JMP012: Backgammon

One Way Anova

Produced by

Dr. DeWayne Derryberry
Idaho State University, Department of Mathematics

Backgammon
One way ANOVA

Key ideas

One-Way ANOVA, testing assumptions, F ratio, R2, and basic computations

Background

A college professor played an online backgammon program eight games per day, during lunch periods, for several days in 1998. Each day he kept score, positive scores when he won, and negative scores when the program won. Each game can be a simple win for the professor (+1) or the program (-1), but it can also be a gammon (± 2) for an extremely one-sided win or loss, and even occasionally a blowout, or backgammon (± 3). Further, there is a doubling cube, which allows the stakes to be doubled and redoubled any number of times. At the end of the 50 days the professor tried to determine if he was better than the program.

The professor then returned in 2011 and played the software 60 more times (eight games/day) and 56 more times in 2012. The purpose of the 2011 games was to determine if the program had been upgraded, resulting in different daily scores. The professor continued to play the game in 2012, creating a third set of data.

The Task

Determine if the program is about the same in 2011 and 2012 as it was in 1998, or if there appear to have been upgrades at some point. Also, determine whether the professor is demonstrably superior in play to the program.

The Data Backgammon.jmp

The daily results were logged in a data table, with one column for each year.

1998 Score	Scores from the 1998 games (sample size = 50)
2011 Score	Scores from the 2011 games (sample size = 60)
2012 Score	Scores from the 2012 games (sample size = 56)

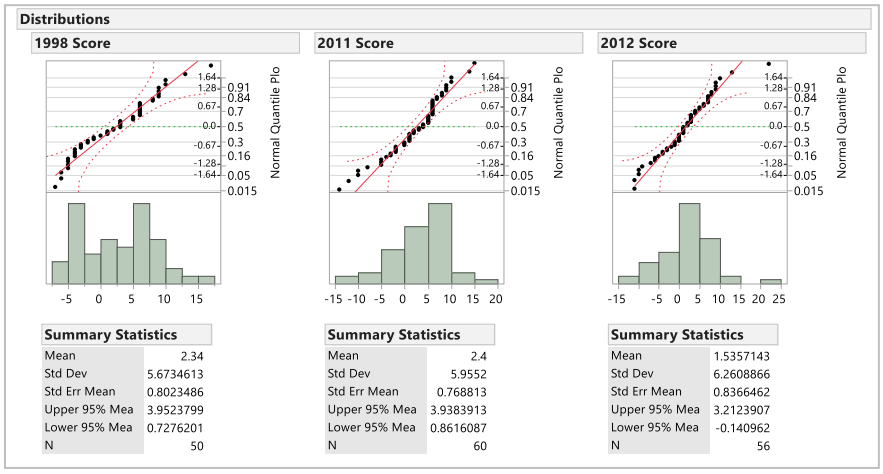
A positive score means the professor got the best of the program on that particular day, and a negative score means that the program won. There are zeros as well – occasionally, at the end of the day, the score was tied.

Analysis

Based on a preliminary look at this data over the three years, it appears that the professor is generally besting the program.

The entire confidence interval for the 1998 scores (0.73 to 3.95) is greater than zero, indicating that in the long-run, the professor will average an overall score of 0.73 to 3.95 better than the program (Exhibit 1).

Exhibit 1 Distribution of Scores



(Analyze > Distribution; enter all three columns as Y, Columns and click OK. Under the red triangle for each variable select Normal Quantile Plot.)

The similarity of scores for 1998 and 2011 is remarkable (Exhibit 1). The average scores and confidence intervals for the two periods are nearly identical. The professor had assumed there were upgrades to the program (i.e., that the program would perform better) in the period from 1998 to 2011. If there were upgrades, they don't appear to have been successful.

The consistency continues in 2012. However, the average score is slightly lower than in the previous years (1.534 versus 2.34 and 2.4) and the resulting confidence interval (-0.14 to 3.21), just barely, includes 0. We will be exploring whether these three samples are sufficiently similar that they can be combined, and if so, how the final confidence interval will compare to the three individual confidence intervals with regard to center and width.

