Fisher's F-ratio illustrated graphically

Introduction

Calculating Fisher's F-ratio is a key step in a number of statistical procedures involving null hypothesis significance testing. This is particularly so in the case of ANOVA (analysis of variance) in its several forms, but even multiple regression includes a test of significance of the overall model which employs an F-ratio. The present paper aims at making the basic ideas behind this common statistic more comprehensible by providing a visual counterpart to, and justification for, its algebraic definition.

As an example of how the definition works, 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 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 for the moment 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			

In teaching ANOVA to students, the calculation outlined above is usually justified as follows. On the assumption that the null hypothesis is true, the between groups and within groups mean squares will each be independent, unbiassed estimates of the same quantity, namely the variance of the underlying population. Both mean squares should therefore, if the null is true, yield roughly equal outcomes, so that their ratio – Fisher's F – is expected to be around one. If the F-ratio is much larger than one, we are therefore justified in rejecting the null hypothesis.

There is usually no discussion of why F-ratios of much larger than one may enable us to reject the null whereas ratios much smaller than one, which seem equally to contradict the null hypothesis, do not. Moreover, the logic of why the mean sums were chosen to compare estimates of population variance, rather than, say, one of the mean sums and the variance of the total dataset, is not explained. The student is to be forgiven if they conclude that statistics is a dark art.

It is, however, possible to represent the F-ratio visually on a two-dimensional diagram in a logical and straightforward way. There is a cost to be paid for this: to explain the procedure it is necessary to abandon a sole focus on the null hypothesis, and to include the experimental hypothesis explicitly. One has also to address the relationship between hypotheses and models. I suggest however that this increase in complexity is worthwhile because it brings with it an ability to perceive what the F-ratio is really doing. There are also unexpected fringe benefits. Effect sizes are emphasised in modern texts on orthodox statistics, but the usual explanation of the standard effect size measures used in ANOVA (eta-squared and its variants) is arcane. Using the diagram, two of them – eta-squared (or equivalently, R-squared) and epsilon-squared (which is identical to the so-called adjusted R-squared) – emerge in plain sight from the geometry of the situation.

Procedure

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 earlier 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. Incidentally, if N is the sample size, the data points and models in vector form can be considered as points in an N-dimensional Euclidean space, and the lofsos of a model is then simply the squared Euclidean distance between the model vector and the data vector.

At this point it is necessary 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).

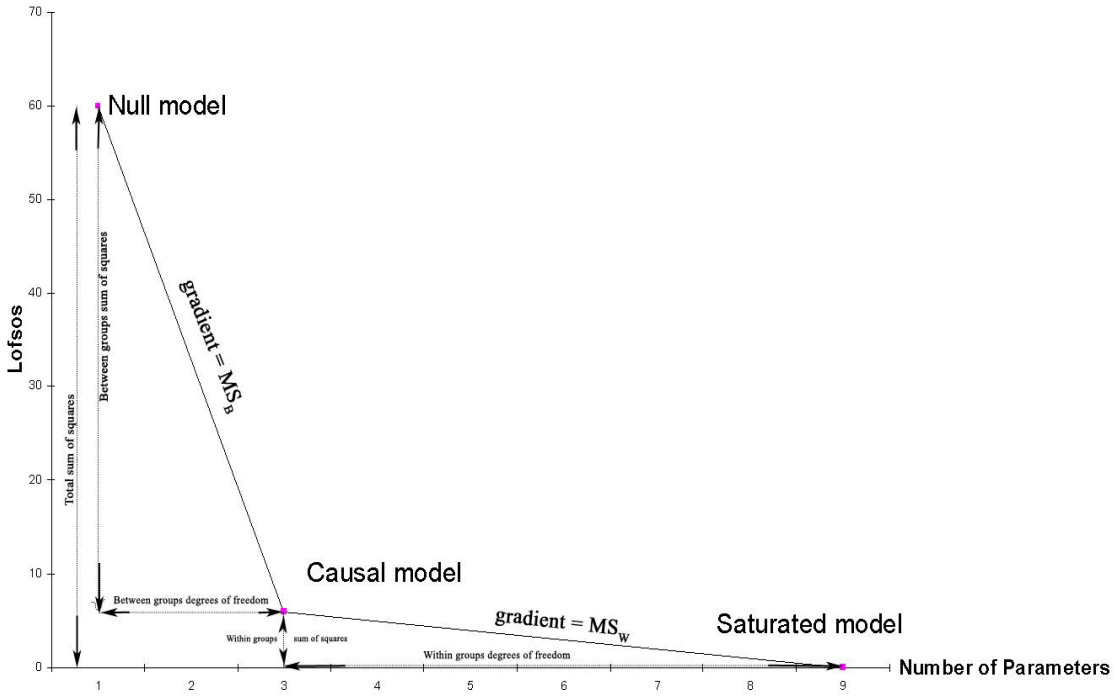

Figure 1: null, causal and saturated models plotted on a lofsos-parameter diagram, with mean squares, sums of squares and degrees of freedom indicated

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.

