DOWNLOAD PDF CURRENT ISSUES IN STATISTICAL INFERENCE

Chapter 1 : Journal of Statistical Planning and Inference - Elsevier

Evaluating Therapeutic Interventions: Some Issues and Experiences Fleming, Thomas R., Statistical Science, Estimating the Causal Effects of Marketing Interventions Using Propensity Score Methodology Rubin, Donald B. and Waterman, Richard P., Statistical Science,

Lecture Notes-Monograph Series, Vol. Essays in Honor of D. Basu , pp. JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. Dawid, University College London, Department of Statistical Science, England Abstract The basic theory of the prequential approach to data analysis is described, and illustrated by means of both simulation experiments and applications to real data-sets. Introduction The prequential approach to the problems of theoretical statistics was introduced by Dawid It is based on the idea that statistical methods should be assessed by means of the validity of the predictions that flow from them, and that such assessments can usefully be extracted from a sequence of realized data-values, by forming, at each intermediate time-point, a forecast for the next value, based on an analysis of earlier values. The main emphasis is on probability forecasting, requiring that one describe current uncertainty about the predictand by means of a fully specified probability distribution. However, point forecasts, or other forms of prediction, can also be accommodated. The purpose of the above paper was to indicate the fertility of the prequential point of view for furthering understanding of traditional concerns of theoretical statistics, such as consistency and efficiency. However, the prequential approach is essentially data-analytic. As such, it is particularly well suited to empirical investigation of the structure and properties of real-world observations, and their sources. In this paper, we shall discuss some of the ways in which prequential assessment may be applied in practical problems, including goodness- of-fit, model choice and density estimation. These methods are illustrated, by means of simulation experiments and applications to real data. Dawid methods on this basis provides a guide albeit imperfect as to their likely relative future performance. Starting from a parametric family of such methods,? This itself needs to be assessed by its prequential loss i? Prequential assessment of past predictive performance is very close in spirit to the method of cross-validation Stone, but bases its prediction for Yk on all previous outcomes, rather than on all outcomes distinct from Yk. In both methods, the intention is to avoid the bias involved in letting Yk contribute to its own prediction, and so to produce an honest assessment of uncertainty. Probability Forecasting One way to choosing the action ak, after observing? Thus if Nature is regarded as generating Y from P, then using? Then the optimal sequence of actions is just the sequence P? We This content downloaded from It is interesting to note that, if the distributions? In this case we shall never achieve an ultimate preference for either PFS, and it seems that we remain forever in a quandary as to which to use for further forecasts. However, a result of Blackwell and Dubins shows that, in this case, the forecasts produced by? Dawid variety of tests can be based on the observed values u. To assess uniformity, we might examine the? This could be tested formally using, say, the Kolmogorov-Smirnov statistic. One should also inspect the m? A simple indicator of trend is provided by the uniform conditional test Cox and Lewis, or y- j? This should give an approximate diagonal line. More formally, we can construct test- statistics such as? Under very weak conditions, not requiring independence,? Seillier and Dawid, and inde- pendent of statistics based on disjoint subsets. It is noteworthy that all the methods described above are applicable given only the two sequences, of outcomes and of their probability forecasts, and make no reference to the structure of? This is in accord with the Prequential Principle Dawid, Consequently, these methods can be used to test the overall goodness-of-fit of a parametric model. If the distribution or model being used fails to describe the data, it may be possible to massage it to provide a better fit. Thus suppose that the w? This distribution could itself be estimated, either parametrically or nonparametrically as in Density estimation below. If the estimate based on? In the case, if previous occasions on which the same probability forecast as? Such adaptive recalibration methods can improve the performance of a badly chosen initial model, although there can be no guarantee that they will, since the recalibration is based on the past but applied to the future. This content downloaded from We might then optimize the choice between these at each time- point. Of course, such a two- stage optimization strategy needs assessing afresh in its own right. The method extends to more stages, and to an arbitrary collection of models at each stage, but clearly less trust can be placed in prequential analyses iterated to more stages: In place of repeated optimization, one can take a Bayesian approach, assigning prior weights a and 1 - a to F and Q. After observing y k , with prequential joint density f y k under? The overall prequential likelihood for this strategy is simply af y n -f 1 - ct g y n. Again the method extends simply to more models and more stages. If one has a finite or countable collection of alternative models, and the data arise from some distribution in one of these, either of the above methods will be consistent and asymptotically optimal, in the sense that their forecasts will tend to those given by the true distribution, and at the fastest possible rate. However, for finite data-sets, the forecasts under the two methods may look rather