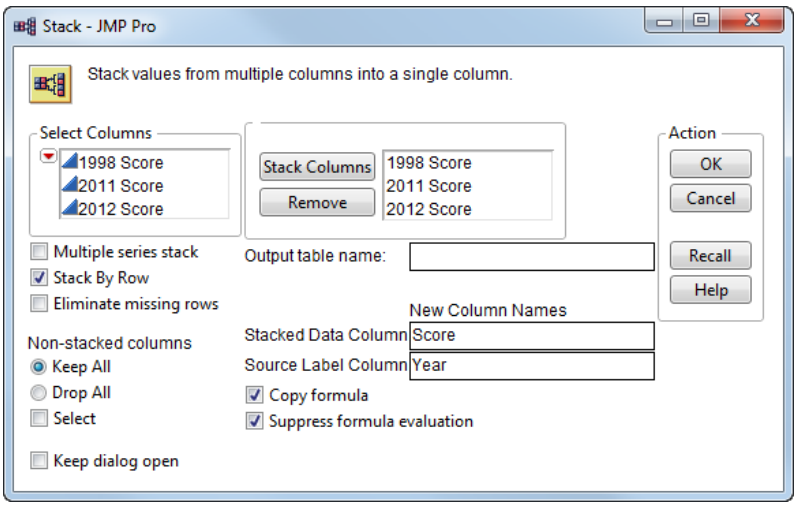
An important result of this initial peek is that the data has a great "shape" – there is no major skew in the data, and there don't appear to be any serious outliers. The standard testing procedures (t-tests and ANOVA) will work quite well here.

Based on our preliminary results, a more formal analysis should show two things: The program is consistent from year to year, and the professor is clearly better than the program.

Formatting the Data for Analysis

Our first analysis will involve comparing average scores across the three years. The data table Backgammon.jmp has three separate columns, where each column contains the scores for each of the years. A formal statistical analysis in JMP will require stacked data. That is, the labels (**Year**) should be stacked in one column, with the values (**Score**) stacked in a separate column (Exhibit 2).

Exhibit 2 Stacking the Three Score Columns



*(Tables > Stack; enter the three score columns under Stack Columns. Change the Stacked Data Column name to **Score** and the Source Label Column to **Year**, and click OK. Note: Since each year has a different number of observations, the resulting table will have some missing values. These rows can be deleted.)*

The stacked data are also found in BackgammonStacked.jmp.

Formal Analysis – Step One, Checking Assumptions

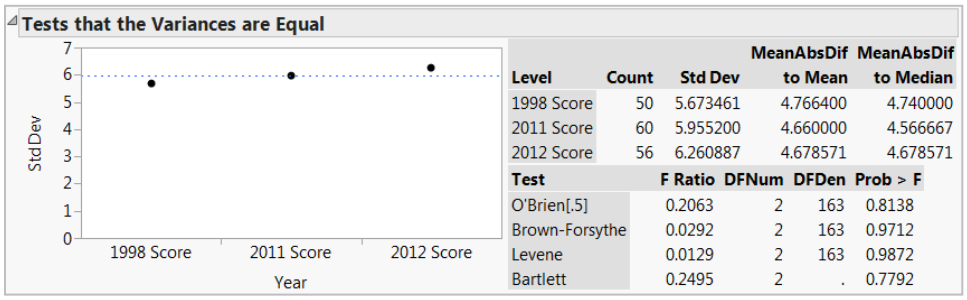
Before performing a formal test, we must check the following assumptions (as much as possible):

- The data is roughly normal (it already looks great from previous plots, and the normal quantile plots in Exhibit 1 provide a more visual confirmation).
- We have independent observations (this will require a separate discussion below).
- The sample is a simple random sample, or can be treated as one (if we hope to generalize beyond this sample to some population).
- Finally, the spreads are roughly equal from year to year. In fact, the standard deviations of all of the years look very similar (Exhibit 1). However, a test of equal spread cannot hurt.

