefore, the absence of such headlines provides evidence that no such clusters have occurred, or that studies reporting this absence have not made it into print.

Another important degree of freedom is the time period under consideration. The literature abounds with examples where excess mortality or morbidity was found in certain periods, but not in others. For instance, the studies from Canada quoted by Greiser (2009), reporting excess incidence of childhood leukemia around Canadian nuclear power plants, cover only years up to 1986. It is rather safe to assume (and confirmed by private information from Canadian authorities) that no excess incidence was observed thereafter.

Then one has to choose a distance from the potential source of radiation. Conventional choices are 6.5 km, 15 km, 20 km, 25 km, 50 km or complete counties, like in most studies from Canada and the U.S.. Again, there is an abundance of examples where excess incidence or mortality was observed for some distances, but not for others.

Then there is the type of cancer (myeloid leukaemia – ML, acute lymphoblastic leukaemia – ALL, acute non-lymphoblastic leukemia, Non-Hodgkin lymphoma, other cancers), which likewise might lead to an excess for one type and a deficit for another. The age group of the children is also important, as many studies report an excess of leukemia for some age groups, and a deficit for others. For details, see Krämer and Arminger (2010).

It is obvious that by judiciously adjusting these parameters it is trivial to establish "significant" effects of any sort. Good examples are Hoffmann et al. (1995) or Körblein and Hoffmann (1999, p. 18), who, being dissatisfied with negative results from another epidemiological study, got what they wanted using the same data set: "A reanalysis of the data … reveals a statistically significant increase in childhood cancers … when the evaluation is restricted to commercial power reactors, the vicinities closest to the plants and children of the youngest age group."

Greiser (2009), using previously published data from the UK, France, Germany, and Canada, plus self-collected date from cancer registries in the U.S., obtains 2127 cases of leukemia for the age group 0-4. This compares to an expected number of only 1969, and would, in the context of a designed experiment, indicate a significant (at 5%) increase in risk.

This significance however appears to be mostly due to data mining and publication bias; it vanishes completely – in fact, the sign of the observed effect is reversed - , once an obvious failure of the underlying model, the total disregard of important confounding factors, is accounted for. According to Ries et. al (1999, figure 6 and table 1.5), and confirmed by many others, important risk factors for childhood leukemia are race and sex. For instance, childhood cancer incidence in the U.S. is 30% higher for boys as compared to girls and almost double for whites as compared to blacks. For leukemia only, the highest incidence rates are observed among hispanics (48.5 per million as compared to 41.6 per million for whites and 25,8 per million for blacks). By far the lowest rates for any type of childhood cancer are observed for American Indians.

Also, leukemia incidence correlates strongly with income – the higher the income of the parents, the larger the risk of leukemia for kids. In Scotland, for instance, the incidence of childhood leukemia between the richest and the poorest subpopulations differs by as much as 50%. Other risk factors which have been identified so far are population density and population mixing, which both might likewise lead to an increased exposure of susceptible individuals to infections and local epidemics which in turn could later promote the onset of cancers of many types.

It would be surprising if these established covariates did not also affect the numbers reported by Greiser (2009). For instance, the plant that contributes most to the surplus of 158 leukemia cases in the Greiser study is San Onofre Nuclear Generating Station in Southern California, in the northwestern corner of San Diego County, south of the city of San Clemente. According to Greiser (2009, p. 21, table 4) there were 281 cases of childhood leukemia close to San Onofre (which in this case means: in San Diego County) in the 2001-2006 time period, compared to only 177 expected cases, an excess of 104. Therefore, this single data point contributes almost all of the 158 excess cases on which the “significant” increase of childhood leukemia in the vicinity of nuclear power plants is based.

Now, looking closer at the San Onofre site, it appears that virtually all confounding factors which have so far been established in the literature are higher there than elsewhere in the U.S. For instance, San Diego County is rather wealthy, with average household income 20 % above the national average. In addition, San Diego county has an above-average population of Hispanics and very few blacks. Also, both population density and population mixing are more pronounced in San Diego county than elsewhere in the U.S.. San Diego is the largest concentration of naval facilities in the world, with a constant moving in and out of families, which is even further accentuated by a large University and many more military facilities such as training camps, airbases, Marine Corps Recruit Depots and coast guard stations. All of these variables correlate strongly with childhood leukemia.