different. In either case, if the true distribution is contained in a model of high-dimensionality, early analysis will generally tend to favor incorrect models of low dimensionality. This is intuitively sensible, since, early on, the mis- modelling bias may well be less of a problem than the imprecision involved in trying to estimate many parameters. As an alternative to allowing such transient behavior to be entirely data- driven, as above, one might build it in directly, by setting out with a strategy for choosing, at each stage, the complexity of the model to be fitted and how it is to be used for prediction. Different strategies, all yielding consistent estimates of the true model and which use each fixed model efficiently will all be asymptotically equally good. However, their transient behaviors, which may be long-lasting, can be very different, with some yielding much larger prequential log-likelihoods or, more generally, much smaller prequential losses than others even though these discrepancies will be bounded as the sample size goes to infinity. More empirical and theoretical work is needed to indicate good forms for such strategies. A sensible super-strategy could be built up from a low- dimensional parametrized family of such strategies, using optimizing or Bayesian This content downloaded from This could combine good transient behavior with sensitivity to the data and avoidance of data-mining. Non-parametric Approximation Many non-parametric problems, such as density estimation or fitting a stationary time-series, can be approached through a sequence of finitely parametrized methods, such as fitting histogram or kernel density estimates with adjustable bin width, or autoregressive models of various finite orders. One can then apply the techniques of the previous section, even though none of the models used is now expected to contain the distribution generating the data. The component models will generally each be characterized by some quantity, such as kernel width w or autoregressive order p , which controls the balance between over-fitting tracking noise in the data and over-smoothing not picking up the signal. Prequential choice of such a quantity will start out with a preference for smoothing large w, small p , and then, as the data-sequence grows longer and can support more detailed modelling, gradually move towards fitting the past data more and more closely w? Such a method will often be prequentially consistent for a wide range of generating distributions, and can provide sensible answers based on finite data-sets, by making the predict i vely optimal compromise between fitting and smoothing. Investigation of the structure of good strategies, for choosing the model to fit at each stage, is still more vital in this context, since the behavior described as transient in the previous section now extends to infinity! Again, much further empirical and theoretical work is required to illuminate this problem area. Results were as follows. This strategy itself had a prequential log-likelihood of , better than that for any fixed k. Its prequential log-likelihood too was Prequential Log- Likelihoods k Optimization Bayes The Bayes strategy ended by assigning probability 0. This strategy had prequential log-likelihood of The Bayes strategy assigned probabilities 0. Simple histogram-type density estimators were constructed from data- values in [0,1], based on a division of the unit interval into k equal sub-intervals. For each initial sub-sequence of data, the current density estimate was used to forecast the next observation. Calibration plots and test-statistics were constructed for various subsets of the data, corresponding to the months of the year, and to specified outcomes of the three previous days. Table I gives, for each month, the overall proportion y of rainy days, and the average forecast probability p. The results are shown in. It is of interest that, for any collection of symptom variables, adding in the location indi- cator Q leads to worse predictions. This offers some empirical support for the This content downloaded from The data-set contained responses to 60 test items from students. At an intermediate stage, a number of items, 1 to k - 1 say, have been accepted, and item k is under test. The fitted probability for this omitted response can then be calculated, and the process repeated with m increased by 1. Comparison of these forecast probabilities with the actual responses where these were not missing then allows assessment of the fit of item k to the model. For testing item 60, with all other items included, the probabilities were grouped into 8 intervals, with counts, average probability and relative frequency of a right answer as given in Table III. The observed This content downloaded from Dawid value of The data comprised inter-failure times ranging between 0 and seconds, and the models used all incorporated reliability growth improved performance after each bug-fix. Some forecasting systems used optimization, some were Bayesian, others com- bined the two methods. The results are summarized in Table IV. It appears that the method of data analysis is less important here than choosing a good model. Measured by prequential likelihood, the optimal system was 6. The authors also considered adaptive recalibration of the above systems, as well as Bayesian and optimizing strategies for combining them, leading in all cases to improvements in performance. However its implementation leaves plenty of scope for variations, and is currently more art than science. Further work should lead to an improved understanding, and give guidance on good strategies of applying the method.

Chapter 2 : STATISTICAL INFERENCE

DESCRIPTION. Prequential Data Analysis Author(s): A. P. Dawid Source: Lecture Notes-Monograph Series, Vol. 17, Current Issues in Statistical Inference: Essays in Honor of D. Basu (),.