Four tests for equal variances are conducted (Exhibit 3). For all of these tests, the hypotheses are:

- Ho: The spreads are equal.
- Ha: At least one group has a different spread.

Exhibit 3 Testing for Equal Spread



Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
1998 Score	50	5.673461	4.766400	4.740000
2011 Score	60	5.955200	4.660000	4.566667
2012 Score	56	6.260887	4.678571	4.678571

Test	F Ratio	DFNum	DFDen	Prob > F
O'Brien[.5]	0.2063	2	163	0.8138
Brown-Forsythe	0.0292	2	163	0.9712
Levene	0.0129	2	163	0.9872
Bartlett	0.2495	2	.	0.7792

*(Analyze > Fit Y by X; Select **Score** as Y, Response and **Year** as X, Factor and click OK. Then, select Unequal Variances from the red triangle.)*

In this case, all of the tests produce large *p*-values, showing no evidence of unequal spread. As usual, we use the awkward double negative language to avoid affirming the null hypothesis.

> "There is no evidence of unequal spread." ≠ "The spreads are equal."

When we have a large *p*-value (in this context) the first statement is correct; the second is incorrect.

In cases where the *p*-values would lead to different conclusions, the preferred test is Levene's test (see The Statistical Sleuth, 2nd edition, Ramsey and Schafer, 2002).

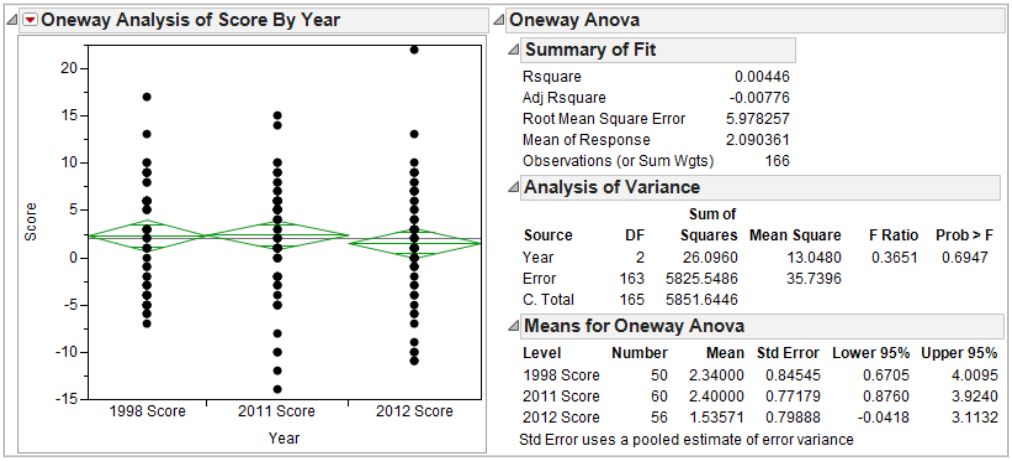
What about the assumption of a simple random sample? To begin with, let's be clear, we do not have a simple random sample. In fact, for a *true* simple random sample you must have a population in the form of a list, and you must choose elements from the population using a pseudo-random number table or a similar method. The population here is hypothetical at best – perhaps all possible sequences of eight games the professor and program could have played.

Part of the notion of a simple random sample is that it should be representative of the population, the second part is that observations should be independent. In this context there is no reason to believe dependence occurred. So we will act as if we have a simple random sample with independent observations.

Formal Analysis – Are There Differences From Year to Year?

We are very comfortable that the assumptions have been met – now to the analysis. Since we have three years of data, the formal analysis is a one-way analysis of variance, or ANOVA (Exhibit 4).

Exhibit 4 ANOVA



(In the Oneway output window, select Means/ANOVA from the red triangle.)

The hypotheses being tested in the ANOVA table are:

Ho: All three years have the same mean.
Ha: At least one year has a different mean.

The test statistic, the F ratio, is 0.3651, resulting in a *p*-value of 0.6947 (under Prob > F in Exhibit 4). There is no evidence of a difference in average score from year to year.

