Palestine Technical University Research Journal, 2018, special issue 6(2), 5-10
(http://www.ptuk.edu.ps)

Breaking the F-barrier: How the use of a visual representation of Fisher's F-ratio can aid student comprehension of orthodox statistics

Rory Allen *,

Goldsmiths University of London, UK

* Corresponding author r.allen@gold.ac.uk

Received 7/08/ 2018, Accepted 7/11/ 2018, Published 20/11/2018

Abstract: Universal laws are notoriously hard to discover in the social sciences, but there is one which can be stated with a fair degree of confidence: "all students hate statistics". Students in the social sciences often need to learn basic statistics as part of a research methods module, and anyone who has ever been responsible for teaching statistics to these students will soon discover that they find it to be the hardest and least popular part of any social science syllabus.

A typical problem for students is the use of Fisher's F-test as a significance test, which even in the simple case of a one-factor analysis of variance (ANOVA) presents difficulties. These are two in number. Firstly, the test is presented as a test of the null hypothesis, that is, that there is no effect of one variable (the independent variable, IV) on the other, dependent variable (DV). This highlights the opposite of what one generally wants to prove, the experimental hypothesis, which is usually that there is an effect of the IV on the DV. Students, if they think about the question at all, may be tempted to ask "why not try to prove the experimental hypothesis directly rather than using this back-to-front approach?"

Secondly, the F-ratio itself is presented in the form of an algebraic manipulation, involving the ratio of two mean sums of squares, and these means are themselves moderately complicated to understand. Even students specializing in mathematics often find algebra difficult, and to non-mathematicians this formula is simply baffling. Instructors do not usually make a serious attempt to remedy this confusion by attempting to explain what the F-ratio is attempting to measure, and when they do, the explanation is not usually very enlightening. Students may struggle with the statement that the F-ratio is the ratio of "two different estimates of the variance of the population being sampled from, under the null hypothesis". So what?

The result is that students frequently end up applying statistical analysis programs such as SPSS and R, without having the faintest understanding of how the mathematics works. They use the results in a mechanical way, according to a procedure learned by rote memory, and may overlook different tests which might be more appropriate for their data. This might be called the cookbook approach to data analysis, and it is the opposite of the ultimate aim of high quality teaching, which is to provide a deep understanding of principles, which will allow the student to use these principles flexibly in real life challenges, without violating the assumptions of the statistical tests being employed.

***Keywords*:** Fisher's F-test, ANOVA, SPSS, Neyman-Pearson statistics

INTRODUCTION

In attempting to make the F-test more comprehensible, I have developed a visual method of presenting the F-ratio, which motivates its use and in addition, provides a concrete realization of a fundamental philosophical principle behind all research methodology in science, namely Occam's Razor or the Principle of Parsimony. The full explanation of why it works is available at Allen (2018), but the aim of the present paper is to summarize the principles on which it works, and provide an incentive for instructors (and students) to adopt a different approach. The method is the outcome of teaching statistics for eight years to psychology masters students, during which time it evolved gradually, largely as a result of feedback and questions from those students. The first step was a realization that the F-ratio test can be seen in a natural way not as a test of null hypothesis on its own, but as a *comparison of two* hypotheses, namely the null and experimental hypotheses.

R. A. Fisher, who was the father of null hypothesis significance testing, maintained aggressively to the end of his life that his method worked by examining exclusively the null hypothesis. It is therefore ironic that his method can be better understood, in my opinion, in the context of comparison of two hypotheses. In fact, it turns out that even this is not quite correct: it is actually a *model* comparison test. And model comparison is the fundamental method used today in both Neyman-Pearson statistics and Bayesian techniques. Approaching the F-test via the model comparison route therefore prepares students mentally in case they ever need to move on to these two more recent developments.

The second step was to appreciate that the actual value of the F-ratio could be seen in terms of the ratio of the slopes of two straight lines in a fairly simple diagram. The diagram includes a third line, which I have named the Occam line in honor of the discoverer of Occam's principle, and which provides a quite specific example of the fundamental role played by this principle in the F-test itself.