Shape and the information in medical images: Bookstein , " This keynote address reviews several techniques f rom morphometries the multivariate biometrics of shape developed mainly in the context of medical image anal-ysis over the last decade. The new techniques pro-vide powerful tools for geometric tasks that arise in the course of most analyses of The new techniques pro-vide powerful tools for geometric tasks that arise in the course of most analyses of medical images in groups. These tasks include standardizing against Euclidean similarity transformations or shear transformations, en-coding informative prior knowledge about shape varia-tion, and detecting, testing, and visualizing linear sta-tzstical patterns of variation within or between groups. I review the features of the present toolkit, the stan-dard underlying data models entailed, and some of the extensions that reach out to the additional information content of medical images for common clznical or sci-entific applications. The paper reviews various topics in shape analysis. In particular, matching configurations using regression is emphasized. Connections with general shape spaces and shape distances are discussed. Matching two configurations and the extension to generalized matching are illustrated with applications in electrophoresis and biology. Shape distributions are briefly discussed and inference in tangent spaces is considered. Finally, some robustness and smoothing issues are highlighted. Show Context Citation Context Inference for random samples is relatively straightforwardand some classical inference procedures were given by Kent and Mardia and Dryden b , for example. In some applications the shap Geologists have long complained of their low status in the wider culture. A few years ago, one major geological society even convened a working party to consider how this status could be improved. This is not just a problem of morale. In-creasingly science is going to be judged not only by its relev In-creasingly science is going to be judged not only by its relevance but also by its popular impact. Perhaps it is time to bring in the image con-sultants. First we need to define the problem: The science is past its philosophical prime One hundred and firty years ago geologists were the lead-ing scientific intellectuals because they grappled with philosophical problems at the frontiers of knowledge: In the course of the 19th Cen-tury all these problems started slipping away from geol-ogy: Yet it continues to offer insights into our world In the 20th Century the high points of the Earth sciences have concerned our world view: Planetary regulation will remain our domi-nant contribution to Society. One pop-sular algorithm begins with an arbitrary but reasonablesshape for instance, that of the first specimen of thessample , Procrustes-fits all the specimens to that one,saverages the fitted fo

Chapter 3 : Issues in Inference on Vimeo

On the history of statistics and probability: proceedings of a symposium on the American mathematical heritage, to celebrate the bicentennial of the United States of America, held at Southern Methodist University, May , / edited by D. B. Owen ; featuring articles by W. G. Cochran, H. O. Hartley, and Jerzy Neyman.

