Teaching Hypothesis Testing. Can it Still be Useful?

Henrik Dahl
Agder College, Department of Mathematics
Tordenskjolds gate 65
Kristiansand, Norway
E-mail: Henrik.Dahl@hia.no

SUMMARY

Hypothesis testing have been with us for more than a century, but has recently come under attack as less useful. It is important that statistics is taught in a way that makes students aware of possible misuse of the methods and how to avoid such misuse. Recently, statistics has become a part of the curriculum in secondary schools (16-18) in Norway, and this has not been without controversy. One criticism is the lack of interesting applications of the theory. An example is presented which helps students build their knowledge. Model-based statistics using modern technology may help students avoid many pitfalls in testing hypotheses.

1. Introduction

Hypothesis testing has been standard apparatus for doing statistics for more than a century, but has recently come under attack as less useful, perhaps because of misuse of the machinery and misinterpretation of the results (Morrison & Henkel, 1970), (Batanero, 1997). Chow (1996) and others have considered the issue of testing from the perspective of a researcher in the social sciences. In this paper we are considering statistical tests from the perspective of the teacher. We are discussing the didactic problems set by the introduction of statistical inference at the secondary level (16-18), presenting some key points in the teaching of the topic, as well as an example which could be useful to help students contextualise and better understand the concepts involved in statistical testing.

2. Teaching inference at secondary school levels (16-18)

Recently statistics has become a part of the curriculum in secondary schools in Norway, including Testing Hypotheses and Interval Estimation. It was introduced in a government reform named Reform 94, and the first students finished school using this approach in 1997. The introduction of statistics in Norwegian secondary schools has not been without controversy. There have been arguments (published in major newspapers) that the introduction of more topics in the mathematics curriculum is the main reason for the lower quality of students entering Norwegian universities. The new textbooks were ready just before teaching started the autumn of 1996. The Norwegian Statistical Society in 1996 formed a committee consisting of Jostein Lillestøl (Norwegian School of Economics), Ivar Heuch (University of Bergen) and Henrik Dahl (Agder College) to evaluate the different textbooks. Separate evaluations were sent to the publishers and also a collective evaluation was published in the Newsletter of the Norwegian Statistical Society (Heuch, 1998).

The main problems were: How to introduce probability, How to introduce and handle the Normal distribution, The distinction between parameters and estimators, The idea of inference (typically the technical problems are given more attention than the philosophical problems). Interval estimation was problematic in a majority of the textbooks and the corresponding problem of interpreting level of significance in testing also occurred. It is to be hoped that the constructive

criticism given by the committee will improve the texts. Other criticisms were the lack of interesting applications of the theory.

3. Some key points in teaching statistical tests

Didactic research has suggested that some of the problems of misapplications of statistical tests could be due to lack of understanding of the key concepts of these tests. It is easy for the students to get lost in technical details. Therefore, we should concentrate not just on the mathematical formulae, but making clear the philosophy behind the methods. In addition to the basic concepts of model, parameter, test statistic, sampling distribution, null and alternative hypotheses and test criterion, a suggested program for teaching hypothesis testing should have the following ingredients:

1. When decisions are based on data, which contain randomness, there is a certain risk of committing errors. When testing hypotheses there are two possible errors: Type 1 error and Type 2 error.
2. It is not possible to make more than one error in a certain situation, but it is necessary to make the choice of statistical method considering both possibilities.
3. It is important to choose a statistical method which has low error probabilities of either type. It is possible to study different statistical methods and evaluate their properties before data is analysed. This study should preferably consider both Type 1 and Type 2 errors. Even if it is not possible or feasible to do all the calculations, both types of errors should be contemplated.

By following this program some of the controversies over testing are avoided. It is important to make students aware that sometimes scientists ignore 1. and 2. before they test hypotheses at a fixed level. This is surely not recommended by Neyman & Pearson (1933) who specifically addressed Type 2 error to choose the optimal test. Even Fisher (1956 p.42) did not recommend this simplistic procedure. The fact that he originally tabulated only the upper 5% percentiles is not to be taken as a dogma (Fisher, 1925). We can discuss with the students that it is natural to sharpen the level of significance when the data set gets bigger: It would be stupid to keep a Type 1 error of 5 % with a big data set when the Type 2 error is negligible (this contradicts 3.). When there are few data, no one can expect to have low error rates. Perhaps Fishers alleged insistence on 5 % level of significance can be interpreted to mean that even with small data sets one should not make the Type 1 error probability too big. The reason for this is to avoid drawing conclusions from insufficient evidence. A consequence of this is that with small data sets non-rejection does not mean a lot: The alternative has not got a fair chance to prove itself!

4. The role of interesting and meaningful examples: How many excellent grades should be tolerated?

It is important to provide students with examples to help to build their knowledge. I offer the following example, which I think will appeal to both teachers and students.

The grade «excellent» is defined as consisting of the best 4 % of the students (of large numbers of students). The problem is how many excellent grades should be tolerated in a class of 25 before suspecting the teacher of being «soft».

This can be viewed as testing the null hypothesis $p = 0.04$ against the alternative $p>0.04$. Using the test statistic X - number of excellent grades in the class of 25 and rejecting (ridiculing the teacher of being «soft») when $X>3$ gives a significance level of 1.7 %. As very few teachers would dare to give more than perhaps two excellent grades in a class of 25, this shows that teachers are perhaps overreacting, thinking that the Law of Large Numbers apply even to small numbers like 25 (Kahnemann, Slovic &Tversky, 1982). It can be argued that a more realistic model than the binomial used would raise the critical value even higher, because of heterogeneity of the classes.