The ANOVA procedure produces a pooled estimate of variance, and the resulting standard deviation is reported as the *Root Mean Square Error* in the Summary of Fit table. Because the standard deviation has been modified, and because this estimate of standard deviation is based on 163 degrees of freedom (50 -1 + 60 – 1 + 56 – 1), the confidence intervals produced in Exhibit 4 are slightly different than the ones produced in the preliminary analysis.

So, our first question has been answered. We have no reason to believe that the backgammon program was upgraded over the years – the performance of the program does not appear to have changed. (We're assuming that the professor's skill level has also not changed.)

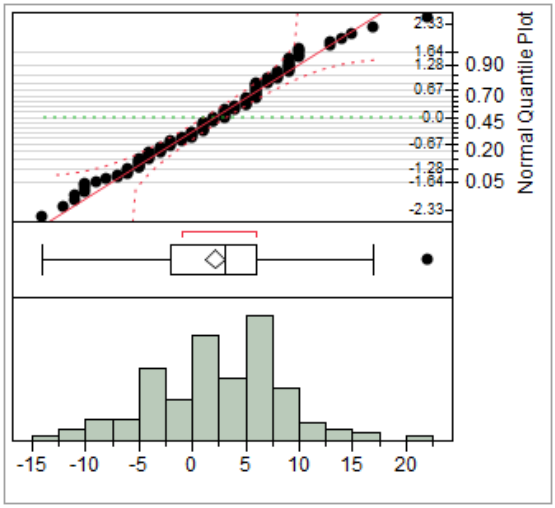
Our next task is to determine if the professor is better than the backgammon program. If the program had changed over time, we would need to estimate the professor's advantage in each time period. But, since the program has not changed, we will throw all of the data together to estimate one interval and perform one test.

Is the Professor Better than the Program?

When small samples of normal data are combined, the resulting samples tend to look even more normal, losing the esoteric patterns common to the original samples (that is, assuming the small samples are all drawn from the same underlying normal population).

This may explain why the combined score data looks so much more normal than the three separate sets of data (Exhibit 5).

Exhibit 5 Distribution of the Combined Score Data



*(Analyze > Distribution; Select **Score** as Y, Columns and click OK. Under the red triangle next to Score select Normal Quantile Plot.)*

The one-sample t-test provides overwhelming evidence (Exhibit 6) that the professor will have a long-run positive score against the program (*p*-value < 0.0001).

Exhibit 6 One-Sample t-Test for the Combined Score

Summary Statistics

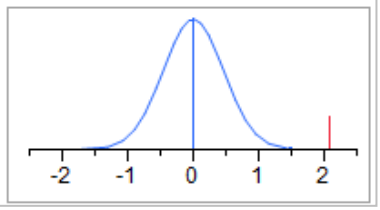
Mean	2.0903614
Std Dev	5.9552089
Std Err Mean	0.4622138
Upper 95% Mean	3.0029776
Lower 95% Mean	1.1777453
N	166

Test Mean

Hypothesized Value	0
Actual Estimate	2.09036
DF	165
Std Dev	5.95521

t Test

Test Statistic	4.5225
Prob > |t|	<.0001*
Prob > t	<.0001*
Prob < t	1.0000



(From the Distribution output window, select Test Mean from the red triangle next to Score, and click OK in the Test Mean dialog window.)

The overall edge appears to be from 1.18 to 3.00 points per day (i.e., per eight-game session). This interval is naturally narrower than the three individual intervals because it is based on roughly triple the sample size.

ANOVA Table Computations

There is a lot to the ANOVA output in Exhibit 4, and making sense of it requires understanding some of what is happening behind the scenes.

ANOVA is a comparison of variances (recall that ANOVA stands for *analysis of variance*). In this analysis, we wanted to know if the backgammon program had changed over the years.

Recall our hypotheses for this test:

Ho: All three years have the same mean.
Ha: At least one year has a different mean.