Shewhart at Bell Laboratories in the early s. Shewhart developed the control chart in and the concept of a state of statistical control. Statistical control is equivalent to the concept of exchangeability [1] [2] developed by logician William Ernest Johnson also in in his book Logic, Part III: The Logical Foundations of Science. Shewhart consulted with Colonel Leslie E. Deming was an important architect of the quality control short courses that trained American industry in the new techniques during WWII. The graduates of these wartime courses formed a new professional society in , the American Society for Quality Control, which elected Edwards as its first president. Common cause and special cause statistics Shewhart read the new statistical theories coming out of Britain, especially the work of William Sealy Gosset , Karl Pearson , and Ronald Fisher. However, he understood that data from physical processes seldom produced a " normal distribution curve"; that is, a Gaussian distribution or " bell curve ". He discovered that data from measurements of variation in manufacturing did not always behave the way as data from measurements of natural phenomena for example, Brownian motion of particles. The notion that SPC is a useful tool when applied to non-repetitive, knowledge-intensive processes such as research and development or systems engineering has encountered skepticism and remains controversial. This implies that SPC is less effective in the domain of software development than in, e. Variation in manufacturing[edit] In manufacturing, quality is defined as conformance to specification. However, no two products or characteristics are ever exactly the same, because any process contains many sources of variability. In mass-manufacturing, traditionally, the quality of a finished article is ensured by post-manufacturing inspection of the product. Each article or a sample of articles from a production lot may be accepted or rejected according to how well it meets its design specifications. In contrast, SPC uses statistical tools to observe the performance of the production process in order to detect significant variations before they result in the production of a sub-standard article. Any source of variation at any point of time in a process will fall into one of two classes. It refers to many sources of variation that consistently acts on process. These types of causes produce a stable and repeatable distribution over time. It refers to any factor causing variation that affects only some of the process output. They are often intermittent and unpredictable. Most processes have many sources of variation; most of them are minor and may be ignored. If the dominant assignable sources of variation are detected, potentially they can be identified and removed. When they are removed, the process is said to be "stable". When a process is stable, its variation should remain within a known set of limits. That is, at least, until another assignable source of variation occurs. When the package weights are measured, the data will demonstrate a distribution of net weights. If the production process, its inputs, or its environment for example, the machines on the line change, the distribution of the data will change. For example, as the cams and pulleys of the machinery wear, the cereal filling machine may put more than the specified amount of cereal into each box. If the manufacturer finds the change and its source in a timely manner, the change can be corrected for example, the cams and pulleys replaced. Understanding the process and the specification limits. Eliminating assignable special sources of variation, so that the process is stable. Monitoring the ongoing production process, assisted by the use of control charts, to detect significant changes of mean or variation. Control charts[edit] The data from measurements of variations at points on the process map is monitored using control charts. Control charts attempt to differentiate "assignable" "special" sources of variation from "common" sources. Using control charts is a continuous activity, ongoing over time. Stable process[edit] When the process does not trigger any of the control chart "detection rules" for the control chart, it is said to be "stable". A process capability analysis may be performed on a stable process to predict the ability of the process to produce "conforming product" in the future. Excessive variations[edit] When the process triggers any of the control chart "detection rules", or alternatively, the process capability is low , other activities may be performed to identify the source of the excessive variation. The tools used in these extra activities include: Ishikawa diagram , designed experiments , and Pareto charts. Designed experiments are a means of objectively quantifying the relative importance strength of sources of variation. Once the sources of special cause variation are identified, they can be minimized or eliminated. Steps to eliminating a source of variation might include: Process stability metrics[edit] When monitoring many processes with control charts, it is sometimes useful to calculate quantitative measures of the stability of the processes. These metrics can also be viewed as supplementing the traditional process capability metrics. Several metrics have been proposed, as described in Ramirez and Runger. Mathematics of control charts[edit] Digital control charts use logic-based rules that determine "derived values" which signal the need for correction.

Chapter 4 : Statistical process control - Wikipedia

A statistical inference will be defined for the purposes of the present paper to be a statement about statistical populations made from given observations with measured uncertainty.