HOW THE PROCEDURE WORKS

Taking a specific example, consider the following very simple set of data comprising an independent variable consisting of three groups, where the values of the dependent variable are 1, 2, 3 for the first group, 4, 5, 6 for the second group and 7, 8, 9 for the third group. The groups could represent three drug treatments, and the numbers, a measure of clinical outcome for each of nine participants. One might represent this set of data as a row vector thus: (1, 2, 3, 4, 5, 6, 7, 8, 9).

The first step with ANOVA is to calculate the so-called "total sum of squares" for these data, which is defined as the sum of squared deviations of the data points from the overall mean. Here, the mean is 5, and the sum of squared deviations from it is 16 + 9 + 4 + 1 + 0 + 1 + 4 + 9 + 16, or 60. This total is then partitioned into two quantities, the "within groups" and "between groups" sums of squares. The within groups sum of squares is found by taking the squared deviations within each group from the mean for that group, and adding these. In this instance each group contributes 2 to the sum, making a total over the three groups of 6. The between groups sum of squares is defined as what is left over when this sum of squares is subtracted from the total sum of squares, namely 54.

From these sums of squares, two "mean squares" are now calculated. The within groups mean square (MS_W) is found by dividing the within groups sum of squares by the within groups degrees of freedom, which is equal to the total number of data points reduced by the number of groups, or 6 with this dataset. The between groups mean square (MS_B) is found by dividing the between groups sum of squares by the between groups degrees of freedom, which is equal to the number of groups reduced by one, in this case 2. Finally, Fisher's F is found as the ratio $(MS_B)/(MS_W)$. The output of such a calculation for the example given above is shown in Table 1. I will ignore the "significance" value of .001 as it is not strictly relevant to the present discussion.

Table 1: Output of ANOVA calculation for the example

ANOVA

Score

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	54.000	2	27.000	27.000	.001
Within Groups	6.000	6	1.000		
Total	60.000	8			

THE PROCEDURE IN GENERAL

Consider the example of a one-way ANOVA, with the independent variable comprising k separate groups and having a total sample size of N. The procedure can be extended to multifactorial ANOVA, and indeed to repeated measures ANOVA, but to illustrate the basic principle this will suffice. A model is defined as

an approximation to the actual data, which involves assigning a value to each sample point, determined by the model. A standard measure of how far a model departs from the data, is given by the lack-of-fit sum of squares (which I abbreviate to lofsos); this is the sum of the squared differences between the actual value of the dependent variable and the value for that data point predicted by the model, taken over the whole sample.

The null hypothesis states that the groups are all drawn randomly from the same population. Corresponding to this hypothesis are a whole continuum of possible models, each consistent with the hypothesis. Each of these models approximates all the data points by a single number, which is called the parameter representing that model. It is well known that out of all such models, the one which fits the data most closely by the lofsos criterion is the model whose parameter is the mean of all the sample data: call it the null model.

In the case of the earlier example, the null model will approximate all the values of the dependent variable by the grand mean of 5. One could represent it as a row vector thus: (5, 5, 5, 5, 5, 5, 5, 5, 5). It can be seen by examining the definitions that the lofsos of the null model is identical to the “total sum of squares” as defined earlier.

Typically, a between-subjects design will be used to test a causal hypothesis, claiming an effect of the differing treatments represented by the various groups on the dependent variable. In its most basic form, the causal hypothesis is the logical contrary to the null hypothesis: it states that the population means from which the groups are sampled are not all equal. The causal hypothesis is, as with the null hypothesis, also compatible with many different models but as before, there is a unique causal model that best fits the data. That model is the one which approximates every data point by the mean of the group to which it belongs, (this group mean being the best estimate of the corresponding population mean).

In the previous example, the causal model will represent all members of each group by that group mean, which appears in row vector form as (2, 2, 2, 5, 5, 5, 8, 8, 8), having three parameters. The lofsos of the causal model is, from the definition, the same as the within group sum of squares. In the general case where there are k separate groups the causal model has k parameters, one for each group, each parameter being equal to its group mean.