However, removing San Onofre from the Greiser (2009) data set, and adding some studies he has overlooked, the initial surplus of leukemia cases turns into a deficit, see Krämer and Arminger (2010). This section therefore shows that one complaint against significance testing raised by Ziliak and McCloskey (2008) – spurious significance due to bad practice – is certainly warranted by current practice in many fields. But it also shows that any claims as to significance of any sort require that the underlying model be reasonably correct.

4. Eleven years of significance testing in the German Economic Review

This section turns to type (b) and type (c) mistakes, i.e. neglecting large effects which are not "significant" and celebrating trivial effects which are significant only due to sample size, by scrutinizing all issues which have appeared so far of the German Economic Review. The German Economic Review is the official Journal of the Verein für Socialpolitik, an association of about 4000 German speaking economists from all over the world. It was inaugurated in 2000 as the English language successor to the venerable Zeitschrift für Wirtschafts- und Sozialwissenschaften, also known as "Schmollers Jahrbuch", with a history dating back to 1871. At the time of this writing it is in its 12th year of existence, so there are

11 complete volumes which in this section will be scrutinized for misleading applications of statistical tests of significance.

Table 1 provides a summary. It shows that about 40% of all articles published so far rely for their results both on data and on formal tests of significance of the confirmatory type (a vast majority being t-tests of the significance of some effect). This percentage has been increasing recently, but is still somewhat less than Ziliak and McCloskey (2008, chapters 6 and 7) find for the American Economic Review 1980-1999.

Table 1: Confirmatory significance tests in the German Economic Review

Volume	Number of articles	Articles with confirmatory significance tests	Number of confirmatory significance tests	Only sign, no effect
1	21	8	421	4
2	24	9	527	0
3	21	7	725	332
4	22	5	176	22
5	20	10	994	40
6	25	10	1359	87
7	22	9	653	0
8	24	11	1375	0
9	24	11	1171	0
10	28	12	1809	0
11	27	18	1365	1
Together	258	110	10575	486

The rather astonishing number of more than 10.000 tests of significance, i.e. about 1000 tests per volume, is of course due to the routine production of such tests by commercial software packages that are used by the authors to fit their models. And more often than not, they result in comments of the type “X has a significantly positive impact on Y”, without any reference to its magnitude. In 486 cases, the exact magnitude of the estimated coefficients are not even reported, the only information given being that they were “significant”.

Even this attribute is doubtful given the prevalence of comments like “table 2 presents the results of our final model estimation”. Obviously, this means that various estimates and tests were computed beforehand, with only the most “significant” results remaining to be shown, so the scientific value of such tests is close to zero: “Cheap t-tests, becoming steadily cheaper with falling computational costs, have in equilibrium a marginal scientific product equal to their cost” (Ziliak and McCloskey 2008, p. 112). And when the costs of tests tend to zero, their informational value seems to follow straight in line.

This is true even if the tests as such were properly done. Table 2 confirms that not even this is true in many cases. It provides information on various additional features of the 110 papers which report tests of significance.

Table 2: Three types of mistakes

Type (c) error: Confusion of economic and statistical significance of estimated coefficients or effects (“significant” used for both? Much ado about statistically significant but economically small coefficients or effects?)	62/110 = 56.4%
Type (b) error: Economically significant and plausible effects or coefficients discarded due to lack of statistical significance?	31/110 = 28.2%
No or only passing discussion of the dependence of “significance” on the correct specification of the model? (Independent and/or identically distributed observations etc.)	78/110 = 70.1%

Among papers that rely on some form of regression, most choose a linear functional form without much or even without any discussion whatsoever. Some also exclude or include variables solely on the basis of statistical significance, paying little attention to relevant economic theory. And only very rarely there was awareness of the multiple testing problem when only final versions of regression models were presented. Table 3 gives the details.

