CASE STUDY

Teaching statistical appreciation in quantitative methods

Peter Mitchell, Department of Animal and Plant Sciences and Department for Lifelong Learning, University of Sheffield, Sheffield, U.K. P.L.Mitchell@Sheffield.ac.UK

Abstract

Statistical appreciation is defined here as the knowledge about statistical tests, how they are chosen, the procedure and interpretation of the results, but without the calculations of the test statistic. This was taught in modules on research skills to students taking part-time degrees at the University of Sheffield. Details of the content, teaching methods and assessment are given here, with stress on the correct understanding of P-values and interpretation of statistical significance. Given that more people need to understand the results and interpretation of statistical tests than to do the calculations, statistical appreciation is of general value, especially to research supervisors. It also provides a firm base for further learning and training in statistics.

Keywords: statistical appreciation, quantitative methods, statistical tests, P-value, statistical significance.

1. Introduction

For over ten years I have taught quantitative methods to students taking part-time degrees in a range of subjects, in a module initially called Developing Research Project Skills and later called Research Methods. Students in the Department for Lifelong Learning (DLL) at the University of Sheffield took this module at Level 2 (usually the third and fourth years of a part-time degree) in preparation for research carried out at Level 3. As is common in service teaching of statistics, the DLL had in mind initially that students would acquire all the knowledge and skills required to carry out whatever statistical analysis might be required in their research. In practice, it was soon realized that this ambition would have to be severely trimmed, for two reasons. First, only 16 hours of contact time were available for quantitative methods, the rest of the module being qualitative. Secondly, the starting points in mathematical ability of these students were varied but mostly low so basic numeracy had to be refreshed, and only a moderate rate of progress could be expected.

After some introductory material on numeracy, use of tables and graphs, and an outline of the research process, I concentrated on descriptive statistics and simple inferences using standard errors and confidence intervals. A firm grasp of these concepts would be useful in all fields of study. This left about four hours of contact time for statistical tests which some students might use in their research, and many would encounter in their reading in their subject. Clearly, there was not time to offer training in the use of a statistical package on the computer, nor to provide full information on t-tests, chi-squared tests, regression, and so on, nor to practise carrying out each test on data and interpreting the results. I decided that the best that could be done was to teach statistical appreciation. These particular part-time degrees ceased enrolling several years ago, with the teach-out period almost finished, so the purpose of this case study is to record what was done and to share more widely my experience of teaching the ideas behind statistical appreciation.

2. Definition

By statistical appreciation, I mean the knowledge *about* statistical tests, including the ability to recognize them and follow the procedure and interpretation, without being able to carry out the actual calculations. Appreciation here is used as in art or music appreciation: many people enjoy art and learning about art without being able to draw or sculpt, and cannot sing or play an

instrument but listen to music attentively and acquire much knowledge about it. For a long time, I used the phrase statistical awareness but this has been used by others with much wider meaning, e.g. Davies et al. (2012). Their view includes competence with calculations for statistical tests. On reflection, appreciation rather than awareness is more suitable, not least for the parallel with art or music appreciation.

I stumbled on statistical appreciation as a matter of expediency but now see that it has value in its own right. Calculation of test statistics is a stumbling block for many students. It is onerous by calculator, except for the smallest sets of data, and nowadays requires training in the use of a computer package. My impression is that students are often repelled by this experience unless there is sufficient time and progress can be gradual and linked with statistical understanding. Paul Wilson (personal communication, Sigma Network meeting, July 2016) pointed out that more people need to know about statistical tests, to understand the results and interpretation, than are actually required to make the calculations from the data. Of course, being able to do the calculations, even if only from the menus of a statistical package, is useful but not at the expense of a sound understanding of the procedure of statistical testing and how to interpret the results.

3. Prerequisites

Statistical testing makes use of probability, and there is much talk of P=0.05 or a one in twenty chance or 5% probability, etc. Consequently, an important prerequisite for statistical appreciation is the ability to understand probability on a scale from 0 to 1 and alternative means of expression, with the ability to convert from one to the other. Descriptive statistics were introduced with small sets of data obtained by students themselves and included the once-in-a-lifetime calculation of standard deviation (SD) by hand and calculator using the standard formula. This was checked using entry of data in statistical mode on a calculator, which brings in discussion of n or n–1 as a divisor, and hence samples and populations. Further calculations for standard errors (SE) and confidence intervals followed, including a brief presentation of the Normal distribution, and the use of a value of Student's t obtained from statistical tables.

