Fox Module 16 analysis of variance practice problems logits

(The attached PDF file has better formatting.)

** Exercise 16.1: One-way ANOVA

Three men (Abraham, Isaac, Jacob) and three women (Rebeccah, Leah, Rachel) receive the test scores:

Candidate	*Sex*	*Score*
Abraham	m	85%
Isaac	m	70%
Joseph	m	55%
Rebeccah	f	75%
Leah	f	80%
Rachel	f	65%

To test whether male and female candidates have significantly different scores, we use one-way ANOVA (analysis of variance), regressing test score on sex of the candidate.

A. What are the logits of these test scores?
B. What is the range of the test scores vs the range of the logits of the test scores?
C. What is the average logit for all candidates?
D. What are the average logits for male vs female candidates?
E. What is the total sum of squares (TSS)?
F. What is the regression sum of squares (RegSS)?
G. What is the residual sum of squares (RSS)?
H. How many degrees of freedom does the total sum of squares have?
I. How many degrees of freedom does the regression sum of squares have?
J. How many degrees of freedom does the residual sum of squares have?
K. What is the R^2 for the one-way ANOVA analysis?
L. What is the *F*-statistic for the one-way ANOVA analysis?

Candidate	*Sex*	*Score*	*Logit(scr)*	*Group*	*TSS*	*RegSS*	*RSS*
Abraham	m	85%	1.7346	0.9275	0.5678	0.0029	0.6514
Isaac	m	70%	0.8473	0.9275	0.0179	0.0029	0.0064
Joseph	m	55%	0.2007	0.9275	0.6090	0.0029	0.5283
Rebeccah	f	75%	1.0986	1.0346	0.0138	0.0029	0.0041
Leah	f	80%	1.3863	1.0346	0.1642	0.0029	0.1237
Rachel	f	65%	0.6190	1.0346	0.1311	0.0029	0.1727
		71.67%	0.9811		1.5038	0.0172	1.4866

Jacob: We test whether the expected test score differs by sex. Why do we call this analysis of variance, when we are comparing means? We don't compare the variances in the two groups.

Rachel: One answer is that the significance of a difference in means depends on the variances. Larger variances reduce the significance of the difference in means.

Illustration: Suppose the means test scores are 80% for men and 85% for women.

- If all men have scores between 78% and 82% and all women have scores between 83% and 87%, the difference in means is probably significant.
- If men have scores between 65% and 95% and women have scores between 70% and 100%, the difference in means is less likely to be significant.

Jacob: You say "one answer." Is there another explanation.

Rachel: The true explanation is that we are comparing two estimates of the variance of the error terms.

- The test works only if the variances of the two groups are the same.
- If the means of the two groups are also the same, the expected values of the two estimates of the variance are the same.
- If the means of the two groups differ, the expected values of the two estimates of the variance differ.

Jacob: You say "expected values of the estimates." Do you mean that "if the means of the two groups are the same, the expected values of the two estimates of the variance are the same"?

Rachel: No: the two estimates of the variance of the error term will differ whether or not the means of the two groups are the same. We examine the ratio of the two estimates.

- For very large samples, the ratio has a compact distribution centered on one.
- For very small samples, the ratio has a skewed, diffuse distribution, but its expected value is still one.

The distribution of this ratio is the F distribution, which depends on the two degrees of freedom. Graphs of this distribution are presented below.

Jacob: How does the number of candidates affect the analysis?

Rachel: The number of candidates affects the significance of the estimated variances. The variances are the residual sum of square divided by the degrees of freedom, which depend on the number of candidates in the group (sex in this example). With twice as many candidates, the variances are about half as large. To be precise: if the number of candidates in the group increases from N to 2N, the degrees of freedom increase from N-1 to 2N-1, and the variance decreases by a factor of (N-1)/(2N-1).

Jacob: The table above does not use the number of candidates in the computations. How does the number of candidates affect the solution?

Rachel: The number of candidates affects the computations two ways. Suppose this number doubles.