Table 3: Selected deficiencies of papers that use some sort of regression model

No detailed discussion of the appropriateness of the chosen model (no theoretical justification for the particular functional form, no or only cursory diagnostic testing etc)	56/98 = 57.1%
Explanatory variables included or excluded exclusively or almost exclusively on the basis of statistical significance?	20/98 = 20.4%
Several models tried, only the final one presented, but no awareness of the multiple testing problem	14/98 = 14.2%

The critique summarized in table 1 to 3 is not meant to denigrate a particular journal or empirical work in the German Economic Review as such. In fact, the particular approach which is criticised here appears to be common to most economic journals in the world and is even more prevalent, if Ziliak and McCloskey (2008) are to be believed, in the American Economic Review, which is the leading journal in the field. Also, there are many fine empirical papers in the German Economic Review which, to convey their message, do not rely on confirmatory significance tests at all.[4] And even if tests are reported, it is sometimes with some sort of tongue in cheek, to pay respect to some tradition which not even the authors do take seriously any more. And it is exactly this what the present paper wants to emphasize: that the

[4] A recent example is Bachmann and Burda (2010), who convincingly summarise and explain labour market dynamics in Germany, using lots of tables and figures, but no t-tests whatsoever.

endless tables of t-values that adorn most empirical papers nowadays are indeed what Ziliak and McCloskey call them - a needless waste of space and time.

5. Specification testing vs. searching for effects

An important point often overlooked in the significance debate is that any such claim – no matter whether it only concerns the existence or also the size of an effect – is only valid if the underlying statistical model describes the data reasonably well. In particular, as was shown in section 3, one has to ensure that all relevant explanatory variables have been properly accounted for. And it is here that statistical test of significance can help researchers along their way a lot. This is best exemplified with the help of the standard linear regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_K x_{iK} + u_i \quad (i=1,\ldots,n),$$

where the y's are to be explained, the x_{ik} are observations on K explanatory variables (regressors, exogenous variables, design variables), and the u_i's are unobservable disturbance terms, presumably uncorrelated, with equal variance and expectation zero. Confirmatory testing in this context means establishing the "significance" of individual regressors, i.e. testing H_0: $\beta_k=0$ for some k = 1,…,K. As was shown in section 4, and is confirmed by independent investigations by Ziliak and McCloskey (2008), well above 99 % of all statistical tests reported in a typical economics journal are of this type, with all the ensuing complications discussed in sections 2 and 3.

What is much more rarely done, but should be standard practise, is testing whether the model that is entertained provides a proper approximation to the data in the first place. Only in that case do tests of the confirmatory sort apply. And as shown in Krämer et al. (1985), most empirical papers, even in decent journals, fail such specification tests, often by wide margins. Among things that can go wrong here are omitted regressors (see section 3), non-linearity, in particular interaction effects, measurement errors, endogeneity or structural changes in the β-coefficients. Only if such deficiencies can be ruled out with some confidence does it make sense to talk about "oomph", i.e. the size of the βs, (and, if one so chooses, to test for their significance). Or to put this differently, one should first test whether the assumptions concerning the (conditional) first moments of the y's are indeed correct before proceeding to establish any kind of effect.[5] So if one takes seriously what most critics of standard statistical significance testing maintain, that it is the size and not the significance of effects which really counts, then one has to do some significance testing first.

Krämer and Sonnberger (1986) provide an overview of the early literature of statistical specification testing. An extremely simple procedure known as the RESET (Regression Specification Error Test) for instance only involves adding artificial regressors like squares, cubes or cross products of the initial regressors and testing whether they are significant. If so, there is evidence that the initial linear functional form is not correct. Or, for time series data, one could simply compare parameter estimates obtained from the initial model to estimates obtained from the same data after first differencing. If the model were correct, both estimators estimate the same things and should be close to each other. If not, there is evidence again that something is wrong with the model (an idea which has been generalized by Hausman (1978) to various other pairs of estimators). Krämer and Sonnberger (1986) collect together a generous toolbox of such techniques for checking model adequacy.

A related class of specification tests do not challenge a given model, because the underlying model is in most cases rather obvious and simple, but test whether or not certain parameters in this model are compatible with established economic theory. An example is the test for weekday anomalies for stock returns, see Krämer and Runde (1992). Financial theory requires that expected excess returns are positive and equal to each other for all days of the week, or that successive returns have autocorrelation zero. Again, one is not interested here in the size of the effect, but rather in whether one exists in the first place, the only problem being the distinction between statistical and practical significance (small deviations from theory cannot

[5] As compared to the desasters resulting from incorrectly specified first moments, the implications of incorrectly specified second or even higher moments (autocorrelation, heteroskedasticity,

be exploited due to trading costs). One might call such procedures "theory-attacking-tests" (as opposed to "theory-confirming-tests", which are the prime target of the Ziliak-McCloskey critique.)