4. Content of statistical appreciation

Allowing for complete beginners, some barely mastering descriptive statistics and confidence interval, I wrote a handout of 7500 words entitled Overview of Statistical Tests (Table 1). This included how to choose a statistical test, the procedure that all tests have in common, interpretation of probability associated with the test statistic (P-value) as a means of determining significance, use of statistical tables, statistical and practical significance, and effect size versus hypothesis testing. Outlines of five tests were given, plus descriptive statistics and simple inference (not a test as such but using the same mathematical foundation). The outline of correlation is given in Table 2 as an example.

Students were warned that the handout was a rather abstract explanation of statistical testing and that they should not expect to understand it immediately. It would need to be read several times, with cross-references followed when necessary. Once they had gained some understanding, during this module, they would be able to return to the handout when needing to refresh their memory because they had encountered a statistical test in their reading or because a particular test was required in their own research. This material would provide the basic information about the test except for how to calculate the test statistic. If this was wanted in their own research, then further help from textbooks, a statistician, colleagues or the help facility on a statistical package should be taken up. This handout is available from the MSOR Connections website (Mitchell, 2018).

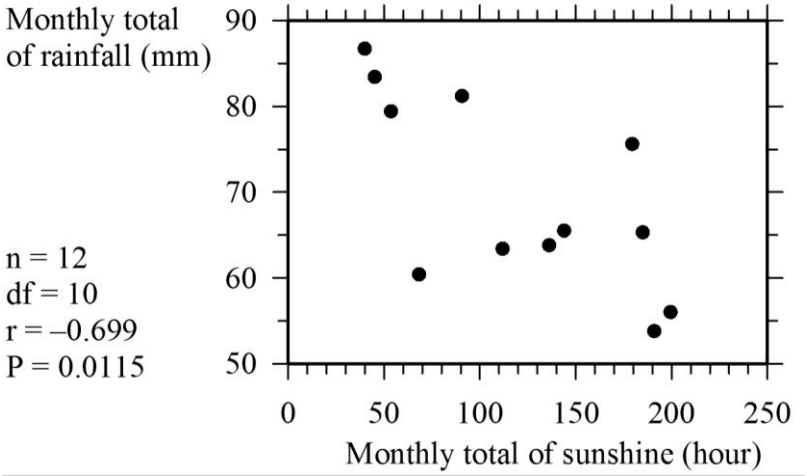
Table 1. Coverage of statistical appreciation in the handout Overview of Statistical Tests.

Heading	Summary
What statistical tests do	Distinguish interesting and uninteresting variation in data, i.e. signal from noise.
Scientific philosophy	The null hypothesis (in words only).
How statistical tests work	Choice depends on type of question and data. Standard key terms: test statistic, degrees of freedom, probability associated with the calculated test statistic, statistical significance. Significance as researcher's interpretation of the probability (P-value); standard thresholds (P=0.05, 0.01, 0.001) and words (significant, very significant, highly significant).
Degrees of freedom	Method of indicating how many independent pieces of information there are in the data.
Correct understanding of P-values	Definition of P-value; recognition that thresholds 0.05, 0.01, 0.001 are arbitrary; a way of converting continuous probability to categories of significance. Use of statistical tables and resources on the internet.
Worked example of using a statistical table	Working across columns to find range of probability for the value of the test statistic.
Proof and probability in statistical testing	You *cannot* prove anything with statistics. You *can* compute the probability of obtaining a set of results like this (or more extreme) if the null hypothesis was true.
Procedure	Consider question being asked, data to be obtained, which statistical test needed, how assumptions will be satisfied; gather data; obtain test statistic; interpret the probability for statistical significance; conclude about the difference or association of the initial question.
Choosing a test	Tests classified by purpose of test. 1. Describe the data and make simple inferences: descriptive statistics and confidence interval. 2. Look for association: classification of attributes (categories) needs chi-squared test; measurable attributes on a scatter graph needs correlation. 3. Compare means: two means needs t-test; two or more means needs analysis of variance. 4. Predict one quantity from another: regression.
A worked example	Mean weight of voles in two populations, using t-test. Key terms highlighted when they occurred in the text. Data contrived to provide a result that was not significant (P=0.126). Discussion of what not significant means, how larger samples would have produced a significant difference (variability held constant).
The conclusion of a statistical test	Clear statement of results always required; examples shown.
What "not significant" means	Say "a difference (or association) could not be detected"; dependence on variability and number of degrees of freedom.
Statistical and practical significance	Statistical significance as "is it likely to be true?"; practical significance as "is it worth taking account of, basing decisions on?".
Effect size versus hypothesis testing	Complementary but effect size provides direction of difference and its confidence interval; links to practical significance.
Outlines of statistical tests	One page each on descriptive statistics and simple inferences, chi-squared test, correlation, t-test, analysis of variance, and regression. Standard headings: purpose, data, null hypothesis, test statistic, degrees of freedom, assumptions, example (with statement of results), variations and elaborations, equivalent non-parametric test.