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 (see Figure 2: it is the line NS) 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 above, 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 at the end of this paper). 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 threshhold 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 significantly 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 distribution of the F-ratio is calculated in any given case. 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.

Incidentally, the notion that the Occam line is related to the set of all sets of subgroup partitions with a particular parameter number suggests that there is a natural distribution-free statistical test for significance, if the normality or other requirements for the standard distribution of the F-ratio are violated. If this happens, consider an F-ratio significant at (for example) the .05 level if and only if the F-ratio for the actual grouping of the data is within the top 5% of all F-ratios obtained by all different possible combinations of the data into the same number of subgroups, of the same sizes. This is evidently equivalent to specifying that the causal model actually obtained is closer (in terms of lofsos distance) to the data than at least 95% of the possible alternative combinations of the same number of subgroups and subgroup sizes. The method appears to ensure limitation of actual type I error rates to nominal alpha levels whatever the distribution of the dependent variable. In the case of our example, it happens that the lofsos of the actual grouping into three groups of three points each is the smallest possible among the 1680 possible such subgroups, meaning that the causal model can be preferred to the null model at a significance level of less than .001. The case for this method has been put with great eloquentce in Mewhort (2005) and Mewhort, Johns and Kelly (2010).

How effect sizes emerge naturally from the diagram

In what follows, to avoid confusion of letters I have referred to the model under test as the full model rather than the causal model.

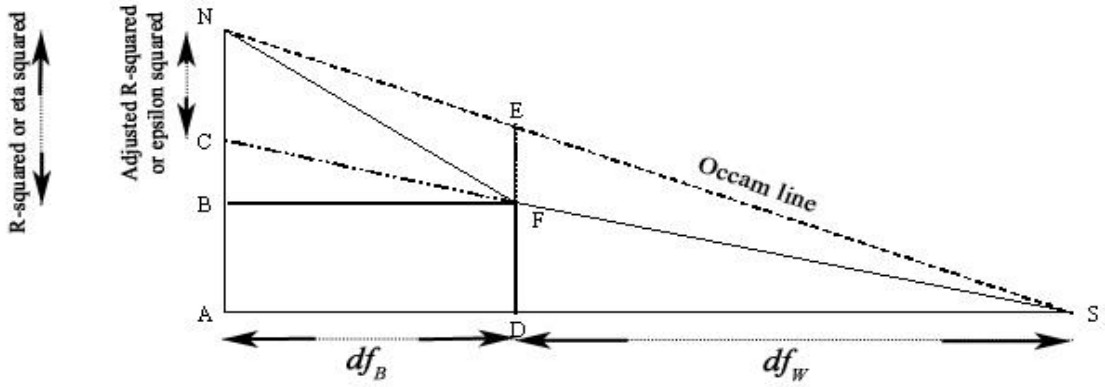

Figure 2: illustrating alternative effect sizes on a lofsos diagram. With AN normalized in length to equal 1, BN = R-squared and CN = adjusted R-squared

In figure 2 the points N, F and S represent the null, full (or causal) and saturated models on a lofsos-parameter diagram, and the line NS is the Occam line. Note that point A is at the point on the horizontal parameter axis corresponding to the number of parameters for the null model (usually 1, but in the case of a repeated measures ANOVA, it will be greater than 1). The lines AN and ED are vertical, and the line BF is horizontal, and parallel to AS. The line CF is an extension of the line FS, so that CFS is one straight line.

The lofsos of the null model is the distance AN, and the lofsos of the full model is the distance FD, which is equal to the distance AB. So the improvement in the lofsos given by the full model is given by AN – AB, or a distance BN. The relative improvement is now just BN/AN.

This measure of effect size has a name, or rather two names. In ANOVA-type analyses, it is usually called η^2 or eta-squared. In multiple regression, it is called R^2 or R-squared. But the two symbols refer to exactly the same thing, conceptually.