Unfortunately, specification tests and theory attacking tests are a distinct minority among statistical significance tests reported in economics journals. Table 4 gives the respective figures for the German Economic Review.

Table 4: Papers with exploratory or "specification" test (i.e. test where an acceptance of Ho is fine)

Number of papers where such tests are done at all	26
Number of papers which discuss the power of such tests	4/26 = 15,4%

Examples of papers from recent volumes of the German Economic Review that rely at least partially on specification tests are Zarzoso et al. (2009. p. 327): "Specification tests also rejected the inclusion of a quadratic aid-term in the estimated equation") or Feld and Reulier (2009), who investigate the effect on a Swiss canton's personal income tax rate of various regressors, including the corresponding rates of neigbouring cantons. In addition to lots of confirmatory testing, they also test whether their regressors are truly exogenous: "Equation (1) cannot consistently be estimated by OLS because there is an obvious endogeneity problem. Hausman tests indicate that the neigbouring tax rates at the local level or at the regional levels are endogenous" (p. 98).

More recently, Holtemöller and Schulz (2010, p. 473) present a Wald-Statistic to check whether investors behave rationally in the housing market, and Feld and Schneider (2010, p. 130) test for the adequacy of an indicator model of the shadow economy. However, tests of these types are still dwarfed in number by mindless batteries of t-tests attached to parameter estimates. In the German Economic Review, the terms "specification test", "specification testing", RESET or "Hausman

nonnormality) appear rather minor indeed and can also easily be remedied.

test” appear less than ten times each in eleven years, that is less than once a year. Therefore, as long as such procedures are not standard in applied econometrics, Ziliak and McCloskey (2008) do have a point.

6. Conclusion

The admonition often heart recently to stop testing in empirical economics is partially mistaken. While it is true that confirmatory testing, where the null hypothesis is only entertained as a dummy to help establishing a prearranged alternative, has a huge potential to mislead, and does indeed mislead in many applications, specification testing is more important than ever. Therefore, the advice should be, not to abandon the concept of significance, but to shift the focus to other types of null hypotheses.

References

Altman, D.G., K.F. Schulz, and D. Moher (2001). "The Revised CONSORT Statement for Reporting Randomized Trials: Explanation and Elaboration." *Annals of Internal Medicine*, 134, 663-694.

Bachmann, R. and Burda, M. (2010): Sectoral transformation, turbulence and labor market dynamics in Germany," *German Economic Review* 11, 37-59.

Beck-Bornholdt, H.-P. and Dubben, H.-H. (2004): *Unausgewogene Berichterstattung in der medizinischen Wissenschaft - publication bias-*, Hamburg (Institut für Allgemeinmedizin des Universitätsklinikums Hamburg-Eppendorf).

Bündnis 90 / Grüne (2009): "AKWs erhöhen das Leukämierisiko", Press Release, Sep. 7.

Denton, F.T. (1985): „Data mining as an industry“, *The Review of Economics and Statistics* 67, 124 – 127.

Feld, L. P and Reulier, E. (2009): "Strategic tax competition in Switzerland: Evidence from a panel of the Swiss cantons," *German Economic Review* 10, 91-114.

Feld, L.P. and Schneider, F. (2010): „Survey on the shadow economy and undeclared earnings in OECD countries“, *German Economic Review* 11, 109-149.

Greiser, E. (2009): Leukämie-Erkrankungen bei Kindern und Jugendliche in der Umgebung von Kernkraftwerken in fünf Ländern, Report prepared by the political party Bündnis 90 / Grüne, see http://www.gruene-bundestag.de/cms/archiv/dokbin/302/302113.studie_leukaemierisiko.pdf

Haller, H. and S. Kraus (2002): "Misinterpretation of significance: A problem students share with their teachers?" *Methods of Psychological Research Online* 7, 1 – 20.

Hausman, J. A. (1978), "Specification Tests in Econometrics", *Econometrica* 46, 1251-1271.