Table 2. The outline for correlation, from the handout Overview of Statistical Tests.

Pearson's Correlation

1. **Purpose.** To look for association (linear correlation) between two measured attributes.
2. **Data.** Two measurements from each case or item or subject, that can be plotted on a scatter graph (does not matter for the statistics which measurement is x). *Always* plot the graph, even if only a sketch.
3. **Null hypothesis.** No (straight-line) association between the two measurements; the points on the graph are a random cloud or occur in a horizontal or vertical line.
4. **Test statistic.** The correlation coefficient, r, measures the strength of association, between 0 for no association to –1 for perfect negative association, to +1 for perfect positive association. Test of significance is usually based on Student's t-test but tables of critical values of r are available for direct assessment.
5. **Degrees of freedom.** For sample of n cases, df = n–2. Note that there are n *cases* (or items or subjects), each case with two numbers.
6. **Assumptions.** Data from a random and independent sample of the population, where there is an underlying straight-line relationship between the measurements. The measurements are distributed bivariate Normally.
7. **Example.** Scatter graph of sunshine and rainfall recorded at Sheffield (monthly means for the period 1981–2010), and the correlation between them.



Statement of results. There is a moderately strong correlation (r = –0.699) between monthly totals of rainfall and sunshine; the correlation is significant (10 degrees of freedom, P = 0.0115). It is a negative correlation, i.e. high sunshine tends to be associated with low rainfall.

8. **Variations and elaborations.** None: correlation is a simple and fairly crude technique. If there is a statistically significant correlation then it may prompt further investigation and gathering of data, for example to predict one measurement from the other using regression.
9. **Equivalent non-parametric test.** There are two methods of correlation when the data are ordinal: Spearman's and Kendall's rank correlations.

5. Teaching and assessment

The learning outcomes are given in Table 3. After a brief explanation of statistical appreciation, i.e. that they would learn *about* statistical tests without any calculations, I found that it was best to plunge in with an accessible and intriguing example. The question was whether men or women spoke more words during the day, and Mehl *et al*. (2007) had collected suitable data and performed a t-test. I presented the summary data (means and SDs) and the result of the t-test. In

this case the result was not significant (P=0.50) despite a small difference in the mean numbers of words spoken, thus leading to an explanation about statistical significance and its interpretation. Having seen one statistical test in action, I then mentioned briefly the other tests and when they were used, and discussed whether statistics could prove anything. A second example test followed, chi-squared, applied to data about whether women who were handed a baby or a parcel held it to the left or right side of the body (Campbell 1989, p. 130). This too engaged students, who wanted to know more than I anticipated initially about calculation of expected values in the computations for the chi-squared test. After these two examples I presented the procedure for statistical testing, pointing out where the key terms (null hypothesis, test statistic, degrees of freedom, probability associated with the test statistic, statistical significance) had occurred in the previous examples. Two tasks in pairs or trios were undertaken: interpreting probabilities as statistical significance, and recognizing the key terms in statements of results from statistical tests.

Table 3. Learning outcomes for the two sessions (4 hours contact time) on statistical appreciation.