- The number of candidates affects the degrees of freedom in the denominator of the *F*-Ratio. If this number doubles, the *F*-Ratio doubles and becomes more significant. To be precise: the *F*-Ratio increases by a factor of (2N-1)/(N-1).
- The number of candidates affects the relative proportional of RSS vs RegSS. If this number is very small, most of the total sum of squares is in the RegSS, even if the true means are the same. As the number of candidates increases, the relation of RegSS and RSS better reflects the true difference in means.

Part A: The logit of a probability π is *ln*[π / (1 – π)]:

$$\text{logit } \pi = \ln\left(\frac{\pi}{1-\pi}\right).$$

Illustration: The logit of Abraham's test score is *ln*(85% / 15%) = 1.73460.

Part B: The range of test scores is 0% to 100% (0 to 1). The range of the logits is $-\infty$ to $+\infty$.

- The logits can be modeled with a normal distribution.
- The test scores themselves are better modeled with a binomial distribution.

Jacob: Does taking logits of the probabilities create a normal distribution?

Rachel: The distribution is not exactly normal, but it is close to normal. The probit transformation creates a normal distribution, and the probit and logit transformations are almost indistinguishable.

Jacob: Is a normal distribution needed for applying the *F* test or using analysis of variance?

Rachel: The *F* test is exact for a normal distribution; it is approximate for other distributions. For skewed, light-tailed, and heavy-tailed distributions, the *p* value derived from the *F* test may not be correct.

Jacob: How can we verify that the logit transformation makes the distribution more normal?

Rachel: Use a quantile comparison test.

- Simulate 10,000 values of a binomial distribution with a π of π_0 (any value between 0 and 1).
- Form logits of the values.
- Examine the qq-plots using the the original values and their logits.

R has built-in functions for simulating distributions and for quantile comparison tests. With Excel, you can

- Auto fill the figures 0, 0.01%, 0.02%, …, 99.99%, 100% in Column A.
- Use the binom.inv built-in function with a given π and the CDF in Column A for the binomial quantiles in Column B.
- Compute the standard deviation σ of the figures in Column B.
- Use the norm.inv built-in function with mean = π, standard deviation = σ, and the CDF in Column A for the normal quantiles in Column C.
- Form a scatterplot of the figures in Columns B and C.

Alternatively, the simulations and quantiles can be formed in VBA.

Jacob: Are you saying that the observed probabilities for the entire sample have a binomial distribution?

Rachel: The observed probabilities for candidates with the same true probabilities have a binomial distribution. If all men (or women) have the same expected test score, the observed test scores for all men (or women) have a binomial distribution.

The module of GLMs with logit link functions uses conditional binomial distributions, where the parameter π is fitted value of the response variable.

Part C: The average logit for all candidates is 0.981086.

Jacob: Why do we use logits of the test scores? Why not just compare the means?

- The mean test score for men is ⅓ × (85% + 70% + 55%) = 70.00%
- The mean test score for women is ⅓ × (75% + 80% + 65%) = 73.33%

Rachel: Test scores are probabilities. They do not have symmetric distributions, so the mean is not a good indicator of the expected test score.

- Test scores close to 50% have high standard errors. Joseph's test score of 55% may be distorted by random fluctuations; perhaps he had a bad day, but his true expected test score is 70%.
- Test scores close to 0% or 100% have low standard errors. Abraham's test score of 85% is not much distorted by random fluctuations. His true expected test score is likely to be 70%.

Jacob: How does this affect the logits?

Rachel: The distribution of the logits is close to normal, so the mean better reflects the expected score.

Jacob: Does the mean of the logits better indicate which group has higher expected test scores than the mean of the test scores themselves does?

Rachel: Suppose we know that all men are alike and all women are alike, except for random fluctuations in their test results, and the observed test results are

Candidate	*Sex*	*Score*	*Logit(scr)*	*Group*	*TSS*	*RegSS*	*RSS*
Abraham	m	99%	4.5951	2.0966	8.1435	0.1261	6.2427
Isaac	m	70%	0.8473	2.0966	0.7995	0.1261	1.5607
Joseph	m	70%	0.8473	2.0966	0.7995	0.1261	1.5607
Rebeccah	f	80%	1.3863	1.3863	0.1261	0.1261	0.0000
Leah	f	80%	1.3863	1.3863	0.1261	0.1261	0.0000
Rachel	f	80%	1.3863	1.3863	0.1261	0.1261	0.0000
		79.83%	1.7414		10.1209	0.7567	9.3641

The mean test scores are 80% for women and 79.667% for men. If the variances of the test scores are the same for men and women (despite the difference in this small sample), we might assume men and women have the same means.

