Section 4B – Categorical/Quantitative Relationships: ANOVA

Introduction

In the last section, we saw that we could use a two-population mean average hypothesis test to determine if categorical and quantitative variables are related or not. If the mean averages were the same in two groups that would indicate that the categorical variable that determines the groups is not related to the quantitative variable mean average.

H_0 : $\mu_1 = \mu_2$ (categorical variable is not related to the quantitative variable)
H_A : $\mu_1 \neq \mu_2$ (categorical variable is related to the quantitative variable)

Many times, categorical data has more than just two options. This would mean that we would need to compare three or more population means.

H_0 : $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \ldots = \mu_k$ (categorical variable is not related to the quantitative variable)
H_A : at least one population mean is ≠ (categorical variable is related to the quantitative variable)

Unfortunately, a T test statistic can only compare two things at a time and cannot handle a hypothesis test involving three or more groups. To compare three or more population means, we will need to use ANOVA.

ANOVA

ANOVA stands for "Analysis of Variance". If you remember, variance is the square of the standard deviation. Variance measures the variability from the mean. There are two specific variances that are compared in an ANOVA test, the variance between the groups and the variance within the groups. The variance between the groups is a measure of how different the groups are. It measures how much variability each of the sample means are from the mean of all the groups combined. The variance within the groups measures the amount of variability that data values in each group are from their own sample mean. In ANOVA, we compare the variance between the groups to the variance within the groups.

ANOVA tests use the F-test statistic. In ANOVA, the F-test statistic divides the variance between the groups to the variance within the groups.

F Test Statistic Sentence: The ratio of the variance between the groups to the variance within the groups.

Calculating and Interpreting the F-test Statistic

As with all difficult calculations in statistics, use a computer program to calculate the F-test statistic. Never calculate it by hand. Always focus more on interpretation than on calculation. Let us see if we can better understand how the F-test statistic works.

Variance divides the sum of squares of the differences by the degrees of freedom.

$$Variance = \frac{Sum\ of\ Squares}{Degrees\ of\ Freedom}$$

To calculate the variance between the groups, the computer calculates the sum of squares between the groups and then divides by the degrees of freedom. The sum of squares between the groups subtracts the mean of all the groups combined $(\bar{\bar{x}})$from the sample means $(\bar{x}_i)$ for each group. It squares the differences and adds them. Since the variance between calculations is based on the number of groups, the degrees of freedom between is the number of groups – 1 or "k – 1".

$$Variance = \frac{Sum\ of\ Squares\ Between}{Degrees\ of\ Freedom\ Between} = \frac{\sum(\bar{x}_i - \bar{\bar{x}})^2}{k-1}$$

To calculate the variance within the groups, we divide the sum of squares within each group divided by the sum of squares within. The “sum of squares within” subtracts each data value minus its own sample mean, squares the differences and adds them up. If we look at the degrees of freedom for each data set $(n_i - 1)$and add them up for each group, we will get the “degrees of freedom within”.

$$Variance = \frac{Sum\ of\ Squares\ Within}{Degrees\ of\ Freedom\ Within} = \frac{\sum(x - \bar{x}_i)^2}{\sum(n_i - 1)}$$

There is a beauty in the mathematics behind the F test statistic. The total number of data values for all of the groups combined minus one is often called the total degrees of freedom. There is also a total sum of squares.

Sum of Squares Between + Sum of Squares Within = Total Sum of Squares

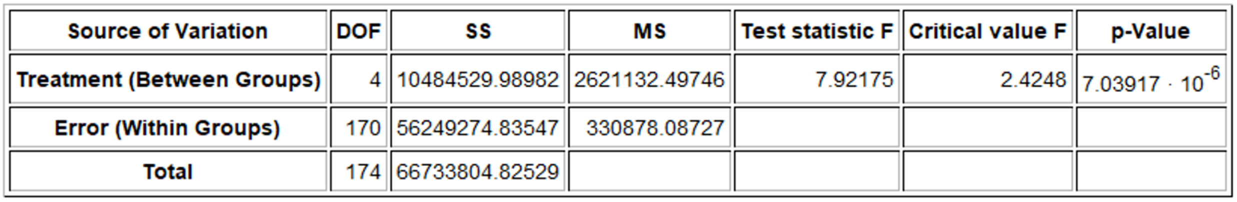
Degrees of Freedom Between + Degrees of Freedom Within = Total Degrees of Freedom

Variance Between + Variance Within = Total Variance

As we said, the F test statistic divides the Variance between the groups by the Variance within the groups. We often say that the F-test statistic is the ratio of two variances. In ANOVA, it is the ratio of the variance between the groups to the variance within the groups.

$$\text{F test statistic} = \frac{Variance\ Between\ the\ Groups}{Variance\ Within\ the\ Groups} = \frac{\left(\frac{Sum\ of\ Squares\ Between}{Degrees\ of\ Freedom\ Between}\right)}{\left(\frac{Sum\ of\ Squares\ Within}{Degrees\ of\ Freedom\ Within}\right)}$$

Computer Programs will often give you sum of squares, degrees of freedom, and variances for the F test statistic. Look at the following printout. This test used a 5% significance level.



Source of Variation	DOF	SS	MS	Test statistic F	Critical value F	p-Value
Treatment (Between Groups)	4	10484529.98982	2621132.49746	7.92175	2.4248	$7.03917 \cdot 10^{-6}$
Error (Within Groups)	170	56249274.83547	330878.08727			
Total	174	66733804.82529				

Let us see if we understand what we are seeing. Notice the MS (mean sum of squares) is the sum of squares (SS) divided by degrees of freedom (df).

“MS Treatment (Between Groups) is the variance between the groups 2621124.8 which was calculated by dividing the sum of squares (SS Between) 10484529.98982 by the degrees of freedom (DOF Between) 4.

“MS Error (Within Groups) is the variance within the groups 330878.08727 which was calculated by dividing the sum of squares (SS Within) 56249274.83547 by the degrees of freedom (DOF Within) 170.

So the F-test statistic is calculated by dividing the variances (MS).