Students will
(a) be able to identify which test to use from the research question and type of data;
(b) recognize that there is a procedure in common for statistical tests;
(c) realize that statistics cannot prove anything but can quantify the uncertainty;
(d) be able to find statistical significance from the probability of a statistical test;
(e) be able to identify key terms used in statistical tests;
(f) be able to state what has been found from a statistical test;
(g) realize that assumptions must be satisfied for calculated probabilities to be reliable; and
(h) be able to calculate predicted values from a regression equation.

The second session started with a review of the procedure for statistical tests. Then examples of correlation, analysis of variance and regression were presented. Regression required much more time than the other tests. Few of these students were familiar with the equation for a straight line and interpreting the coefficients as a slope and an intercept. This had to be introduced for beginners, leading to the ability to calculate a predicted value (y) from any input value (x) supplied (used for work in class: each student computed a predicted value from the regression equation given a value for x, to assemble a set of results for the class). The final topic was the importance of assumptions in statistical testing, pointing out that these were stated in the outline of each test given in the handout (see Table 2 for an example). Omitted in class for want of time were statistical and practical significance, and effect size versus hypothesis testing; students were referred to the handout as a starting point, in case these topics occurred in their reading or research project.

Assessment was by coursework to avoid needless anxiety and unrealistic time constraints as in a test or examination. The single assignment of six questions to be completed individually over a period of three or four weeks comprised a mix of questions requiring numerical and narrative answers. Narrative answers were often merely a word or two, sometimes a sentence of explanation to test understanding of a concept. Some questions were scenarios or sets of data where the responses sought were which test would be suitable, what initial analysis of the data could be undertaken, but no computation of test statistic was required. Other questions presented the results of a test and asked for interpretation or simple calculations from the results (e.g. predicted values from a regression), or requested examination of the assumptions for the test. Assessment by coursework allowed students to work at whatever pace suited them and to consult any resources such as handouts and class notes, textbooks, and the internet. Despite this freedom, a wide spread of marks was always encountered.

6. Discussion

My aim was to produce students who could work out which test they needed for a research question and data, could pick out the key terms of the procedure, could state what a P-value meant and interpret it for statistical significance, and could answer the research question from the results of the test. If they acquired this knowledge then they could appreciate statistical tests in the literature, and do everything for their own use of a test except calculate the test statistic (which is arguably a trivial ability compared with the knowledge above). Statistical appreciation should provide a firm starting point for further learning on how to carry out the test and for training on a statistical package.

There are two implications for support services for students (not taking mathematics or statistics degrees) who require statistics as part of a project, and also for the service teaching of statistics.

1. If students have acquired statistical appreciation (i.e. correctly identified the test required, obtained suitable data, and can interpret the P-value and answer the research question), should it become routine for the support service to provide the test statistic?

2. Should statistical appreciation, as a minimum, be expected of any research supervisor? It requires nothing more than basic numeracy, and no training on statistical packages. Supervisors could then discuss with students use of statistical tests in the literature and their own project, and obtain further help, or plan further learning, for calculation of the test statistic when required.

Is there a role for the **sigma** Network in promoting statistical appreciation? There are no doubt weaknesses and imbalances in my coverage so I welcome further discussion, and am happy to share teaching materials.

7. Acknowledgements

Dr Verity Brack first recruited me to teach quantitative methods and was an inspiring colleague. I am grateful to the students who formed successive interesting classes. Dr Paul Wilson's talk "Statistics first, software later" at the **sigma** Network meeting, 1st July 2016, encouraged me to think further about statistical appreciation. Comments from a reviewer improved the manuscript.

8. References

Campbell, R.C., 1989. *Statistics for Biologists (third edition)*. Cambridge: Cambridge University Press.

Davies, N., Marriott, J. and Martignetti, D., 2012. A statistical awareness curriculum for STEM employees. *MSOR Connections*, 12, pp. 12–16.

Mehl, M.R., Vazire, S., Ramirez-Esparza, N., Slatcher, R.B. and Pennebaker, J.W., 2007. Are women really more talkative than men? *Science*, 317, p. 82.

Mitchell, P., 2018. Overview of the Statistical Tests, *MSOR Connections* 16(1). Available from: https://journals.gre.ac.uk/index.php/msor/index