The missing reference to Chow has been included in this edition. Guidelines and Explanations" , it is hard to believe that anyone could object very much. Be careful to distinguish between mere contrast groups and true control groups. Report reliability and validity figures when using a psychometric questionnaire. Pay attention to possible biases introduced by experimenter expectancies, and report the steps you took to attenuate them. Report the process that led to your choosing the particular size of your sample, and attend to the power of your test. Examine plots of your data and try to include graphical representations of them in your report, not just numerical statistics. Random selection of samples is not necessary. Contrast groups do not provide adequate experimental control. Experimenter effects are chimerical. Effect sizes are primarily useful only for buttressing highly questionable meta-analytic procedures. Null-hypothesis significance testing remains the foundation of good statistical analysis. Power analysis is irrelevant to significance testing. As a result, sample size recommendations generated by power analysis are largely irrelevant as well. How are psychologists to sort out these conflicting claims? How could informed commentators disagree so strenuously on such fundamental aspects of so well-worn a topic as basic statistical analysis of behavioral scientific data? In part, I believe that the differences are not so great in some areas as it would at first appear. In part, I believe that Chow has adopted a radical posture in order to counterbalance what he views as an incautious assault on traditional statistical methods. I cannot comment on all the issues raised by Chow in the space allotted me here, so I will focus on four areas of disagreement between him and the Task Force: Chow argues, therefore, that there is no EEE effect for the Task Force to warn researchers about. There are a number of difficulties here. The point was that the expectancies of "data-collectors" whether authentic experimenters or not can have an adverse impact on the data actually collected. Chow found no such effect and attributed these findings to his modification of the procedure but, as we all know, one should not assert the null merely on the basis of a failure to reject it note, the issue here is not whether the null can be true, but rather what conclusion we should draw when we fail to reject it. I cannot tell with certainty whether Chow accepts this form of inference despite the "conventional wisdom. More important, perhaps, is the question of what Chow would have us do instead -- utterly ignore the possibility that the experimenter is subtly passing his or her expectancies on to her subjects? If so, Chow is, in effect, arguing against half of the standard "double-blind" research design itself, and we have far too much evidence in a wide array of disciplines that this procedure is a wise precaution for us to abandon it on the basis of a single study that failed to replicate a well-established result. Control Groups I believe that Chow simply misconstrued the intent of the Task Force on the matter of control and contrast groups. If we can neither implement randomization nor approach total control of variables that modify effects outcomes , then we should use the term "control group" cautiously. In most of these cases, it would be better to forgo the term and use "contrast group" instead. Wilkinson, , 3rd para. Null Hypothesis Testing Chow has made something of a name for himself over the past decade-and-a-half attempting to defend null-hypothesis significance testing NHST against its legions of critics. His campaign reached its zenith with the publication of Statistical Significance , and an open debate with many leaders in the field in Behavioral and Brain Sciences After reading through much of this material, I wonder whether more heat has been generated than light. Let us consider two passages. However, to say that something is not due to chance is really not saying very much at the theoretical level, particularly when NHSTP says nothing about whether or not an explanatory hypothesis or an empirical generalization receives empirical support. Some had hoped that this task force would vote to recommend an outright ban on the use of significance tests in psychology journals. Although this might eliminate some abuses, the committee thought that there were enough counterexamples e. Wlikinson, , 2nd para. Both sides seem to agree that NHST is frequently misused and misinterpreted in the psychological literature as it now stands. Both sides seem to agree that these abuses should be corrected posthaste. Both sides seem to agree that there is a legitimate, if limited, role for NHST. If only the participants in all sharply divided debates in psychology agreed on so many fundamentals! What are we going to do in those cases where significance testing is not appropriate or not sufficient? Naturally, these have their limitations, as Chow is quick to point out. The goal is to get the most relevant information to the reader in the most easily-apprehendable format. There is no reason I can see, nor that Chow presents, that NHST should remain the "default" statistic, or that effects sizes or graphical displays of data should be banned. As Gigerenzer , p. Power Analysis It is here that Chow is at his most radical. He only asks rhetorically if we should be "oblivious" to the fact. It is not that the power-analytic graph of the sampling distribution under Ho and under a predicted or desired form of H1 invokes a new "level of abstraction" in any substantive sense, but rather that it completes parts of the picture left vacant by the NHST graph: Is there any good logical or scientific reason for developing an elaborate analysis of the probability of the one error but not of the other? Chow worries that there are indefinitely many H1s that might be used to generate power numbers. Contrary to his claim that power analysis is "mechanical" Chow, , , the scientist must use discretion in selecting pertinent H1s to work with. Now, if Chow thinks that the style of analysis popularized by Cohen is suboptimal, he is, of course, at liberty to develop another. But to reject the analysis and control of Type II error altogether while continuing to insist on the careful analysis and control of Type I error is difficult to countenance. In a related vein, Chow presents the results of a simulation of significance testing under different sample sizes in order to refute the oft-made claim that significance is dependent on sample size. He demonstrates what I think most people already know -- as one increases sample size most aberrantly large sample differences will be diluted out. However, what I think most people actually mean when they claim that significance is dependent on sample size is, rather, that for a given difference between sample means, a larger sample generates a smaller standard error, making the obtained t larger. In addition the larger sample, via degrees of freedom, makes the critical t smaller. In this way, a difference that is not significant under one sample size, can be rendered significant under a larger sample size. This is demonstrated in Table 1.

Chapter 5 : Issues in Statistical Inference - Cogprints

Enter your mobile number or email address below and we'll send you a link to download the free Kindle App. Then you can start reading Kindle books on your smartphone, tablet, or computer - no Kindle device required.