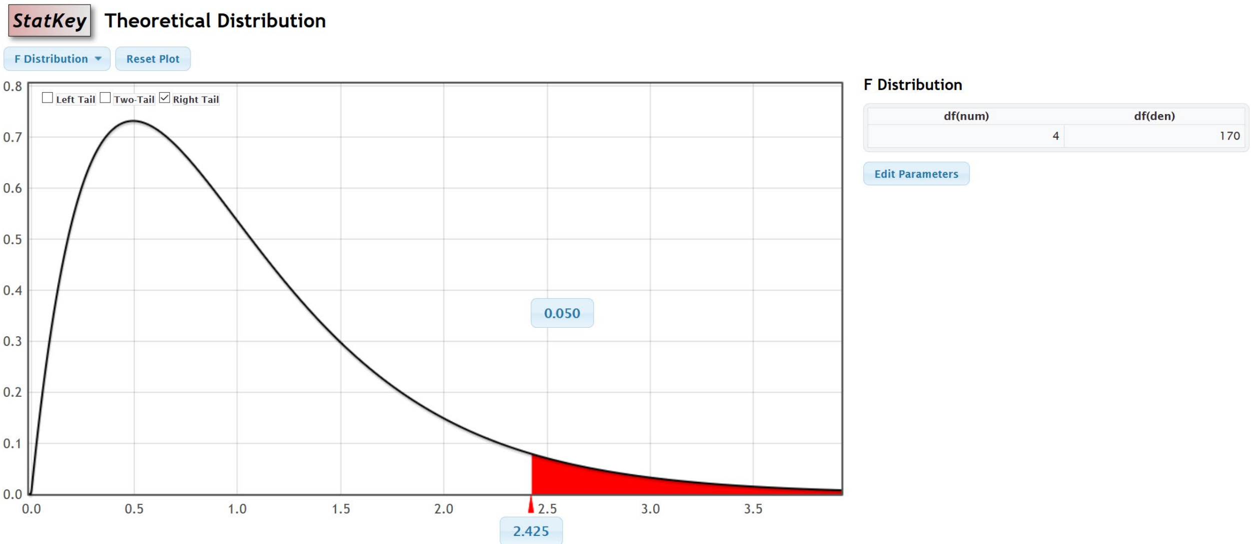
$$\text{F test statistic} = \frac{Variance\ Between\ the\ Groups}{Variance\ Within\ the\ Groups} = \frac{2621124.8}{330878.08727} = 7.92175$$

So the variance between the groups is almost 8 times greater than the variance within the groups.

Is this F test statistic significant?

Notice Statcato has calculated a critical value to compare the test statistic to. ANOVA is always a right tailed test. Remember the test statistic needs to be in the tail determined by the critical value in order to be significant. Statcato thinks that the F test statistic has to be 2.4248 or higher to be significant. Our test statistic is 7.9217, which is definitely larger than the critical value 2.4248 and falls in the right tail. So our F test statistic is significant. Therefore, the F test statistic is significantly large and the variance between the groups is significantly greater than the variance within the groups. As with all test statistics, this also tells us that the sample data significantly disagrees with the null hypothesis.

We could have looked up the critical value with StatKey as well. You will need to know the degrees of freedom between (numerator degrees of freedom) which was four and the degrees of freedom within (denominator degrees of freedom). Go to www.lock5stat.com and open StatKey. Under "Theoretical Distributions" click on "F". Put in the numerator degrees of freedom as 4 and the denominator degrees of freedom as 170. Since ANOVA is always a right tailed test and the significance level is 5%, simply click on "Right Tail" and enter 0.05 for the right tail proportion. Notice the number on the bottom is the critical value. It is about the same as what Statcato gave.



☐ Left Tail ☐ Two-Tail ☑ Right Tail

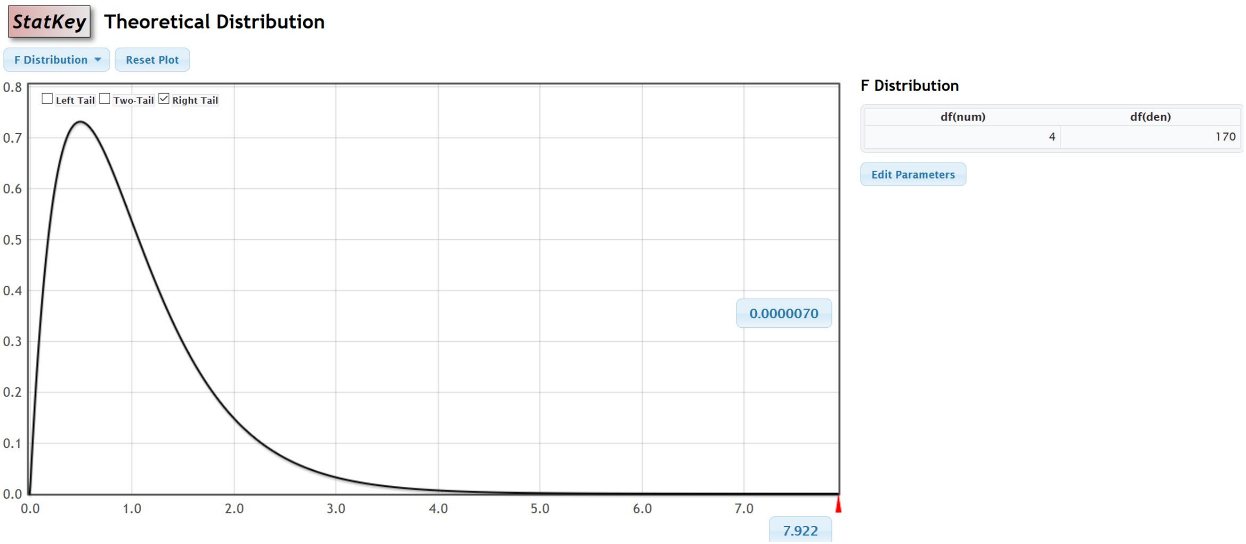
0.050

2.425

F Distribution

df(num)	df(den)
4	170

Notice we can also use the same F theoretical curve to calculate the P-value with StatKey. Remember the numerator degrees of freedom is 4 and the denominator degrees of freedom is 170. Now just put the F test statistic in the bottom box in the right tail. The proportion is the P-value. Notice the P-value calculated by StatKey is very close to the P-value calculated by Statcato.