I now have to introduce one final model: the saturated model, which approximates the dataset by itself. The saturated model can be represented by the same row vector as the original set of data: in the previous case, (1, 2, 3, 4, 5, 6, 7, 8, 9). Since each value of this vector is given by the data, there are in general N numbers required to specify the model: it has N parameters. The lofsos of the saturated model is evidently zero. The point of the saturated model will appear presently.

In Figure 1, I have plotted these three models derived from the example, with lofsos on the vertical axis and the number of parameters on the horizontal axis. The figure includes vertical lines indicating the size of the total sum of squares (the lofsos of the null model: 60), within groups sum of squares (the lofsos of the causal model: 6) and the between groups sum of squares (54), as well as the between groups degrees of freedom (2) and within groups degrees of freedom (6).

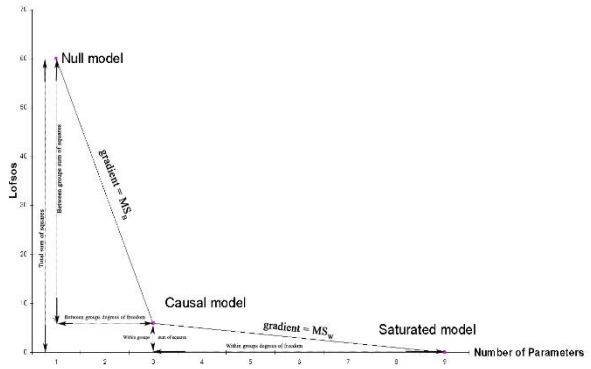

Figure 1: null, causal and saturated models plotted on a lofsos-parameter diagram, with mean squares, sums of squares and degrees of freedom indicated

Given the definitions of the mean squares as the ratio between the appropriate sum of squares to the appropriate degrees of freedom, it is clear that MS_B is the gradient of the line joining the null and causal models, and MS_W is the gradient of the line connecting the causal and saturated models (the point of the saturated model should now be clear: it was needed so that both these statistics could be represented on the same diagram). Fisher’s F-ratio appears as the ratio of these two gradients.

It is evident from this diagram that the causal model for our example lies below the line joining the null and saturated models. A moment’s thought will confirm that this will be the case when, and only when, the gradient MS_B is steeper than the gradient MS_W. This condition is clearly equivalent to the statement that $MS_B/MS_W > 1$. It follows that the plot of the causal

model lies below the line joining the null and the saturated models in the lofsos-parameter diagram if, and only if, Fisher's F is greater than one.

Figure 2 below gives the diagram for another dataset; this time I have included the Occam Line. The key point is that the point representing the causal model (corresponding to the experimental hypothesis) plots well below the Occam Line. Of course, a statistical test is needed to show if it is "far enough" below the line: this is provided by the F-ratio test. The F-ratio is the ratio of the slopes of (1) the line joining the null and causal models and (2) the line joining the causal and saturated models.

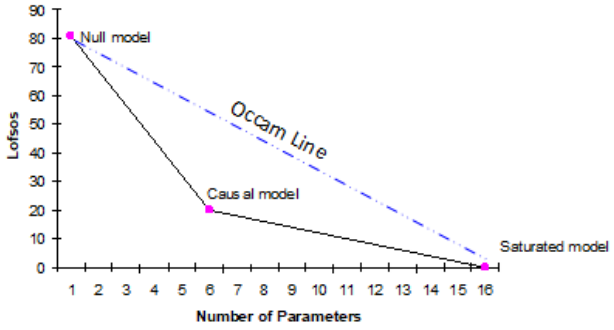

Figure 2: the lofsos-parameter diagram, showing the three models under comparison and the Occam Line

INFORMAL JUSTIFICATION

Why should this be "significant", in the non-statistical sense of the word? The null and saturated models are both lacking in interest, in terms of what they tell us about the data. The null model fails to distinguish in any way between the data points, and so does not tell us whether (or in what direction) any one of the group means differs from any of the others. The saturated model is equally unhelpful, but in the opposite direction. A model which uses the data to represent themselves has perfect fit, but at the expense of lacking any predictive validity.