However, there is still a problem. To understand this, imagine that F was actually situated on the Occam line, vertically above its present position, at point E. Using the previous definition, the full model would still have a positive effect size, because point E has a smaller lofsos than N, even though it is doing no better than would be expected if the null hypothesis were true. In other words, this measure of effect size takes no account of Occam's principle. Surely, we would want any point on the Occam line to have an effect size of zero. This represents the well-known fact that eta-squared or R-squared suffers from a positive bias (as we will see below).

Looking at the diagram, there is, however, a natural definition of effect size for F which does not have this handicap. Surely the improvement in lofsos for F that we should be measuring is not the improvement relative to the null model, but that relative to the equivalent model with the same number of parameters as the full model. This equivalent model is located by definition on the Occam line, at position E. The lofsos for this model is the distance ED. The improvement achieved by F is now not NB, but FE, which is smaller. So the correct measure of effect size is EF/ED. This is precisely the effect size known as adjusted R-squared in regression calculations, or as epsilon-squared (ε^2) in the context of analysis of variance.

To see how this measure relates to the original R-squared measure of effect size, extend the line SF, until it hits the line AN at C. By similar triangles, EF/ED = NC/NA. The correct effect size is therefore NC/NA, compared with the original estimate of effect size, of NB/NA. Since NC is smaller than NB, the real effect size is smaller than the apparent one, confirming that R-squared was indeed biased positively.

To find a formula for the new effect size in terms of R-squared, proceed as follows. To simplify the calculations, I am going to take NA to be of unit length (this amounts to dividing all original lengths by the length of NA, leaving all ratios and therefore effect sizes unchanged).

If the length of AN is 1, then the length of NB must be R-squared (since NB/NA = R^2). So by subtraction, the length of AB is $1 - R^2$, and this is equal to the length of FD since ABFD is a rectangle (the equality of AB and FD is obvious from the diagram).

Since the triangles SDF and SAC are similar, the length of AC is AS/DS times the length of FD. You can see this as before by imagining FD projected from S onto AN. The length of the image is proportional to its distance from S.

But AS represents the difference between the number of parameters of the null model and the parameters of the saturated model, which is the definition of the degrees of freedom of the null model. By a similar argument, DS is the number of degrees of freedom of the full model, or equivalently the within groups degrees of freedom.

So with the obvious notation, the length of AC is $df_N / df_W.(1 - R^2)$, and the length of NC is $1 - \{df_N / df_W.(1 - R^2)\}$, which is the standard formula for adjusted R^2, sometimes called Wherry's formula.

It is possible for adjusted R-squared to turn out to be negative. This is the case if, and only if, the plot for the full model is above the Occam line. In that case, the model is clearly has no merit.

In ANOVA, the formula given for epsilon-squared is:

$$\varepsilon^2 = (SS_B - df_B.MS_W)/SS_T$$

where the first SS is the between sum of squares (corresponding to the distance NB in the diagram), the df is the "between" degrees of freedom, namely the parameter difference AD, the MS term is the mean within sum of squares (which is the slope of the line FS, ie the denominator in the F-ratio) and the final SS term is the total sum of squares, which is the distance AN.

It can be seen that this is equivalent to the formula for adjusted R-squared as follows. Since MS_W is the gradient of the line CF, and since df_B is the length BF, $df_B.MS_W$ is the vertical rise of CF over the distance BF, namely BC. So $(SS_B - df_B.MS_W)$ is simply the length NB – CB = NC, and the whole expression is NC/NA, which is identical to the previous formula for adjusted R-squared.

Incidentally, R-squared and adjusted R-squared are terms borrowed from the technical language of multiple regression, whereas eta-squared and epsilon-squared are terms for the same concept when used in the ANOVA context. The two areas of statistics were developed independently to analyse observational and experimental data respectively, and the consequent overlapping terminology has never been rationalized. Perhaps as a result, it does not seem to be widely known that these identities hold. Adjusted R-squared is in general use as a bias-corrected effect size measure in regression work. The most widely recommended measure for ANOVA is not, however, its equivalent, epsilon-squared, but a different statistic known as ω^2 or omega-squared. Omega-squared goes some way towards correcting the positive bias

of eta-squared, but the analysis above would suggest that it does not do so as well as epsilon-squared. There is evidence to support the claim that of the three measures, epsilon-squared is indeed the least biased (Albers & Lakens, submitted for publication; Okada, 2013).

Mathematical proof that the Occam line represents the lofsos plots of the average models for each parameter value

Suppose the dataset is of size N, and designate the lofsos of the null model by lofsos(null). The geometry of the Occam line is such that by well known principles of similar triangles, the vertical distance of the Occam line above the horizontal axis at the point corresponding to a parameter number of k is:

lofsos(null).{(N – k)/(N – 1)}.