□ Left Tail □ Two-Tail ☑ Right Tail

0.0000070

7.922

F Distribution

df(num)	df(den)
4	170

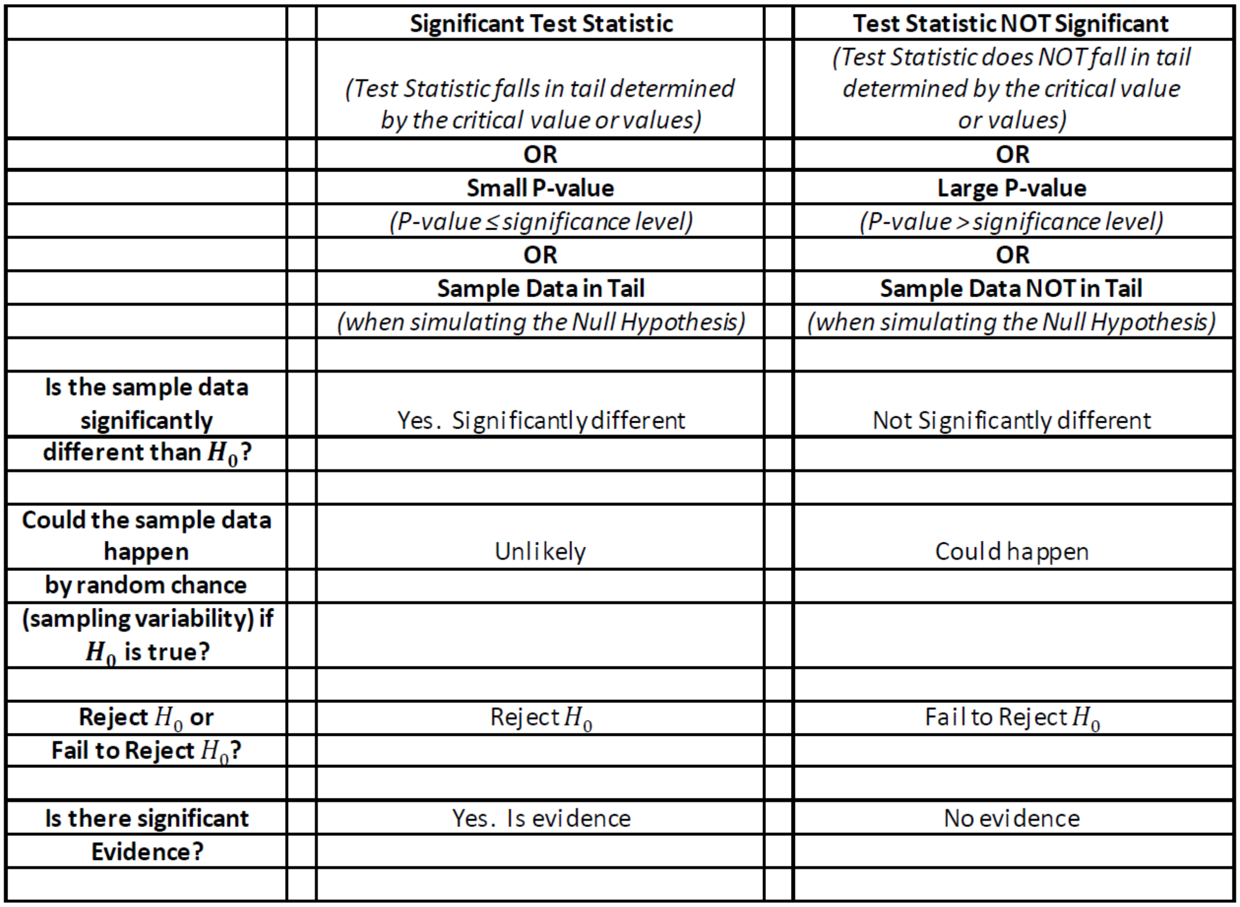
Notes about the F-test statistic

- In a fraction, when the numerator is significantly larger than the denominator, the overall fraction is large. If the variance between the groups is much larger than the variance within the groups, this will give a large F-test statistic (and a small P-value) and indicates that the sample means are significantly different. A small P-value indicates that the sample data is unlikely to happen by random chance. We will be reject the null hypothesis that the population means are the same. We are also rejecting that the categorical and quantitative variables are not related and supporting the alternative hypothesis that the variables are related.

- In a fraction, when the numerator is the same of smaller than the denominator, the overall fraction is small. So if the variance between the groups is much smaller than the variance within the groups, this will give a small F-test statistic (and a large P-value) and indicates that the sample means are not significantly different. A large P-value indicates that the sample data could have happened by random chance. We will fail to reject the null hypothesis that the population means are the same. In other words, the population means might be the same and the categorical and quantitative variables are probably not related.

- The F-test statistic can also be used in a two-population variance or two-population standard deviation hypothesis test. In that case, it compares the variance from two populations.

Here is the summary table from last chapter to remind you of the key decisions in a hypothesis test.



		Significant Test Statistic		**Test Statistic NOT Significant**
		(Test Statistic falls in tail determined by the critical value or values)		*(Test Statistic does NOT fall in tail determined by the critical value or values)*
		OR		**OR**
		Small P-value		**Large P-value**
		(P-value $\leq$ significance level)		*(P-value $>$ significance level)*
		OR		**OR**
		Sample Data in Tail		**Sample Data NOT in Tail**
		(when simulating the Null Hypothesis)		*(when simulating the Null Hypothesis)*
Is the sample data significantly		Yes. Significantly different		Not Significantly different
different than H_0?				
Could the sample data happen		Unlikely		Could happen
by random chance				
(sampling variability) if H_0 is true?				
Reject H_0 or		Reject H_0		Fail to Reject H_0
Fail to Reject H_0?				
Is there significant		Yes. Is evidence		No evidence
Evidence?				

Assumptions for an ANOVA hypothesis test

- The quantitative samples should be collected randomly or be representative of the population.
- Data values within the sample should be independent of each other.
- Data values between the samples should be independent of each other.
- The sample sizes should be at least 30 or have a nearly normal shape
- No standard deviation for any sample is more than twice as large as any other sample.

Notice that we must have a random or representative sample. As with all mean average hypothesis tests, we require the sample size to be at least 30 or have a normal shape. Data values within the samples and between the samples should be independent of each other. This again is a difficult assumption to assess. If we have a small simple random sample out of a very large population, then the data values are unlikely to be related. ANOVA is based on variance, so variability in the samples is very important. If one sample has a lot more variability than the others do, this can be a problem. Therefore, we want all of our sample standard deviations to be close. An often-used rule is that no sample standard deviation can be more than twice as large as any other can. Notice that to check these assumptions, we need to look at the sample sizes, sample means and sample standard deviations for each of our groups. We should also look at the shape of the samples with histograms or dot plots.

ANOVA Example 1: Mean Average Salaries for people living in five states in Australia.