But suppose candidates differ, and the best estimate of each candidate's expected test score is the observed test score. A test score of 70% has a variance of 70% × 30% = 21%; a test score of 80% has a variance of 80% × 20% =16%. With these large variances, the observed test score may vary considerably. An expected test score of 99% has a variance of 99% × 1% = 0.099%.

The logics convert the binomial distribution, which has low variances at both ends, into a distribution that is more like a normal distribution. The 99% test score with a low variance becomes a very high logit of 4.5951. Men have slightly lower test scores than women, but their mean logit is much higher than that of women (in this illustration).

Jacob: Your reasoning seems backwards. The variance depends on the expected probability, which we do not know. Perhaps men and women all have expected test scores of 80%, and the 70% and 99% are random fluctuations.

Rachel: You are correct; your comment underlies the difference between transformations and GLMs.

- Transformations adjust the observed response variables. The logit transformation converts the 99% test score to a value of 4.5951 and applies classical regression analysis to these logits.
- GLMs adjust the fitted values by the link function. Classical regression analysis has a closed form solution method: we compute α and β algebraically. No closed form solution method exists for GLMs, and iterative weighted least squares methods are used.

Logistic regression is similar to a GLM with a logit link function, but it is not the same. Before high-powered computers with GLM software, statisticians used logistic regression for dichotomous response variables. Now statisticians use GLMs with logit link functions.

Jacob: Why do we study logistic regression in this course?

Rachel: The module on logistic regression explains the concepts, and this practice problem shows how logit transformations affect regression analysis and ANOVA. Working though logistic regression and ANOVA using logits helps you understand GLMs with logit link functions.

Jacob: The logits in this exercise are all positive, ranging from 0.201 to 1.735, with no negative numbers. Is this reasonable for a normal distribution?

Rachel: A standard normal distribution has a mean of zero. The distribution here has with a mean of 0.981. Three candidates have logits below the mean and three have logits above the mean.

Jacob: Is the average logit the same as the logit of the average?

Rachel: No. For transformations that affect the skew of the distribution, the mean of the transformed values does not equal the transformed value of the mean.

Illustration: The average test score is 71.667%, whose logit is *ln*(71.667% / 28.333%) = 0.92799.

Part D: The average logit is 0.927523 for men and 1.034649 for women.

Jacob: Does this indicate that women have higher average test scores than men?

Rachel: The *F* test examines whether the observed difference in average test scores (or logits of test scores) is significant. The *F* test calculates the probability that the observed difference in average test scores (or logits of test scores) reflects random fluctuations, not true a difference in the expected test scores.

Jacob: Are you saying that the F-Ratio tests whether the groups have different means and variances?

Rachel: The *F* test assumes the variance is the same in all groups, whether or not the mean is the same, just as classical regression analysis assumes a constant variance of the error term for all points.

Jacob: Is the assumption of the same variance reasonable? If men and women have different means, won't they also have different variances?

Rachel: Your comment applies to all classical regression analyses, which assume the same variance for all data points, regardless of their fitted values. In truth, the variance is often a function of the fitted value. GLMs use variances that depend on the fitted values and the conditional distribution of the response variable, but they must solve for the regression parameters by maximum likelihood or iterated weighted least squares.

In many applications, the assumption of a constant variance is not too restrictive if the fitted values by group are not that different. For a binomial distribution, the variance is $\pi \times (1 - \pi) \times$ the number of observations. The difference in the variance between mean probabilities of 62% and 64% is small, and would not materially affect most analyses. The variances for π of 62% vs 64% are

- $62\% \times (1 - 62\%) = 0.236$
- $64\% \times (1 - 64\%) = 0.230$

The difference in variance for a two percentage point difference in the probability π is lowest at $\pi = 50\%$ and highest at $\pi = 0\%$ or 100%. The variance at $\pi = 97\%$ is almost three times the variance at $\pi = 99\%$.