Call this lofsos(k) for short. For k = 1, this of course gives a value of lofsos(null), as we would expect since the null model has one parameter, and for the saturated model with N = k, it reduces to the correct distance of zero.

This value has been claimed to represent in some sense the value of lofsos for a k-parameter model which is telling us nothing useful about the dataset. This is the meaning of the Occam penalty, which tells us to treat any model which plots on this line in the lofsos-parameter diagram, as equivalent to both the null and the saturated models. I will now justify this statement.

By the definition of lofsos, if the total dataset is represented as $\{x_i : 1 \leq i \leq N\}$,

$\text{lofsos(null)} = \Sigma_N(x_i - \bar{x})^2$,

where $\bar{x}$ is the overall mean, and any text on mathematical statistics shows that this can be simplified to

$\Sigma(x_i^2) - (\Sigma x_i)^2/N$.

Expanding the right hand term, this in turn reduces to

$\{(N - 1)/N\}.\Sigma(x_i^2) - \{2/N\}.\Sigma_{i \neq j}(x_i x_j)$

Therefore the height of the Occam line for parameter value k is

$[\{(N - 1)/N\}.\Sigma(x_i^2) - \{2/N\}.\Sigma_{i \neq j}(x_i x_j)].\{(N - k)/(N - 1)\}$

$= \{(N - k)/N\}.\Sigma(x_i^2) - \{2/N\}.\{(N - k)/(N - 1)\}.\Sigma_{i \neq j}(x_i x_j)$……(1)

Suppose now that I divide the actual dataset of N points into k disjoint subsets, of size $n_1, \ldots, n_k$.

the division into the k subsets can be written
$\{x_{11}, x_{12}, \ldots\}, \{x_{21}, x_{22}, \ldots\}, \ldots, \{x_{k1}, x_{k2}, \ldots\}$, where the size of the r-th subset is n_r.

What is the lofsos value for this particular division?

It is the sum of the values of the form $\Sigma_n(x_i - \bar{x})^2$, where n runs through the values $n_1, \ldots, n_k$, and the values of x_i for, say, the r-th subset runs through the n_r values $\{x_{r1}, x_{r2}, \ldots\}$ and $\bar{x} = \bar{x}_r$ is the mean of the r-th subset. I write the summand in the simplified form $\Sigma_n(x_i - \bar{x})^2$ to prevent a proliferation of subscripts.

Each such summand can be written, using the same reasoning as above, in the form

$\{(n - 1)/n\}.\Sigma(x_i^2) - \{2/n\}.\Sigma_{i \neq j}(x_i x_j)$, where now the summation is taken over only the values x_{ri} and x_{rj} in the r-th subset, omitting the r subscript for simplicity. So the total lofsos can be written in the form

$\Sigma[\{(n - 1)/n\}.\Sigma(x_i^2) - \{2/n\}.\Sigma_{i \neq j}(x_i x_j)]$ ……(2)

where the leftmost summation is taken over the subsets numbered from 1 through k, and the interior summations are taken *within* these successive subsets.

I will now show that when these expressions are taken over all the possible combinations into which the original set can be thus divided, the average value of all these expressions lies on the Occam line. By "all the possible combinations" I mean all distinct subdivisions of the original set of size N into subsets of sizes n_1 through n_k.

The main result I will need is that the result of averaging all possible expressions of the form (2) must be of the form

$\alpha.\Sigma(x_i^2) + \beta.\Sigma_{i \neq j}(x_i x_j)$ …….(3)

The summations in (3) are taken over all the values of $\{x_i : 1 \leq i \leq N\}$ in the original dataset.

A heuristic explanation for (3) is that when taking the average of the expressions (2) over all possible combinations of subsets, each datapoint is treated precisely equally: the symmetry of the situation means that the coefficient of each term x_i^2 must be equal to that for any other term x_j^2 for any values of i and j. Likewise, any pair (i, j) with $i \neq j$ must have the same coefficient in the final average for the term $(x_i x_j)$, as does any other pair (p, q) with $p \neq q$, for the term $(x_p x_q)$.

In more mathematical terms, consider the set of expressions of the form (2) over all the $N!/(n_1!n_2! \ldots n_k!)$ possible subdivisions of the original dataset. This set can be operated on by members of the complete group of permutations on N objects, the symmetric group S_N, in the obvious way, with elements of S_N permuting the subscripts in the algebraic

expressions. Since the set of expressions corresponds to the whole set of possible subdivisions of the dataset, the set is closed under the effect of S_N. This means that the sum of the algebraic expressions is invariant under the action of any member of S_N: any symmetric group element merely permutes the terms of the sum among themselves, leaving the sum itself unaltered. So the average of the terms, which is merely the sum of the terms divided by a constant, is also invariant. It is immediately obvious that the only expression in the terms x_i^2 and (x_ix_j) with $i \neq j$ that satisfies these criteria is the one where all the terms x_i^2 have the same coefficient as one another, and likewise all the terms (x_ix_j) with $i \neq j$ have the same coefficient, whence the form of (3) follows.