Single mean against value specified as z in Ho Single proportion against value specified as z in Ho Difference between two means Significance of contingency table Single correlation Sampling Distribution of the Mean. Suppose that instead of taking a single random sample of 1, people, one took two such samples and determined the mean of each one. With 1, cases, it is likely that the two samples would have means that were close together but not the same. The larger the samples are, the more stable the mean is from one sample to the next. With only 20 people in the first and second samples, the means may vary a lot, but with , people in both samples, the means should be almost identical. Mathematically, it is possible to derive a distribution of the means of all possible samples of a given n even though only a single sample is observed. It can be shown that the mean of the sampling distribution of means is the population mean and that the standard deviation of the sampling distribution of the means is the population standard deviation divided by the square root of the sample size. The standard deviation of the mean is called the standard error of the mean: Four distributions used in statistical inference: A remarkable property of the sampling distribution of the mean is that with a large sample size, it will be normally distributed even though the population and sample distributions are skewed. One gets a general idea of how the sample did by seeing where the sample mean falls along the sampling distribution of the mean. Thus, one infers that the sample mean probably is based on a nonrandom sample. With cases, one should not be surprised to get a sample mean this far from the population mean. Being able to compare the sample mean to the population mean by using the sampling distribution is remarkable, but statistical theory allows more precision. One can transform the values in the sampling distribution of the mean to a distribution of a test statistic. The appropriate test statistic is the distribution of the normal deviate, or z-distribution. It can be shown that If the z-value were computed for the mean of all possible samples taken at random from the population, it would be distributed as shown in Panel D of Figure 2. It will be normal, have a mean of zero, and have a variance of 1. Thus, it is extremely unlikely that the sample is purely random. With the same sample mean but with a sample of only people, Using tabled values for a two-tail test, the probability of getting the sample mean this far from the population mean with a sample of people is. One should not infer that the sample is nonrandom, since these results could happen 25 percent of the time by chance. The four distributions can be described for any sample statistic one wants to test means, differences of means, proportions, differences of proportions, correlations, etc. While many of the calculations will be more complex, their logic is identical. Since most studies include multiple tests, interpreting results can become extremely complex. If a researcher conducts tests, 5 of them should yield results that are statistically significant at the. A conservative approach is to divide the Type I error by the number of tests conducted. This is known as the Dunn multiple comparison test, based on the Bonferroni inequality. For example, instead of doing nine tests at the. To be viewed as statistically significant at the. There are many specialized multiple comparison procedures, depending on whether the tests are planned before the study starts or after the results are known. Brown and Melamed describe these procedures. Sociologists use statistical inference to minimize the chance of accepting a main hypothesis that is false in the population. Still, by minimizing the chances of this error, sociologists increase the chance of failing to reject the null hypothesis when it should be rejected. Table 2 illustrates these two types of error. Type I, or a, error is the probability of rejecting H0 falsely, that is, the error of deciding that H1 is right when H0 is true in the population. If one were testing whether a new program reduced drug abuse among pregnant women, the H1 would be that the program did this and the H0 would be that the program was no better than the existing one. Type I error should be minimized because it would be wrong to change programs when the new program was no better than the existing one. The focus on Type I error reflects a conservative view among scientists. Type I error guards against doing something new as specified by H1 when it is not going to be helpful. If one failed to reject the null hypothesis that the new program was no better H0 when it was truly better H1 , one would put newborn children at needless risk. Type

II error is the chance of missing something new as specified by H1 when it really would be helpful. Power is 1 â€” B. Power measures the likelihood of rejecting the null hypothesis when the alternative hypothesis is true. Thus, if there is a real effect in the population, a study that has a power of. The power of a statistical test is measured by how likely it is to do what one usually wants to do: Using the example of a treatment for drug abuse among pregnant women, the power of a test is the ability to demonstrate that the program is effective if this is really true. Power can be increased. First, get a larger sample. The larger the sample, the more power to find results that exist in the population. Second, increase the a level. Rather than using the. The larger a is, the more powerful the test is in its ability to reject the null hypothesis when the alternative is true. There are problems with both approaches. Increasing sample size makes the study more costly. If there are risks to the subjects who participate, adding cases exposes additional people to that risk. An example of this would be a study that exposed subjects to a new drug treatment program that might create more problems than it solved. A larger sample will expose more people to these risks. However, sociologists are hesitant to raise a since doing so increases the chance of deciding something is important when it is not important. With a small sample, using a small a level such as. Many small-scale studies have a Type II error of over. This is common in research areas that rely on small samples. For example, a review of one volume of the Journal of Abnormal Psychology this journal includes many small-sample studies found that those studies average Type II error of. This means the psychologist had inadequate power to reject the null hypothesis when H1 was true. When H1 was true, the chance of rejecting H0 i. Some areas that rely on small samples because of the cost of gathering data or to minimize the potential risk to subjects require researchers to plan their sample sizes to balance a, power, sample size, and the minimum size of effect that is theoretically important. For example, if a correlation of. If a correlation is substantively and theoretically important only if it is over. Procedures for doing a power analysis are available in Cohen ; see also Murphy and Myous Power analysis is less important for many sociological studies that have large samples. With a large sample, it is possible to use a conservative a error rate and still have sufficient power to reject the null hypothesis when H1 is true. Therefore, sociologists pay less attention to B error and power than do researchers in fields such as medicine and psychology. When a sociologist has a sample of 10, cases, the power is over. When tests are extremely powerful to detect small effects, researchers must focus on the substantive significance of the effects. Statistical inference does not ensure substantive significance, that is, ensure that the result is important. With a sample of cases, this correlation will not be statistically significant; with a sample of 10, cases, it will be statistically significant. The smaller sample shows a weak relationship that might be a zero relationship in the population. The larger sample shows a weak relationship that is all but certainly a weak relationship in the population, although it is not zero. In this case, the statistical significance allows one to be confident that the relationship in the population is substantively weak. Whenever a person reads that a result is statistically significant, he or she is confident that there is some relationship. The next step is to decide whether it is substantively significant or substantively weak. Power analysis is one way to make this decision. One can illustrate this process by testing the significance of a correlation. A population correlation of. In other words, if a correlation is statistically significant but. It is just as important to explain to the readers that the relationship is substantively weak as it is to report that it is statistically significant. By contrast, if a sample correlation is. Figure 3 shows power curves for testing the significance of a correlation. These curves illustrate the need to be sensitive to both statistical significance and substantive significance. The curve on the extreme left shows the power of a test to show that a sample correlation, r, is statistically significant when the population correlation, p rho , is. With a sample size of around , the power of a test to show statistical significance approaches 1. This means that any correlation that is this strong in the population can be shown to be statistically significant with a small sample. What happens when the correlation in the population is weak? Suppose the true correlation in the population is. A sample with cases almost certainly will produce a sample correlation that is statistically significant, since the power is approaching 1. Figure 3 shows that even if the population correlation is just. Thus, any time a sample is 1, or larger, one has to be especially careful to avoid confusing statistical and substantive significance. The guidelines for distinguishing between statistical and substantive significance are direct but often are ignored by researchers: If a result is not statistically significant, regardless of its size in the sample, one should be reluctant to generalize it to the population. If a result is statistically significant in the sample, this means that one can generalize it to the population but does not indicate whether it is a weak or a strong relationship.