Assuming a constant variance allows us to use classical regression analysis and analyses of variance. If the differences in variances are large, we use GLMs.

Jacob: Does the logit transformation correct this difference in variance?

Rachel: The logit transformation helps. The logit of 99% is much higher than the logit of 97%, which offsets the lower variance at 99%.

Intuition: A π of 64% does not differ much from π of 62%. They have large variances, so observed differences may stem from random fluctuation. A π of 99% differs much from π of 97%, since the variances are low. The logits differ greatly for these two probabilities.

F test*: null hypothesis and measures of variance*

Jacob: What is the intuition for the *F* test? What does the F-Ratio represent?

Rachel: The numerator and denominator of the F-Ratio are two ways of measuring variance. We test a null hypothesis vs the alternative hypothesis.

- The null hypothesis is that the two groups have the same mean.
- The alternative hypothesis is that the two groups do not have the same mean.

Both hypotheses assume the two groups have the same variance.

The ratio of the two estimates of the variance is the test of the null hypothesis.

- If the null hypothesis is true, the expected values of the two ways of measuring variance are the same, and the expected value of the F-Ratio is one.
- If the null hypothesis is false, the expected values of the two ways of measuring variance differ, and the expected value of the F-Ratio is greater than one.

Jacob: What is the null hypothesis?

Rachel: The null hypothesis is that the groups have the same expected values, and the observed differences reflect random fluctuations. The alternative hypothesis is that some (or all) expected values differ by group.

Intuition: We test whether the two groups have the same mean. To test this null hypothesis, we examine the ratio of the two estimates of the variance.

Jacob: What are the two ways of measuring variances?

Rachel: Consider first the denominator of the F-Ratio, which uses the residual sum of square (RSS).

We estimate σ^2_ε as RSS / degrees of freedom.

- We calculate the RSS for all candidates and divide by N - k - 1, where k is the number of parameters (or 1 on this Illustration: men vs women).
- We calculate the RSS separately for men and women.
  - Dividing the men's RSS by M-1 (where M is the number of men) gives one estimate of the variance.
  - Dividing the women's RSS by W-1 (W is the number of women) also estimates the variance.

Intuition: The *F* test assumes the variance is the same in all groups, so the estimated σ^2_ε (the variance of the error term) is the same for men, for women, and for both groups combined.

Jacob: The F-Ratio does not use separate estimates of the variance for men and women. Where is the computation of the variances for men and for women?

Rachel: The F-Ratio uses a weighted average of the estimates for men and women. In this illustration with men and women as the two groups, M + W = N. The weighted average of σ^2_ε is

$$\{ (M-1) \times RSS_M / (M-1) + (W-1) \times RSS_M / (W-1) \} \div ((M-1) + (W-1) \} = (RSS_M + RSS_W)/ (N-2)$$

Jacob: Is $RSS_M + RSS_W$ equal to the total RSS? Shouldn't we divide by N-1, not N-2?

Rachel: Adding the split between men and women adds one parameter, which slightly reduces the total RSS. The degrees of freedom are N-k-1 = N-2, so this weighted average equals the expected σ^2_ε.

Jacob: Why do we use M-1 and W-1 as the weights? Why not M and W (the number of men vs women)?

Rachel: The variance depends on the degree of freedom, not the number of data points.

Intuition: With 5 men and 1 woman, the computed RSS for women is always zero. It gets no weight, since it has no degrees of freedom. With 5 men and 3 women, but one other explanatory variable in each group (such as IQ score), the degrees of freedom are 4 and 2. The variances of the error term are reduced in each group by the other explanatory variable, and so are the degrees of freedom.

Jacob: How does the denominator of the F-Ratio differ if the alternative hypothesis is true?

Rachel: The derivation above does not assume that the expected values of men and women are the same. As long as the variances of men and women are the same, the *denominator of the F-Ratio gives an unbiased estimate of the population variance*.

Jacob: If the estimate of the variance does not depend on the hypothesis, how does the F-Ratio work?