One way to think about this is to consider whether one mean, or three means, best fits the data:

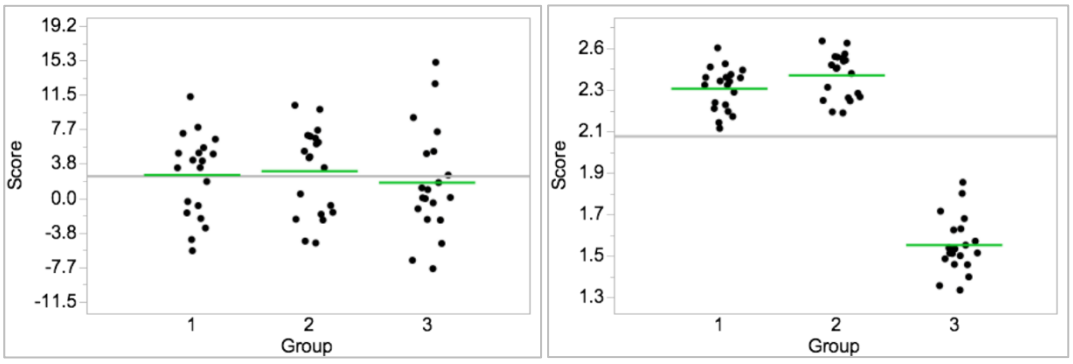
- If one mean fits the data well – this is consistent with no change to the program.
- If we must fit more than one mean – this is consistent with a change in performance of the program. (Remember that we're assuming that the professor's ability has remained constant.)

When we fit one mean to the data we get a sample variance, s^2. This is the overall variance of all of the data. When we fit three means to the data, we get three sample variances – each variance is a measure of the variation *within* a particular group. These within group variances can be pooled to produce the pooled sample variance, s_p^2. This quantity, which is a measure of the residual or random error variation, is called the *Mean Square Error*, or *MSE*.

We also have a measure of the variance *between* the means, referred to as *Mean Square Groups*, or *MSG*. This is often referred to as Mean Square Model (MSM) or Mean Square Treatment (MST). If the null hypothesis is true (the means are equal), then these two quantities, MSG and MSE, will be roughly equal. That is, the variation between groups will be about the same as the random variation within groups. If the null hypothesis is not true, then MSG will be larger than MSE - the variation between the groups will be larger than the random variation.

Two examples are shown in Exhibit 7. In the graph on the left we see that the differences in group means are small relative to the random variation within groups. In the graph on the right, the differences between the group means appear large in comparison to the random variation.

Exhibit 7 Backgammon ANOVA Table



So, how do we formally test the null hypothesis? We compare MSG and MSE by taking the ratio (MSG/MSE). This produces a test statistic, the F Ratio, which follows the F Distribution. We use a *p*-value to represent how extreme our resulting F Ratio is with regard to the F Distribution (for the given degrees of freedom). As with all *p*-values, a small value (generally < 0.05) provides evidence against the null hypothesis.

Here's what we can expect to see:

- When the null hypothesis is true, the between group variance (MSG) ≈ within group variance (MSE). The F Ratio will be small (near 1) and the *p*-value will be large.

- When the alternative hypothesis is true, the between group variance (MSG) >> the within group variance (MSE). The F Ratio will be large (greater than 1) and the *p*-value will be small (< 0.05).

These results are summarized in an ANOVA table. The ANOVA table for our backgammon example is shown in Exhibit 4, and again in Exhibit 8. Mean Square Groups (Mean Square Years in this example) is 13.0480, and Mean Square Error is 35.7396.

The ratio of these two quantites, the F Ratio, is 0.3651, and the resulting *p*-value is 0.6947. As we discussed earlier, this high *p*-value indicates that there is no evidence that the average yearly scores are different.

Exhibit 8 Backgammon ANOVA Table

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Year	2	26.0960	13.0480	0.3651	0.6947
Error	163	5825.5486	35.7396		
C. Total	165	5851.6446			

Completing an ANOVA Table

Let's take a closer look at how the F Ratio and other values of an ANOVA table are computed. We'll use an example in which the data suggest that the alternative hypothesis is true.