Chapter 6 : Holdings : Current issues in statistical inference : | York University Libraries

History and Philosophy of Psychology Bulletin Volume 14, No. 1, Issues in Statistical Inference Siu L. Chow Department of Psychology, University of Regina.

If it does not appear to be in cogprints you will be forwarded to the paracite service. Poorly formated references will probably not work. The nature and history of experimental control. American Journal of Psychology, 67, Experimental and quasi-experimental designs for research. Meta-analysis of pragmatic and theoretical research: Journal of Psychology, , Some reservations about statistical power. American Psychologist, 46, German Journal of Educational Psychology, 8, Rationale, validity and utility. Rationale, Validity and Utility. Statistical power analysis for the behavioral sciences Revised Edition. Things I have learned so far. American Psychologist, 45, American Psychologist, 49, An introduction to logic and scientific method. Remarks in defense as well as critique. Minnesota studies in the philosophy of science pp. University of Minnesota Press. Operationism and the concept of perception. Psychological Review, 63, On a distinction between hypothetical constructs and intervening variables. Psychological Review, 55, Theory testing in psychology and physics: Philosophy of science, 34, Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology, 46, A system of logic: University of Toronto Press. The logic of scientific discovery 2nd edition, originally published in The growth of scientific knowledge originally published in Three experiments in experimenter bias. Psychological Reports, 12, The effect of experimenter bias on the performance of the albino rat. Behavioral Science, 8, Experimenter effects in behavioral research Enlarged edition. Statistical methods in psychology journals: American Psychologist, 54 8 ,

Chapter 7 : CiteSeerX â€" Citation Query Current issues for statistical inference in shape analysis

Note: Citations are based on reference standards. However, formatting rules can vary widely between applications and fields of interest or study. The specific requirements or preferences of your reviewing publisher, classroom teacher, institution or organization should be applied.