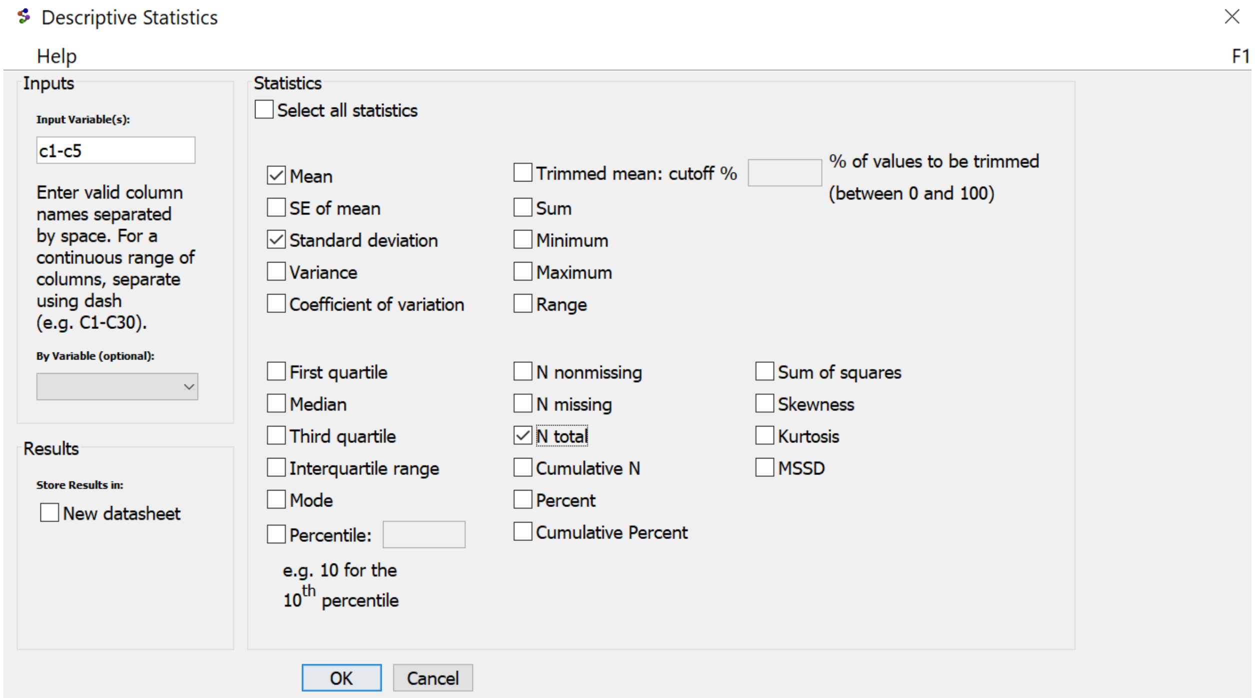
Suppose we want to compare the mean average weekly salary for people living in five states in Australia. The states are Northern Territory, New South Wales, Queensland, Victoria, and Tasmania. We claim that the mean average salary of people is related to where they live. To support this claim, we will need to show that the mean average salaries are different in these states. As with all multiple population hypothesis tests, you should label the populations. To perform this test, adults were randomly selected from each of the five states. We will be using a 5% significance level.

μ_1 : Northern Territory
μ_2 : New South Wales
μ_3 : Queensland
μ_4 : Victoria
μ_5 : Tasmania

Here is the null and alternative hypothesis for the ANOVA test. Remember an ANOVA is a multiple μ test for three or more groups. Notice that if the population mean average salary is the same, then it does not matter which state the person lives in. This implies that the state (categorical variable) is not related to the salary (quantitative variable). If at least one population mean average salary is different, then it does matter which state the person lives in. This implies that the state (categorical variable) is related to the salary (quantitative variable). Again, we see that “not related” is the null hypothesis and “related” is the alternative.

Ho: μ1 = μ2 = μ3 = μ4 = μ5 *(states in Australia are not related to salary)*
Ha: at least one is ≠ (CLAIM) *(states in Australia are related to salary)*

When doing an ANOVA test, it is good to find the sample size (n), the sample mean of each group, and the standard deviation for each group. We also will need to create histograms to check the shape of our samples. Go to www.matt-teachout.org and click on the “statistics” tab and then “data sets”. You can either open the Australia Salary data Statcato file in Statcato or open the excel file and copy and paste it into Statcato. The adults in this sample data were randomly selected. To calculate the sample sizes, means and standard deviations, go to the “statistics” menu in Statcato, then click on “basic statistics” then “descriptive statistics”. To create histograms, go the “graph” menu and click on histogram. In small data sets like these, I prefer three bins (bars). If makes it easier to see the shape. In addition, if you click on “Show Legend” the computer will also make a title for the graph. Here is the sample statistics and graphs from Statcato.



Descriptive Statistics

Inputs

Input Variable(s): c1-c5

Enter valid column names separated by space. For a continuous range of columns, separate using dash (e.g. C1-C30).

By Variable (optional):

Results

Store Results in:

□ New datasheet

Statistics

□ Select all statistics

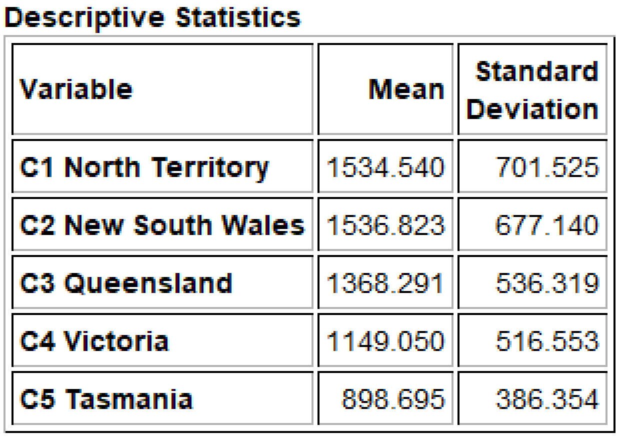
☑ Mean
□ SE of mean
☑ Standard deviation
□ Variance
□ Coefficient of variation

□ Trimmed mean: cutoff % — % of values to be trimmed (between 0 and 100)
□ Sum
□ Minimum
□ Maximum
□ Range

□ First quartile
□ Median
□ Third quartile
□ Interquartile range
□ Mode
□ Percentile: — e.g. 10 for the 10th percentile

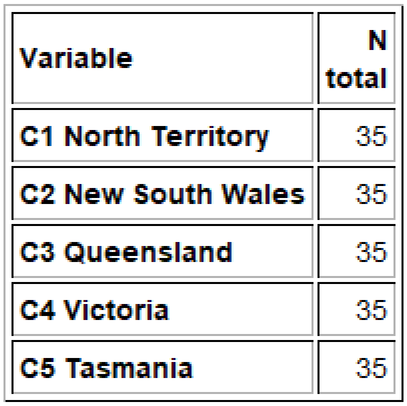
□ N nonmissing
□ N missing
☑ N total
□ Cumulative N
□ Percent
□ Cumulative Percent