This suggests that the line joining the null and saturated models might represent the point plots of all models which share, with the models at both extremities of the line, the property of being *without value in terms of conveying useful information* about the underlying structure of the data. In fact it can be shown that this line represents something quite concrete. Taking the example in the diagram in Figure 1, there are 9 sample points. Consider a model with three parameters. The line joining the null and saturated models has a slope of 60/8 or 7.5, so the point on this line vertically above the three parameter mark, which is two parameter units to the right of the plot of the null model, is at a vertical lofsos value of 60 – 2 x 7.5 or 45. Now suppose that I take all possible ways of dividing the original dataset into three groups, and for each such combination, I calculate the lofsos for that model, in which the data are approximated by the group means. Then the grand average of the lofsos values for all these combinations will be precisely 45.

This result is quite general (a proof is given in Allen, 2018). This means that the line joining the null and saturated models represents, for each value of parameter on the horizontal axis, a lofsos value that would be obtained on average by choosing appropriate numbers of subgroups of the dataset totally at random and calculating the corresponding models. Clearly, a prospective model should fit the data better than this – in other words, it should plot below this line – if it is to improve on the average performance of a model obtained in this random manner, and so to have any merit.

The null model-saturated model line slopes downwards to the right, meaning that the more complex models, with higher parameter values, have (as their complexity increases) a more severe threshold to overcome if they are to plot below this line, like the steadily dropping bar in a limbo-dancing contest. Complexity, measured by number of parameters, is penalized in a linear manner. The line therefore represents a numerical representation of Occam's razor. It might perhaps therefore fairly be dubbed the "Occam line" for this dataset.

The criterion that the causal model should lie below the Occam line on the lofsos-parameter diagram if it is to be preferred to the null model, is the same as specifying that the F-ratio for a dataset be greater than one, if the null hypothesis is to be rejected. This viewpoint shows why an F-ratio that is less than one is not of interest: this represents a model that lies above the line, and so fits the data worse than the null model once the penalty for complexity has been imposed. Clearly such a model is undesirable.

This does not of course suffice to show how the statistical distribution of the F-ratio is calculated in any given case: for that, one still has to use the statistical packages (or look it up in a book of statistical tables). But it does provide a logical foundation for an explanation of what the F-ratio is really doing. The presence of random error in the sampling of data from a population or populations means that the F-ratio must not only be greater than one, but significantly greater

than one for the causal hypothesis to be preferred, in order to limit the type I error rate.

Besides showing in a qualitative way why the F-ratio works, this approach has two further benefits. The concept of *degrees of freedom* is often hard to understand. In the present approach, it arises naturally. A degree of freedom is just the difference between two other numbers, namely the number of parameters in a pair of models. For example, the between groups degrees of freedom is the difference between the number of parameters in the causal model, and the number in the null model. The within groups degrees of freedom is the difference in parameter numbers between the causal and saturated models.

The second benefit is that an unbiased measure of effect size arises in a natural way from the diagram. It turns out that adjusted R-squared, or equivalently, epsilon-squared, is the obvious one to take when you look at the lofsos-parameter picture (see Allen, 2018 for details).

This approach has been applied more widely to explain the analysis of a range of statistical procedures based on the ANOVA method, in textbook format (Allen, 2017). This book demonstrates that the method is not simply a theoretical ideal with no real world application. It is hoped that this will introduce the method to a wider audience. Meanwhile the present paper may serve to alert teachers of statistics to a new view of the basics of the subject, which may be of value in their own practice.

REFERENCES

Allen R. (2017). *Statistics and Experimental Design for Psychologists: A model comparison approach.* London: World Scientific Publishing Europe Ltd.

Allen, R. (2018). Fisher's F-ratio illustrated graphically. *The Mathematical Gazette, 102*(553), 50-62.