Hoffmann, W., Kuni, H. and Ziggel, H. (1996): "Leukämiesterblichkeit in der Nähe von japanischen Atomkraftwerken doch erhöht“, *Strahlentelex* 238, 2 – 5.

Holtemöller, O. and Schulz, R. (2010): „Investor rationality and house price bubbles: Berlin and the German reunification“, *German Economic Review* 11, 465-486.

Körblein, A. and Hoffmann, W. (1999): "Childhood cancer in the vicinity of German nuclear power plants“, *Medicine & Global Survival* 6, 18-23.

Kruskal, W. (1968): "Tests of statistical significance." In: David Sills et al. (ed): *International Encyclopdedia of the Social Sciences*, New York (McMillan), 238-250.

Krämer. W. (2008): *Denkste – Trugschlüsse aus der Welt des Zufalls und der Zahlen*. 8th paperback edition, München (Piper).

Krämer, W. (2010): *So lügt man mit Statistik* (11th paperback edition), München (Piper).

Krämer, W., Sonnberger, H., Maurer, J. and Havlik, P. (1985): „Diagnostic checking in practice," *Review of Economics and Statistics* 68, 118-123.

Krämer, W. and Sonnberger, H. (1986): *The Linear Regression Model under Test,* Heidelberg (Physica-Verlag).

Krämer, W. and Runde, R. (1992): "The holiday effect: yet another capital market anomaly?" in Schach, S. and Trenkler, G. (Hrsg.): *Data analysis and statistical inference: Festschrift in honour of Friedhelm Eicker*, Bergisch-Gladbach (Eul-Verlag), 453 – 462.

Krämer, W. and Gigerenzer, G. (2005): „How to confuse with statistics. The use and misuse of conditional probabilities," *Statistical Science* 20, 223-230.

Krämer, W. and Arminger, G. (2010): "True believers or numerical terrorism at the nuclear power plant," Forschungsbericht Nr. X , Fakultät Statistik, Universität Dortmund.

Lovell, M.C. (1983): "Data Mining", *Review of Economics and Statistics* 65, 1 – 12.

McCloskey, D. (1985): "The loss function has been mislaid: The rhetoric of significance tests", *American Economic Review* 75, 201-205.

McCloskey, D (2002): *The Secret Sins of Economics*, New York (Wiley).

Ries, L. A. G., Smith, M. A., Gurney, J. G., Linet, M., Tamra, T., Young, J. L. and Bunin, G. R. (1999): "Cancer Incidence and Survival among Children and Adolescents: United States SEER Program 1975-1995", National Cancer Institute, Bethesda, MD.

Rozeboom, W.W. (1960): "The Fallacy of the Null Hypothesis Significance Test", *Psychological Bulletin*, 57, 416-428.

Schuchard-Fischer, C., K. Backhaus, H. Hummel, W. Lohrberg, W. Plinke and W. Schreiner (1982): *Multivariate Analysemethoden – Eine anwendungsorientierte Einführung.* 2nd edition, Berlin (Springer).

Sterling, T. R. (1959): "Publication decisions and their possible effects on inferences drawn from tests of significance – or vice versa," *Journal of the American Statistical Association* xxx, 30-34.

Stern J.M., Simes, R.J. (1997): "Publication bias: evidence of delayed publication in a cohort study of clinical research projects", *British Medical Journal* 315, 640 – 645.

Tang, D., Geller, N. L., and Pocock, S. J. (1993): „On the design and analysis of randomized clinical trials with multiple endpoints," *Biometrics* 49, 23-30.

Taubes, G. (1995): "Epidemiology faces its limits", *Science* 269, 164-169.

Todhunter, I. (1949): *A History of the Mathematical Theory of Probability from the Time of Pascal to that of Laplace,* New York (Chelsea) (first edition 1865).

Tyler, R. W. (1931): "What is statistical significance?" *Educational Research Bulletin* 10, 115-118, 142.

W. Wyss (1991): *Marktforschung von A – Z.* Lucerne (Demascope).

Zarzoso; I., Nowak-Lehmann, F., Klasen, S. and Larch, M. (2009): "Does German development aid promote German exports?" *German Economic Review* 10, 317-338.

Ziliak, S. and McCloskey, D. (2008): *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice and Lives*, University of Michigan Press.