□ Sum of squares
□ Skewness
□ Kurtosis
□ MSSD

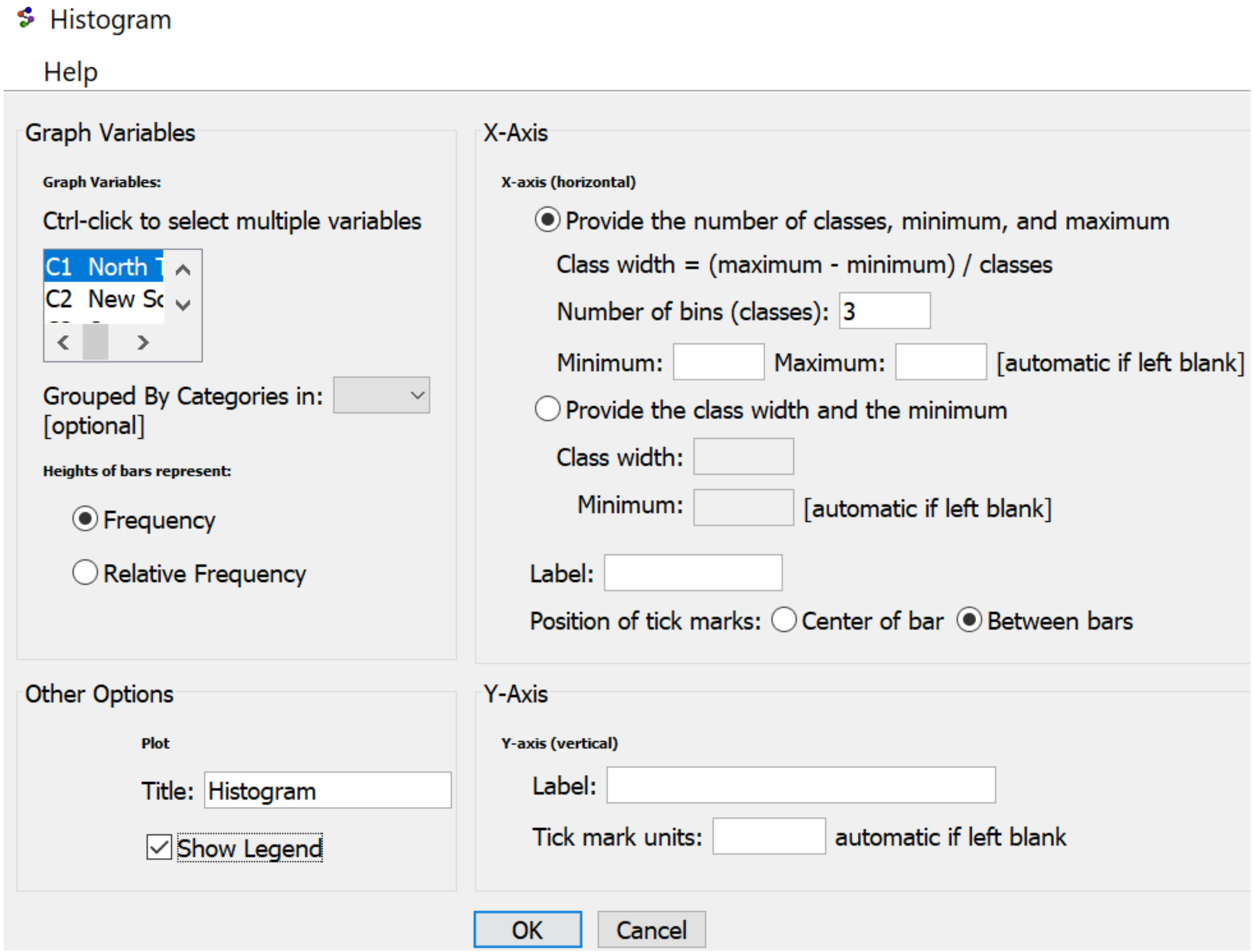


Descriptive Statistics

Variable	Mean	Standard Deviation
C1 North Territory	1534.540	701.525
C2 New South Wales	1536.823	677.140
C3 Queensland	1368.291	536.319
C4 Victoria	1149.050	516.553
C5 Tasmania	898.695	386.354



Variable	N total
C1 North Territory	35
C2 New South Wales	35
C3 Queensland	35
C4 Victoria	35
C5 Tasmania	35



Histogram

Help

Graph Variables

Graph Variables:

Ctrl-click to select multiple variables

C1 North T

C2 New Sc

Grouped By Categories in: [optional]

Heights of bars represent:

● Frequency

○ Relative Frequency

X-Axis

X-axis (horizontal)

● Provide the number of classes, minimum, and maximum

Class width = (maximum - minimum) / classes

Number of bins (classes): 3

Minimum: Maximum: [automatic if left blank]

○ Provide the class width and the minimum

Class width:

Minimum: [automatic if left blank]

Label:

Position of tick marks: ○ Center of bar ● Between bars

Other Options

Plot

Title: Histogram

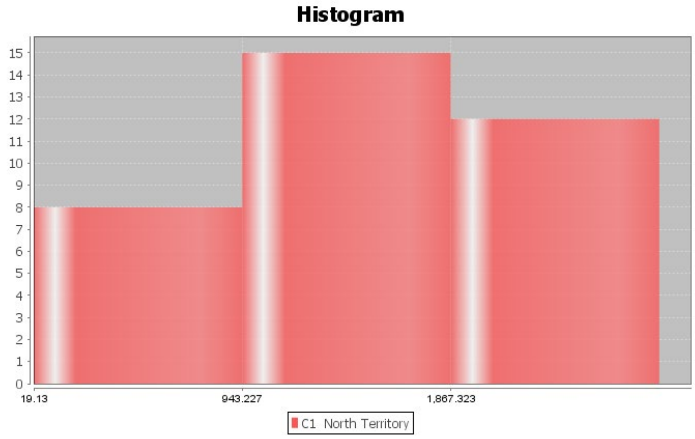
☑ Show Legend

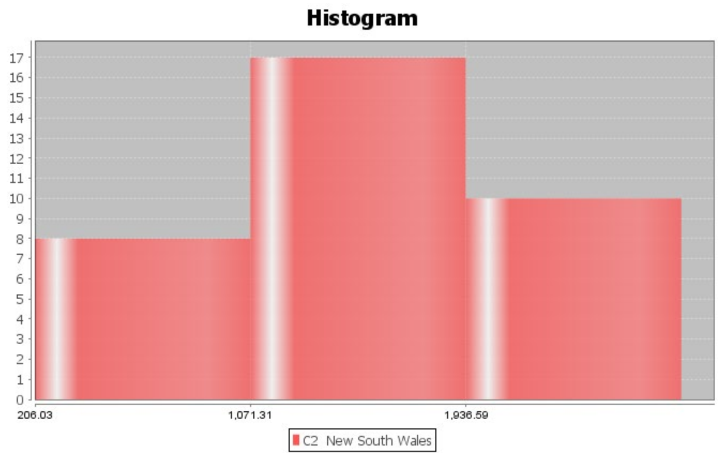
Y-Axis

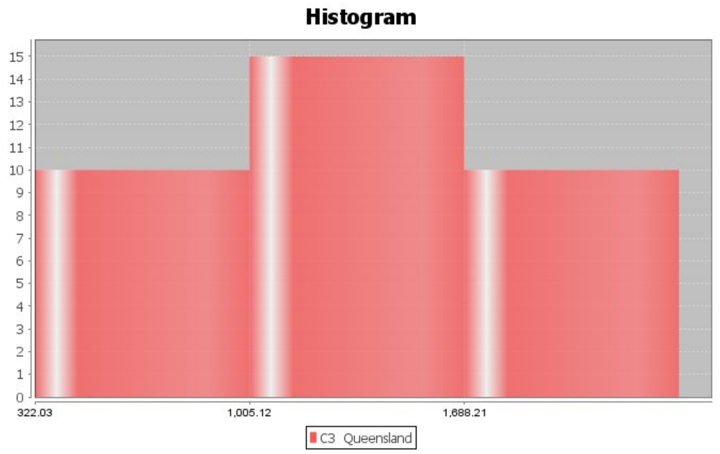
Y-axis (vertical)

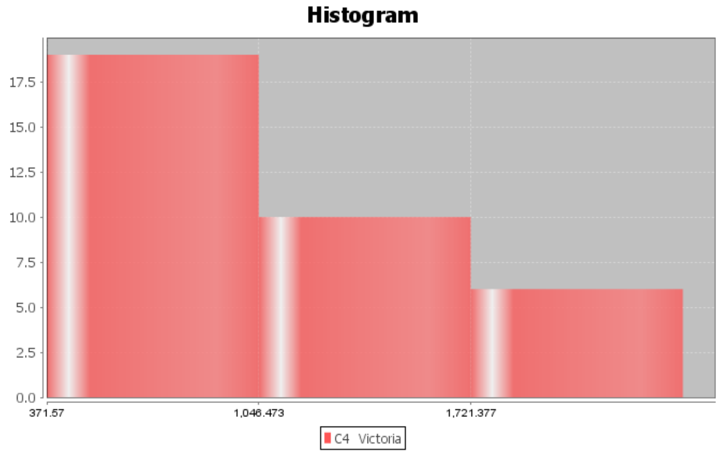
Label:

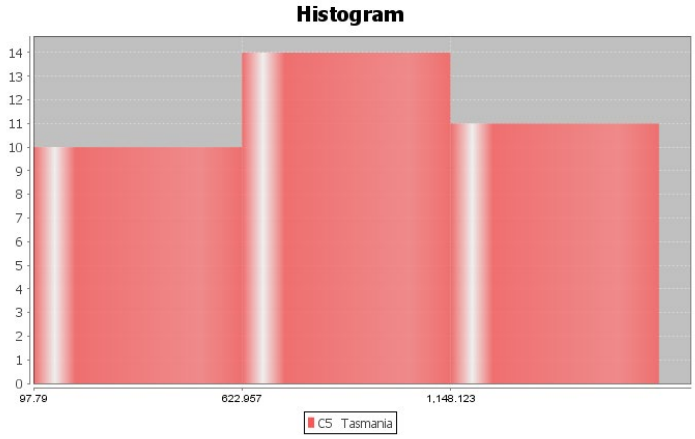
Tick mark units: automatic if left blank




Histogram
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
206.03
1,071.31
1,936.59
C2 New South Wales


Histogram
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
322.03
1,005.12
1,688.21
C3 Queensland


Histogram
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
371.57
1,046.473
1,721.377
C4 Victoria



Assumptions: Notice that this data passes all of the assumptions for the ANOVA hypothesis test.

1. The sample data should be random or representative of the population. Yes. The sample data sets were collected randomly.

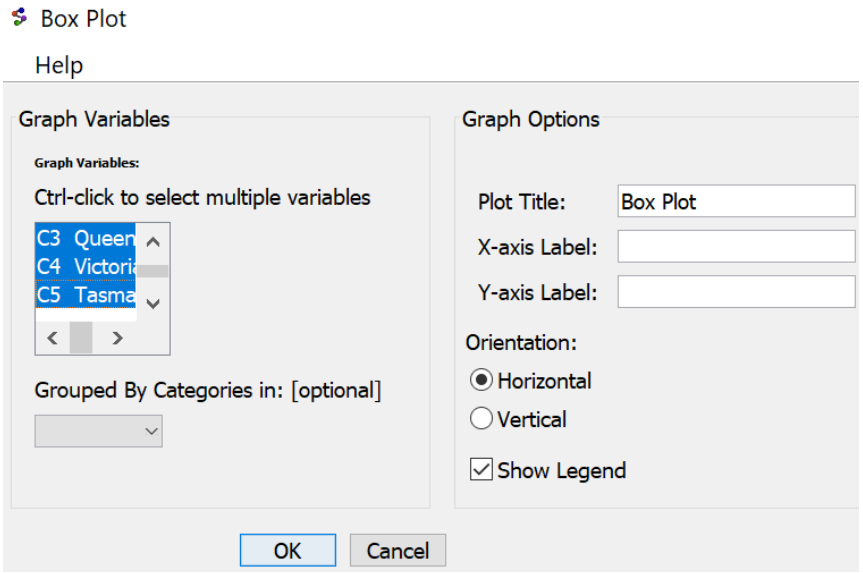
2. Each sample should have a sample size of at least 30 or be nearly normal. Yes. All of the samples had a nearly normal shape except for the data from Victoria, which was skewed right. The sample size for all of the samples was 35. So even though data from Victoria was skewed right, its sample size was still over 30. All of the other samples sizes were over 30 and normal.

3. Data values within the samples and between the samples should be independent of each other. Yes. Since we are dealing with small random samples out of millions in the populations, it is unlikely that these data values are related.

4. The sample standard deviations for the groups should be close. No standard deviation should be more than twice as large as any other should.   Yes. The sample standard deviations are close. No sample standard deviation is more than twice as large as any other sample standard deviation. Notice that the smallest standard deviation was 386.3 and the largest was 701.5 and all of the others are in between.

Some data scientists like to create a side-by-side boxplot when performing an ANOVA test.   This is surprising since the ANOVA test looks at means and standard deviations for center and spread, yet box plots look at the median and interquartile range (IQR). The boxplot can still show us general tendencies about shape, center and spread.