Rachel: The numerator of the F-Ratio depends on the hypothesis. The numerator reasons:

Intuition: The variance of the mean is variance of each data point divided by N (the number of observations in the group). To estimate the variance of the error term, take the variance of the mean and multiply by N.

Jacob: Do we know the variance of the mean? If we don't know the variance of the error term, how can we know the variance of the mean?

Rachel: If the groups have the same means, then the variance of the observed means by group is a proxy for the variance of the mean.

Intuition: The variance of the mean is the variance of the computed mean if we observe many simulations.

In practice, we observe only one simulation (the actual observations), so we can't observe the variance of the mean. But if several groups have the same variance, the variance of the groups means is a proxy for the variance of each group mean.

Illustration: Suppose we want to know the variance of motor insurance claim frequency for male drivers. We have male drivers with last names sorted in alphabetic order. We divide 100,000 male drivers into 100 groups of 1,000 drivers based on alphabetic order. The variance of the sample of 100 group maens is a proxy for the variance of each group mean.

Jacob: Is this relation true if the groups have the same variance but different means?

Rachel: Suppose mean and women have the same variance of one but expected values of 200 and 300. We form groups of N = 100 men and N = 100 women.

- The expected group means are 200 and 300, and each has a variance of 1/N = 1%.
- The expected variance of the two group means is $(50^2 + 50^2) / 1$.

In general, we have *k+1* groups, where *k* is the degrees of freedom in the numerator (equal to the number of constraints or the number of explanatory variables).

Illustration: For men vs women, we observe two group means, and we have *k*=1 explanatory variable.

Intuition: The numerator of the *F*-ratio has *q* degrees of freedom, where *q* is the number of groups minus one, the number of constraints, or the number of dummy variables in the analysis of variance.

Testing the null hypothesis

Jacob: How does the F-Ratio test the null hypothesis of equal expected values?

Rachel: If the groups have different expected values, the variance of the observed group means is greater than the variance of each group mean. Larger differences of group means increase the numerator of the *F*-ratio but not the denominator, so the *F*-ratio itself increases.

Jacob: This implies that the *F*-ratio is always more than one, even if the null hypothesis is true (that the β's are zero). Suppose one forms small samples of two groups of normally distributed random variables with means of 0 and variances of 1. Will the *F*-ratios all be about one?

Rachel: The F distribution is positively skewed for small samples. Most randomly generated *F* values will be less than one, and few will be much greater than one. The overall mean will be one. See the graphs below.

Part E: The total sum of squares uses the square of (the observed value minus the overall average).

Illustration: For Abraham, this is $(1.734601 - 0.981086)^2 = 0.56778$

Part F: The regression sum of squares uses the square of (the group mean minus the overall average).

Illustration: For Abraham, this is $(0.981086 - 0.927523)^2 = 0.00287$

Part G: The residual sum of squares uses the square of (the observed value minus the group mean).

Illustration: For Abraham, this is $(1.734601 - 0.927523)^2 = 0.65137$

Part H: The total sum of squares has 6 data points minus one parameter (the mean) = 5 degrees of freedom.

Part I: The regression sum of squares has two groups. If a candidate is not male, she is female, so the groups have one relation. 2 data points minus one relation = 1 degree of freedom.

Jacob: What about the expected values for men vs women? Don't we have three parameters: the mean of the men, the mean of the women, and the relation that each person is either a man or a woman?

Rachel: The mean for the men is the observed value, not a parameter. Similarly, the mean for the women is an observed value, not a parameter. The parameter is the constraint that a person is either male or female.

Intuition: The RSS computes the variance from the individual data points with the mean as a parameter. The RegSS computes the variance from the group means, with the constraint among the groups as a parameter.

Part J: The degrees of freedom for the residual sum of squares in the denominator of the F-Ratio is the remaining degrees of freedom: 5 – 1 = 4. That is, the degrees of freedom for the RSS in a regression analysis with one parameter is N - k - 1 = 6 - 1 - 1 = 4.

Part K: The R^2 is the RegSS / TSS = 0.01721 / 1.503816 = 1.14%.

Part L: The *F*-statistic is (0.01721 / 1) / (1.486603 / 4) = 0.0463.