This is a demonstration of several issues raised above: When fixing the level of significance at 1.7 %, the power of the test is weak. If p = 0.1, that is, the teacher gives 10 % of the students excellent grades in the long run (this will surely make her popular), the power of the test X>3 is only 23 %, resulting in a Type 2 error probability of 77 %. This could be an argument to lower the critical value. Let us look at the test «Ridicule when X>2». The significance level of this test is 7.7 %. If we choose this method, 7.7 % of teachers who follow the correct grading standard will be ridiculed as being too soft! I think it is obvious that this practice is unsound. We have to conclude that with a material consisting of grading of a single class of 25 students it is not possible to find out a lot. Even so, it is important not to ridicule honest people because of random fluctuations!

Consider the situation of a teacher who has graded 10 classes of 25, that is 250 students. As above, we use the somewhat simplistic assumption that the sampling distribution of the test statistic Y - number of excellent grades among the 250 students is binomial with
n = 250 and p = 0.04 under the null hypothesis of fair grading. If we here seek a test with significance level near 1.7 %, we find that Y>16 has a significance level of 1.8 %. Is it reasonable to accept this error probability in this certain situation? Before a decision is made on this issue, one has to evaluate the possibility of Type 2 error. Similar to the calculations done above it is possible to find the power of the test «Ridicule when Y>16» in the alternative p = 0.1 as 97.7 %, making the Type 2 error 2.3 %. If we think the harm of ridiculing an innocent and letting a culprit with p = 0.1 go free are approximately equivalent, then the test «Ridicule when Y>16» is reasonable. My own opinion is that ridiculing is so severe that perhaps Y>17 with significance level 0.8 % would be a better choice, even at the expense of a Type 2 error when p = 0.1 of 8.5 %.

This is just an example to make the students understand the ideas of Type 1 and Type 2 errors, the relations between their probabilities and the relevance of sample size. Other interesting contexts could be found in quality control and medical diagnose.

5. Model-based statistical inference

To get students to avoid many of the pitfalls of statistical practice, I think it is important to stress that all statistical conclusions are relative to a model. Typically, you can never be absolutely sure that the model is correct, but it is possible to check some of the consequences of the model and so get an impression of the situation. A normal-plot will indicate whether you have serious non-normality. Residual-plots check for other model assumptions. Typically, a statistical computer package is a very big help when teaching model assessment. When using modern computer technology it is natural to use the different statistical approaches in an integrated way. A normal plot can be seen as an estimate of the underlying distribution. A normality test help to tell you whether a seemingly non-normal plot may be reasonably explained as randomness. If the data set is small, non-rejection may be due to insufficient power.

The use of computers in the teaching of statistics is receiving increasing attention from teachers and researchers, as is shown in IASE Round Table Conference on the impact of new technology in teaching and learning statistics (Garfield, & Burrill, 1977). Research suggests that computers may not only extend what statistical topics are taught, but may also affect how statistics is learnt, because technology provides students with powerful resources and multiple representations, which can help them to widen the meaning of statistical concepts.

REFERENCES

Batanero, C. (1997). Should we get rid of statistical testing? The significance test controversy. ISI Newsletter, 21,(2), 19.

Chow, L. S.(1996) Statistical significance. Rationale, validity and utility. London: Sage Publications.

Fisher, R. A. (1925) Statistical Methods for Research Workers. Edinburgh, Oliver & Boyd.

Fisher, R. A. (1956) Statistical Methods and Scientific Inference. Edinburgh: Oliver and Boyd.

Garfield, J. B. & Burrill, G. (Eds.) (1997) Research on the role of technology in teaching and learning statistics. Proceeding of the IASE 1996 Roundtable Conference.. Voorburg, The Netherlands: International Statistical Institute.

Heuch, I. (1998) Evaluation of textbooks in secondary schools (in Norwegian). Tilfeldig Gang. Newsletter of the Norwegian Statistical Society, 15,(1), 3.

Kahnemann, D., Slovic, P. & Tversky, A. (1997). Judgement under uncertainty: Heuristics and biases. New York: Cambridge University Press.

Morrison, D. E., & Henkel, R. E. (Eds.). (1970). The Significance test controversy. A reader. Chicago: Aldine.

Neyman, J. & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. Phil. Trans.A 231, 289-337

FRENCH RÉSUMÉ

Les tests statistiques ont été utilisés pendant plus d'un siècle. Toutefois, l'utilité de ces tests a été récemment remise en question. Il est important pour l'enseignement de la statistique de prévenir les étudiants sur la possibilité d'utiliser incorrectement les méthodes, y compris les moyens d'éviter le mauvais usage. Récemment, la statistique a été introduite dans le programme des écoles secondaires en Norvège, non sans controverses. Une critique qui a été faite concerne le manque d'applications intéressantes de la théorie. Cet article présente un exemple qui facilite l'aquisition des connaissances par les étudiants. La statistique basée sur des modèles, en utilisant la technologie moderne, peut aider les étudiants à éviter les difficultés relatives aux tests statistiques.