Suppose s = 1.21 and s_p = 0.87, k = 4 (there are four groups) and n = 61 (the total sample size is 61). This is all of the information required to complete an ANOVA table.

The ANOVA table:

Source	DF	Sums of squares	Mean squares	F ratio
Groups	k -1	SSG = SST - SSE		
Error	n - k	SSE		
Total	n - 1	SST		

We use s and s_p to calculate SST (Sum of Squares Total) and SSE (Sum of Squares Error).

$SST = (n-1)s^2$ and $SSE = (n-k)s_p^2$

In this case:

k - 1 = 3, n - k = 57, and n - 1=60

$SST = (61-1)(1.21)^2 = 87.846$ and $SSE = (61-4)(0.87)^2 = 43.143$

So, Sum of Squares Groups, or SSG is:

$SSG = 87.846 - 44.703 = 43.143$

When we plug these values into the ANOVA table, we get:

Source	DF	Sums of squares	Mean squares	F ratio
Groups	3	44.703		
Error	57	43.143		
Total	60	87.846		

The Mean Squares and the F Ratio can now be computed as follows:

Source	DF	Sums of squares	Mean squares	F ratio
Groups	a	A	MSG = A/a	MSG/MSE
Error	b	B	MSE = B/b	
Total	a+b	A + B		

So we have…

$Mean\ Square\ Groups\ (MSG) = \frac{44.703}{3} = 14.902$, and $Mean\ Square\ Error\ (MSE) = \frac{43.143}{57} = 0.757$

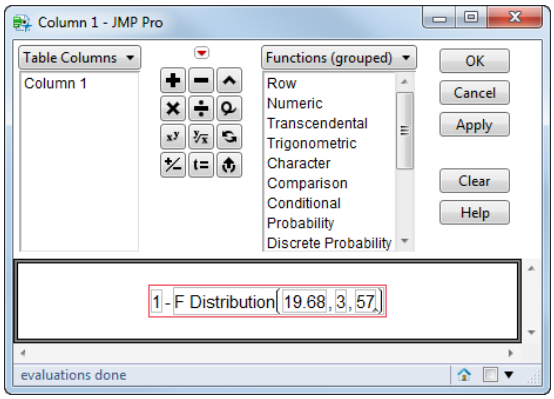
The F Ratio, then, is the ratio of MSG and MSE: $F\ ratio = \frac{14.902}{0.757} = 19.68$

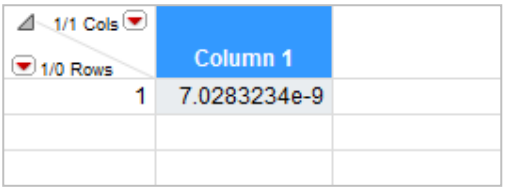
The completed ANOVA table:

Source	DF	Sums of squares	Mean squares	F ratio
Groups	3	44.703	14.901	19.68
Error	57	43.143	0.757	
Total	60	87.846		

The result is an F ratio with 3 degrees of freedom in the numerator and 57 degrees of freedom in the denominator. The F Distribution is scaled based on these degrees of freedom.
Using a look up table in the back of a textbook or the formula editor in JMP, the *p*-value is 7.028323e-9 (Exhibit 9).

Exhibit 9 Distribution of the Combined Data





(Create a new column. Then, right-click on the column header and select Formula.In the Formula Editor, enter a "1" and then use the key pad in the editor to enter "-". Under Functions (grouped), select Probability > F Distribution. Enter the values as shown, and click OK. If working with a new data table use Rows > Add Rows and type "1" to add 1 row. The p-value will display.)

In this example, the *p*-value is extremely small, providing evidence against the null hypothesis. It appears the alternative hypothesis is true (we can reject the null hypothesis that the means are equal).

Note that $R^2 = A/(A + B)$, or SSG/SST. For this example, R^2 (44.703/87.846) is 0.509. (We'll come back to this in a moment.)

Now consider the case when the data is consistent with the null hypothesis.