View Large There are 2 important and related questions that have not yet been answered with respect to this study. The first question relates to the importance of the difference that has been detected. Although a mean difference in pain relief between the 2 treatments has been shown to be real, based on probabilistic statistics, it may be of little practical importance. On a cm VAS, a difference between treatments of 3. The outcomes of statistical tests need to be considered in the context of the situation to which they relate, and outcomes of clinical research must be subjected to clinical judgment. It is worth noting that the converse of the situation just outlined may also arise. In such a case, although the observed difference in pain relief is real for this sample, it cannot be assumed to reflect a real difference in the population. In order for such a finding to be applicable to general clinical practice, the observed difference must be shown to be real for the population. This will require a reduction in the risk of a Type 2 error, through an increase in sample size, a more precise method of measurement, or other means of reducing random error. Another unanswered question relates to the magnitude of this difference in pain relief between groups. This difference is an estimate of the difference that would exist if the full population of patients with FMS were studied. All we know is that the difference found in this particular sample is sufficiently great for it to be attributed to a genuine difference between the treatments rather than to chance variation in sampling or allocation to groups. We do not know how good an estimate it is of the true population difference ie, the difference we would find had we tested the treatments on the whole population of individuals with FMS. As Abrams and Scragg 11 point out, a probability value conveys no information about the size of the true effect. This is information, however, that we need in order to inform clinical practice. The CI assists in addressing these questions as to the clinical importance and magnitude of an observed effect and remedies some of the shortcomings of more conventional approaches to hypothesis testing. These points will be considered in detail following an account of interval estimation. The Nature of Confidence Intervals A sample statistic, such as a sample mean, provides an estimate of a population parameter. It provides a single estimate of the specific value of the parameter on the basis of the observed value of the statistic. As an adjunct to a single estimate, an interval estimate can be calculated. This interval estimate specifies a range of values on either side of the sample statistic within which the population parameter can be expected to fall with a chosen level of confidence. It is also the case that values toward the extremes of this interval are rather less likely to represent the population mean than those nearer the center. The population mean is either included or not included. A CI is used when examining a characteristic of a sample in this case, the mean pretest VAS score in terms of its degree of variability in the corresponding population. Width of Confidence Intervals For a given level of confidence, the narrower the CI, the greater the precision of the sample mean as an estimate of the population mean. There are 3 factors that will influence the width of a CI at a given level of confidence. First, the width of the CI is related to the variance of the sample scores on which it is calculated. If this sample variance can be reduced eg, by increasing the reliability of measurements , the CI will be narrower, reflecting the greater precision of the individual measurements. Selecting a sample that is more homogeneous will reduce the variance of scores and thereby increase their precision. This increase in precision occurs because the variance of a statistic, as expressed by its standard error, decreases as sample size increases. Figure 1 shows 4 samples of a progressively greater size drawn from a single population of physical therapists and the mean period of postqualification experience for each sample. The mean is precisely the same in each case, but the CI becomes narrower as the sample size increases. Means and confidence intervals of years of qualified experience for progressively larger samples of physical therapists drawn from a single population. View large Download slide Means and confidence intervals of years of qualified experience for progressively larger samples of physical therapists drawn from a single population. Third, the chosen level of confidence will influence the width of the CI. With a higher level of confidence, the interval needs to be wider in order to support the claim of having included the population parameter at the chosen level of confidence. It is not the case, however, that, at a given confidence level, a narrow CI is any more or less likely than a wider CI to be one that contains the population parameter. The probability of including the parameter is determined by the chosen confidence level, not by the width of the particular CI concerned. The probability of inclusion, however, is the same in both cases. The width of a CI is indicative of its precision ie, the degree of random error associated with it , but it does not convey its accuracy ie, whether it includes the population parameter , which is determined by the chosen level of confidence. The usefulness of a CI depends on the point statistic eg, the sample mean on which it is based being an unbiased point estimate. If systematic error is present in a study, the point estimate will lie at some distance from the true value of the parameter. In such a case, a CI based on a large sample will, paradoxically, be more misleading than one based on a small sample. Imagine that the point estimate of 9. It is evident that, unlike the 2 wider CIs, the narrow CIs, based on the larger samples, actually exclude this value. In the presence of systematic error, the lesser precision afforded by a wide CI actually increases the likelihood of its including the true population value. This example illustrates the fundamental point that increases in sample size will only assist in dealing with random, not systematic, error. Systematic error is usually an issue in study design rather than a function of the statistics used. The t statistic corresponds to a particular probability level and thus a confidence level , the degrees of freedom are determined by sample size, and the standard error of the mean represents sample variance. Confidence intervals can also be constructed for sample statistics other than the mean and in relation to samples that do not satisfy the assumptions of parametric statistics.

Chapter 8 : Kempthorne: Intervention experiments, randomization and inference

Don't report statistics you don't understand, and make sure the assumptions of the statistics you employ have been satisfied. Causation can be a thorny issue; don't be over-awed by the output of "causal modeling" software.

Chapter 9 : Barlow, Irony: Foundations of statistical quality control

These issues are especially relevant when the number of simultaneous tests can be very large, as is increasingly the case in the analysis of data from DNA microarrays. [citation needed] Bayesian statistics: A list of open problems in Bayesian statistics has been proposed.