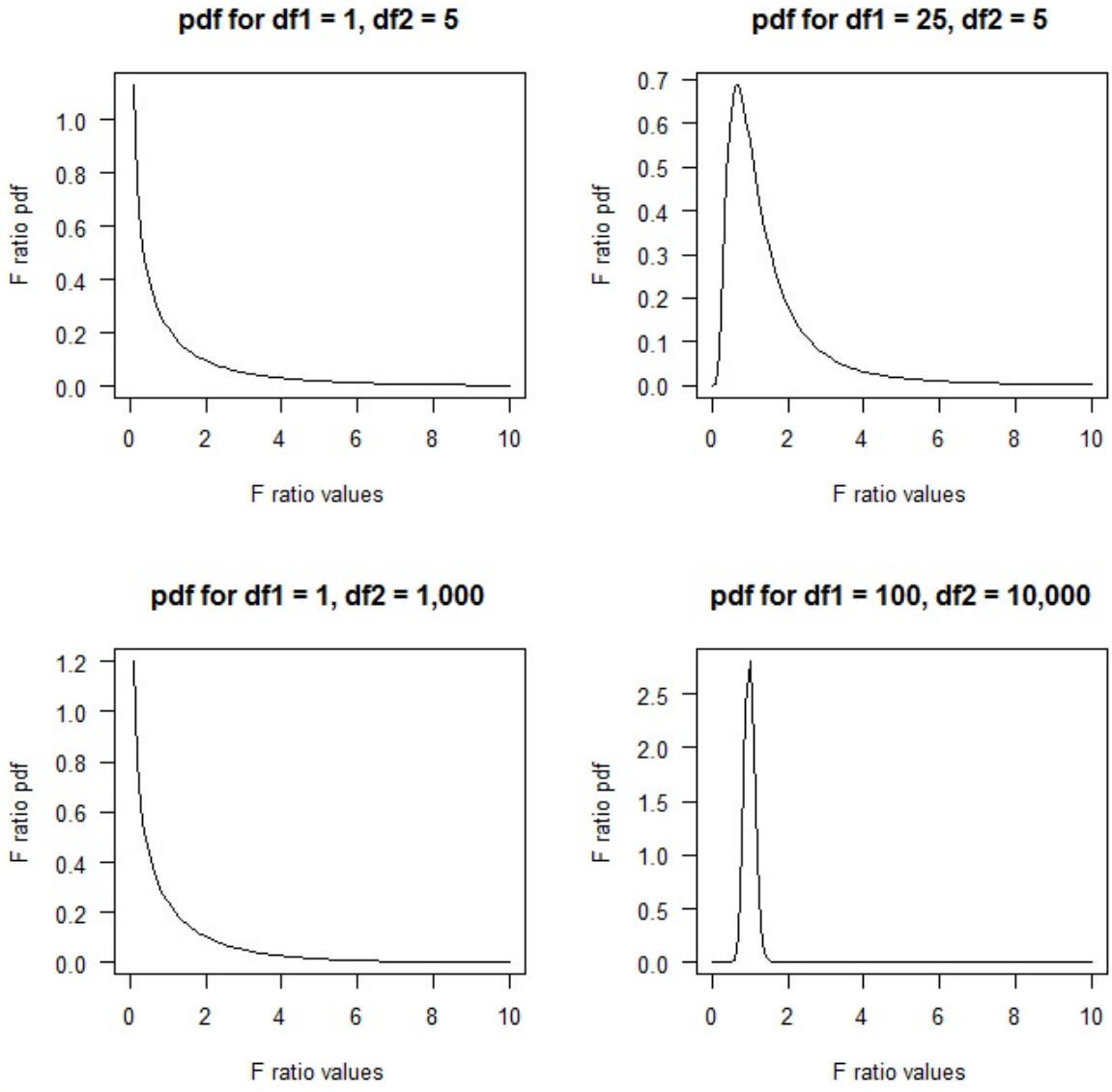
Jacob: The expected value of the F-Ratio under the null hypothesis of no difference between men and women is one, not zero. How do we get an *F*-statistic so close to zero?