Suppose s = 1.21 and s_p = 1.145, k = 5 (there are five groups) and n = 46 (the total sample size is 46).

Source	DF	Sums of squares	Mean squares	F ratio
Groups	4	5.578	1.395	1.064
Error	41	53.751	1.311	
Total	45	59.329		

The result is an F ratio with 4 and 41 degrees of freedom. Using the calculator in JMP, the *p*-value is 0.4186. The R^2 is 0.094. In this case, we do not have evidence against the null hypothesis (we can't conclude that the means are different).

Summary

Statistical Insights

It is useful to understand the relationship between the F ratio and R^2. It turns out that:

$$F = \frac{R^2}{1 - R^2} \times \frac{n - k}{k - 1}$$

This illustrates a couple of important ideas. If the null hypothesis is true, then the population R^2, denoted ρ^2, is zero, and the sample R^2 stays close to zero. Of course, R^2 cannot actually be zero, and the sample sizes and R^2 just balance out in such a way that the F ratio stays around 1.0. The *p*-value will be uniform (0,1) no matter the sample size.

However, if the alternative hypothesis is true, ρ^2 has some fixed positive value, and as the sample size increases, R^2 approaches ρ^2. This means that the F ratio will grow approximately in proportion to $n - k$. As the sample size grows, the F ratio grows and the *p*-value becomes smaller and smaller.

Implications

Suppose the null hypothesis is true, then:

$s^2 \approx s_p^2$, $F \approx 1.0$, *p*-values are uniform (0,1), and $R^2 \approx 0.0$ for any sample size.

Suppose the alternative hypothesis is true, then:

$s^2 >> s_p^2$, $F >> 1.0$, *p*-values are biased toward smaller values relative to a uniform (0,1), and $R^2 > 0.0$. As the sample size increases, F grows in proportion to $n - k$ and the *p*-values become more biased toward low values.

JMP® Features and Hints

In this case we used the Distribution platform to compare distributions and confidence intervals for three years of the professor's backgammon scores, then used Tables > Stack to stack the scores into one column for analysis. We used Fit Y by X and ANOVA to test the hypothesis that the average scores were the same across the three years, after using Unequal Variances to test the assumption that the population variances were equal.

We used the Distribution platform and a one-sample t-Test on the combined data to test the hypothesis that the professor scores higher than the program. Finally, we used the formula editor to compute the area under the curve for the F Distribution (i.e., *p*-values) for F ratios computed by hand.

Note: An alternative to the Formula Editor for calculating *p*-values is the Probability and Distribution Calculator Expanded, which can be downloaded from jmp.com/modules. This calculator, which was written using JSL (the JMP Scripting Language), can be used to calculate the area under the curve for a number of probability distributions.

Exercises

1. A partial ANOVA table is provided below:

Source	DF	Sums of squares	Means squares	F ratio
Groups	7			
Error		455.17		
Total	64	534.34		

   a. Complete the ANOVA table.
   b. What are s and s_p?
   c. What is R^2?
   d. What is the sample size?
   e. How many groups are there?
   f. Should you reject the null hypothesis?

2. Suppose s = 2.32, s_p = 1.65, k = 3, and n = 210.

Source	DF	Sums of squares	Means squares	F ratio
Groups				
Error				
Total				

   a. Complete the ANOVA table.
   b. What is R^2?
   c. Should you reject the null hypothesis?

3. Open the file Drug.jmp from the Sample Data directory in JMP. To locate this file, go to the help menu in JMP, select Sample Data, then click Open the Sample Data Directory button.

   Thirty subjects were given one of three drugs. The response, "y" is a measure of some response. We are interested in testing whether there are differences between the three drugs.

   a. State the null and alternative hypotheses for this test.
   b. Test the assumption that the population variances are equal. What can we conclude?
   c. Use ANOVA to test the hypothesis that the average response was the same for the three drugs. What is the *p*-value for the test? Based on this *p*-value, what decision can we make relative to the null hypothesis? In nonstatistical terms, what can we conclude?