It may appear that we are not much further forward. But there is a very useful property of the original expressions of the form (2) which we can now exploit.

Consider one of the expressions of this form, corresponding to a particular subdivision of the dataset: it does not matter which one. Suppose we add the coefficients of the terms of the form x_i^2. I show that this sum does not depend on the subdivision. In the expression

$$\Sigma[\{(n - 1)/n\}.\Sigma(x_i^2) - \{2/n\}.\Sigma_{i \neq j}(x_ix_j)]$$

Consider any particular summand, which can be written

$$\{(n - 1)/n\}.\Sigma(x_i^2) - \{2/n\}.\Sigma_{i \neq j}(x_ix_j).$$

There are n terms of the form x_i^2, each with coefficient $(n - 1)/n$, so the sum of the coefficients is $n - 1$. This is the case for each of the k summands, so the overall sum of the coefficients of squared terms is

$$(n_1 - 1) + \ldots + (n_k - 1) = n_1 + \ldots + n_k - k = N - k.$$

Since this sum of coefficients does not depend on the particular subdivision chosen, the sum of these coefficients in the average taken over all possible subdivisions must also be $N - k$.

From (3), this sum is also equal to $\alpha.N$, so $\alpha.N = N - k$, and finally,

$$\alpha = (N - k)/N.$$

The value of β is found similarly. The sum of the coefficients for x_ix_j in

$\{(n - 1)/n\}.\Sigma(x_i^2) - \{2/n\}.\Sigma_{i \neq j}(x_ix_j)$ is $-(2/n).n.(n - 1)/2$, or $-(n - 1)$,

since there are $n.(n - 1)/2$ ways of choosing terms x_ix_j out of n possible terms, and each such term has the same coefficient $-2/n$. As before, summing this over all values of $n = n_i$ for this particular subdivision gives a total of $-(N - k)$.

The sum of the coefficients in the expression for the average over all subdivisions, expression (3), must therefore also be $-(N - k)$.

The second term in (3), namely $\beta.\Sigma_{i \neq j}(x_i x_j)$, is taken over all such pairs from the set $\{x_i : 1 \leq i \leq N\}$, of which there are $N(N - 1)/2$ terms, each with the same coefficient β. These coefficients sum to $\beta.N(N - 1)/2$ which I have shown is equal to $-(N - k)$.

Therefore $\beta = -2(N - k)/N(N - 1)$.

Substituting these values in (3), we have: the result of averaging all possible expressions of the form (2) must be of the form

$\{(N - k)/N\}.\Sigma(x_i^2) - 2\{(N - k)/N(N - 1)\}.\Sigma_{i \neq j}(x_i x_j)$ …….(4)
and this expression is seen to be identical to (1), which lies on the Occam line.

Therefore this point on the Occam line represents the lofsos that we would expect for a k-parameter model consisting of k groups, if the k groups were chosen entirely at random subject to the first one having n_1 points, the second one having n_2, and so on.

Clearly if this is the case, then any model which plots on the Occam line and which consists of k groups with n_1, etc points, is doing no better than if the points had been grouped entirely at random. Such a model is certainly not telling us anything useful about the data.

What about the general k-parameter model with k groups? Any such model must specify in advance the size of the subgroups, and therefore can be analyzed as above, with the values of n_i as given. Since the point on the Occam line depends only on the value of k and not on the sizes of the subgroups, the same analysis applies, and the same conclusion follows: the model must plot below the line in order to have any kind of merit.

References

Albers, C., & Lakens, D. (submitted for publication). Biased sample size estimates in *a priori* power analysis due to the choice of the effect size index and follow-up bias.

Mewhort, D. J. K. (2005). A comparison of the randomization test with the F test when error is skewed. *Behavior Research Methods, 37*(3), 426-435.

Mewhort, D. J. K., Johns, B. T., & Kelly, M. (2010). Applying the permutation test to factorial designs. *Behavior Research Methods, 42*(2), 366-372.

Okada, K. (2013). Is omega squared less biased? A comparison of three major effect size indices in one-way ANOVA. *Behaviormetrika, 40*(2), 129-147.