Rachel: The F distribution is highly skewed for small samples under the null hypothesis. (The sentence above is poorly phrased. The F distribution *assumes* the null hypothesis, but we added the redundant phrase *under the null hypothesis* to make sure the intent is clear.) The null hypothesis gives many values close to zero and a few very large values. The expected value is one.

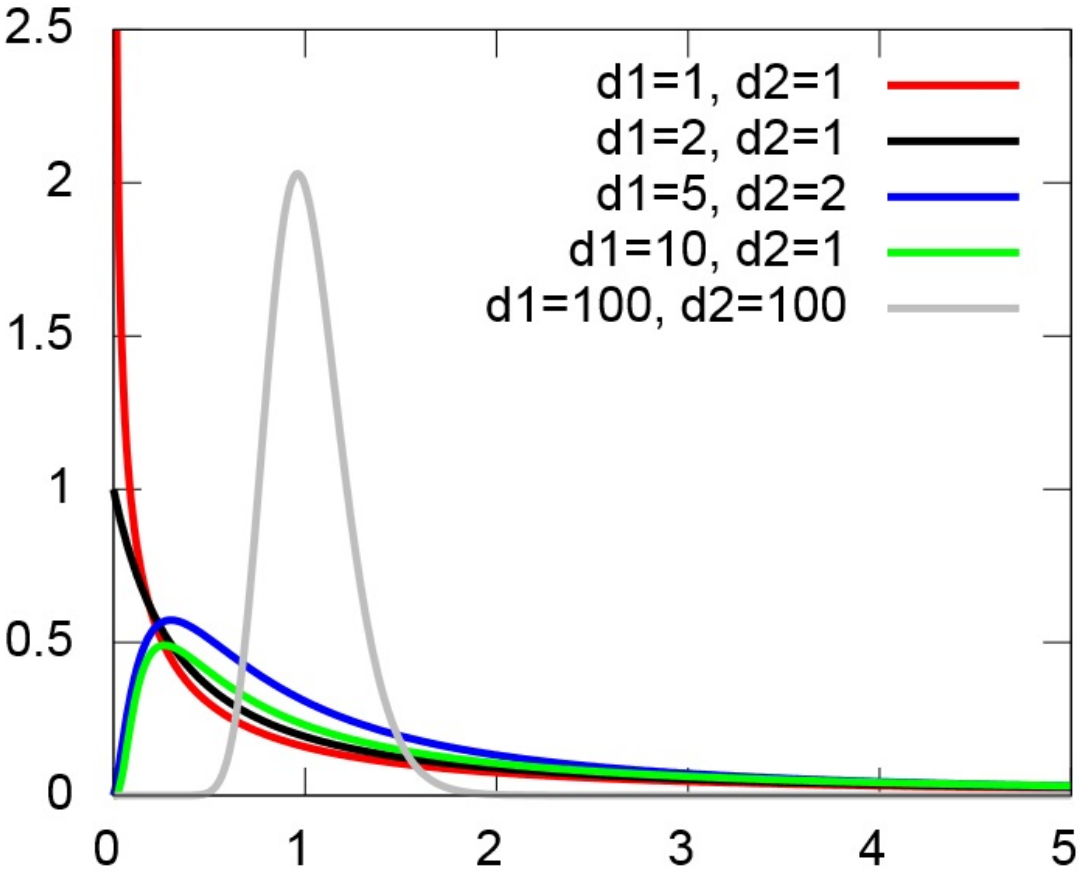
Part M: The scores of the six candidates range from 55% to 85%, with an average of 70% for men and 73.33% for women. If we took candidates at random and divided them arbitrarily into two groups of three, the average scores are likely to differ by random fluctuations. A small difference in a small sample does not indicate a statistically significant result.

The graphic below shows the probability density function for the *F* distribution with varying degrees of freedom.

- With one degree of freedom in the numerator (df1), the pdf is downward sloping (left hand graphics).
- As the degrees of freedom in the numerator increases, the pdf is compact and centered on x = one (right hand graphics).



The color graphic below from Wikipedia shows other combinations of degrees of freedom in the numerator and denominator of the *F* ratio.



{See Fox, Table 8.1}