In Statcato, go to the “Graph” menu and click on “Box Plot”. Hold the control key down and highlight all five of the data sets. You can create a vertical or horizontal box plot. “Show Legend” will create a title. Do NOT click on the “Group By” button. Here is the box plot from Statcato.



Box Plot

Help

Graph Variables

Graph Variables:

Ctrl-click to select multiple variables

C3 Queen
C4 Victori
C5 Tasma

Grouped By Categories in: [optional]

Graph Options

Plot Title: Box Plot

X-axis Label:

Y-axis Label:

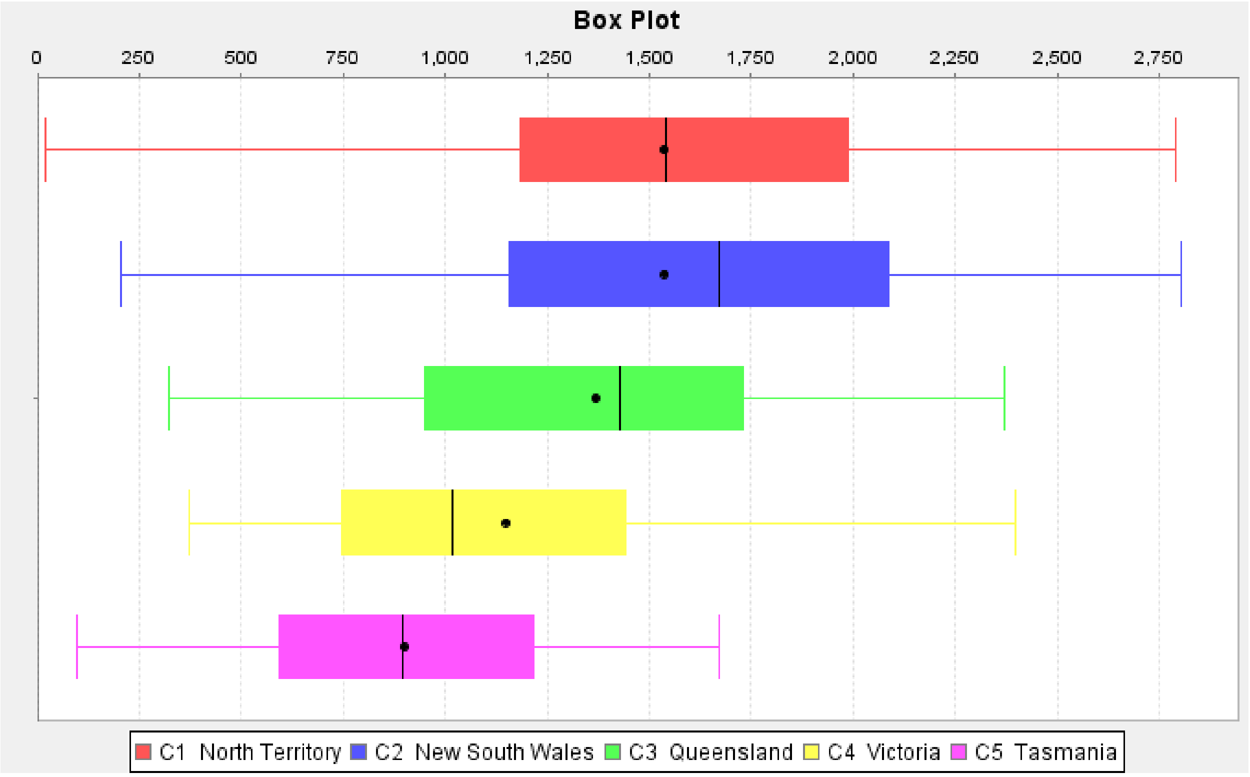
Orientation:

● Horizontal

○ Vertical

☑ Show Legend

OK Cancel



This graph tells a lot. We can see that the centers are quite different. The length of the box is a measure of spread (IQR). The lengths of the boxes are all pretty similar. If one box was more than twice as long as another was, this might indicate that one group has a lot more variability than another does. We can also get a sense of the shapes of these data sets, though separate histograms are better. So this side-by-side boxplot shows us that the variability is similar in the groups but the centers are quite different. Only the data from Victoria looks skewed right.

The key question: Are these sample means different because of sampling variability (random chance) OR are they different because at least one of the populations really is different?

To answer this, we need the F test statistic and a P-value.

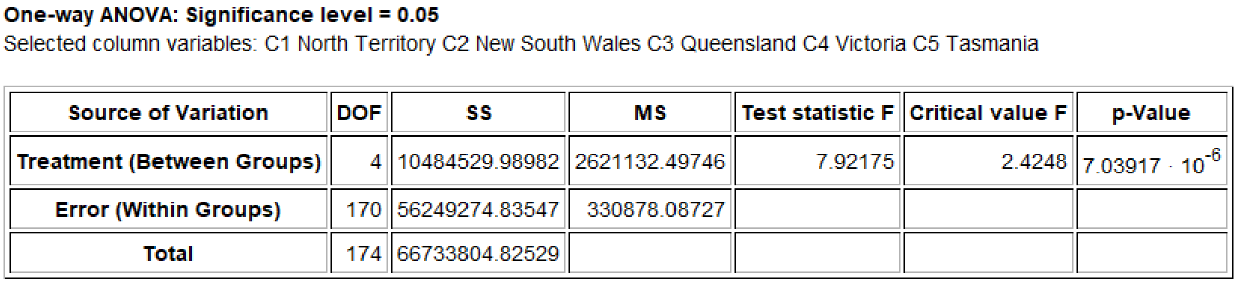
<u>How to do an ANOVA test with Statcato</u>

Copy and paste your raw quantitative data from each group into some columns in Statcato.

To calculate the F-test statistic and P-value, go to the "statistics" menu, then "Analysis of Variance", then "One-Way ANOVA".

Statistics ➔ Analysis of Variance ➔ One-Way ANOVA

Hold the control key down to select the columns where your data is and push "add to list". Select your significance level and push "OK". Here is the printout we got. Notice this is the same printout we were looking at before.



One-way ANOVA: Significance level = 0.05

Selected column variables: C1 North Territory C2 New South Wales C3 Queensland C4 Victoria C5 Tasmania

Source of Variation	DOF	SS	MS	Test statistic F	Critical value F	p-Value
Treatment (Between Groups)	4	10484529.98982	2621132.49746	7.92175	2.4248	$7.03917 \cdot 10^{-6}$
Error (Within Groups)	170	56249274.83547	330878.08727			
Total	174	66733804.82529				

Let us see if we understand what we are seeing. Notice the MS (variance) is the sum of squares (SS) divided by degrees of freedom (DOF).

MS (Treatment) is the variance between the groups (2621124.8)

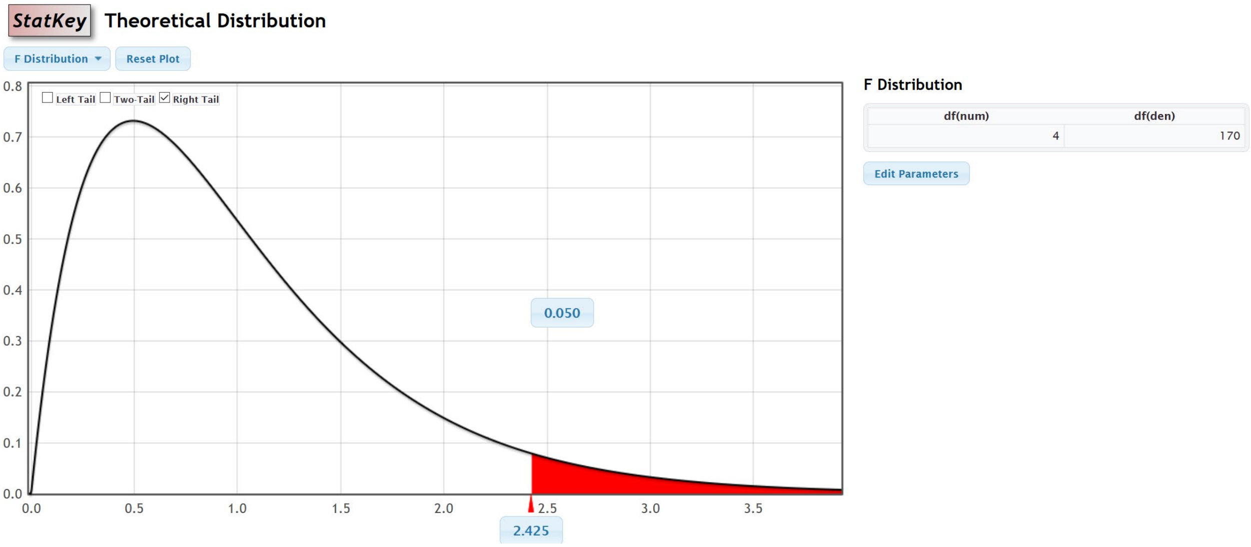
MS (Error) is the variance within the groups (330878.43)

So the F-test statistic is calculated by the formula

F test statistic = $\frac{Variance\ Between\ the\ Groups}{Variance\ Within\ the\ Groups} = \frac{2621124.8}{330878.08727}$ = 7.92175

So the variance between the groups is almost 8 times greater than the variance within the groups. Is this significantly large for an F?

Notice Statcato has calculated a critical value to compare the test statistic to. Remember the test statistic needs to fall in the tail determined by the critical value to be significant. ANOVA is always a right tailed test. Look at the following picture created by StatKey. We see that our test statistic is 7.9217 falls in the right tail. So the F test statistic is significant. This also tells us that the sample data significantly disagrees with the null hypothesis and that the <u>variance between</u> the groups is <u>significantly greater</u> than the <u>variance within</u> the groups. Otherwise, the F test statistic would not have fallen in the right tail.



Theoretical Distribution

□ Left Tail □ Two-Tail ☑ Right Tail

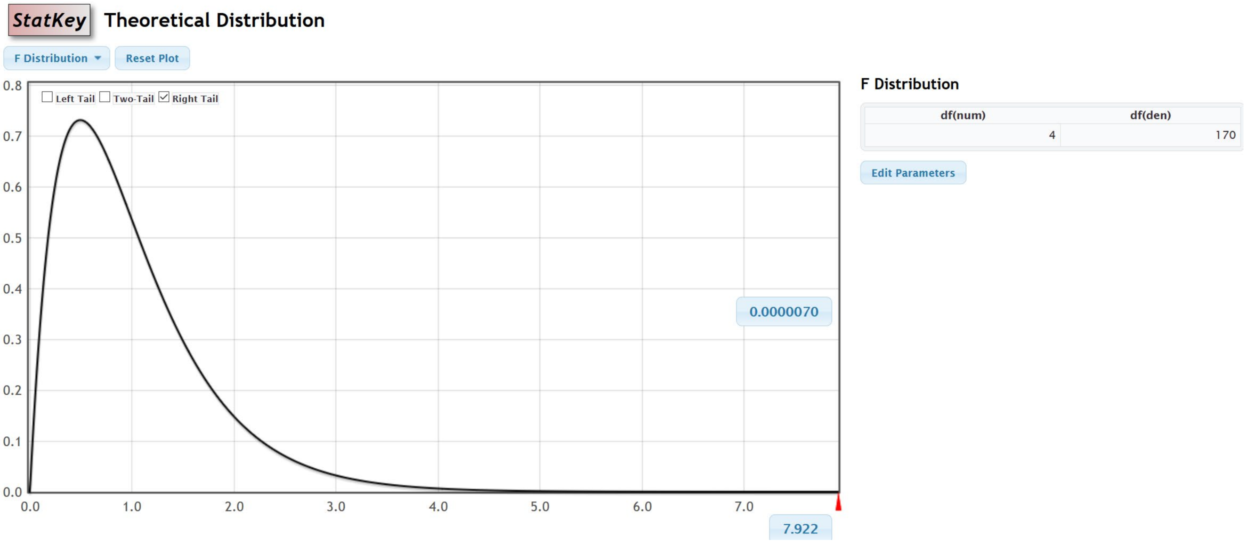
F Distribution

df(num)	df(den)
4	170

↑

F = 7.92175

Notice we can also use the same F theoretical curve to calculate the P-value with StatKey. Remember the numerator degrees of freedom is 4 and the denominator degrees of freedom is 170. Now just put the F test statistic in the bottom box in the right tail. The proportion is the P-value. Notice the P-value calculated by StatKey is very close to the P-value calculated by Statcato.



Theoretical Distribution

□ Left Tail □ Two-Tail ☑ Right Tail

F Distribution

df(num)	df(den)
4	170

The test statistic fell in the tail determined by the critical value, so the sample data does significantly disagree with the null hypothesis.

Notice that in our printout from Statcato, we got the following P-value: “7.039 x 10^-6”. This is scientific notation. Move the decimal six places to the left to get the P-value as a decimal.

P-value = 0.00000704

The actual P-value is very close to zero and much lower than a 5% significance level. From our study of P-values, we know this is very significant and unlikely to happen by random chance (sampling variability).

Since the P-value is less than our significance level, we should reject the null hypothesis.

Conclusion

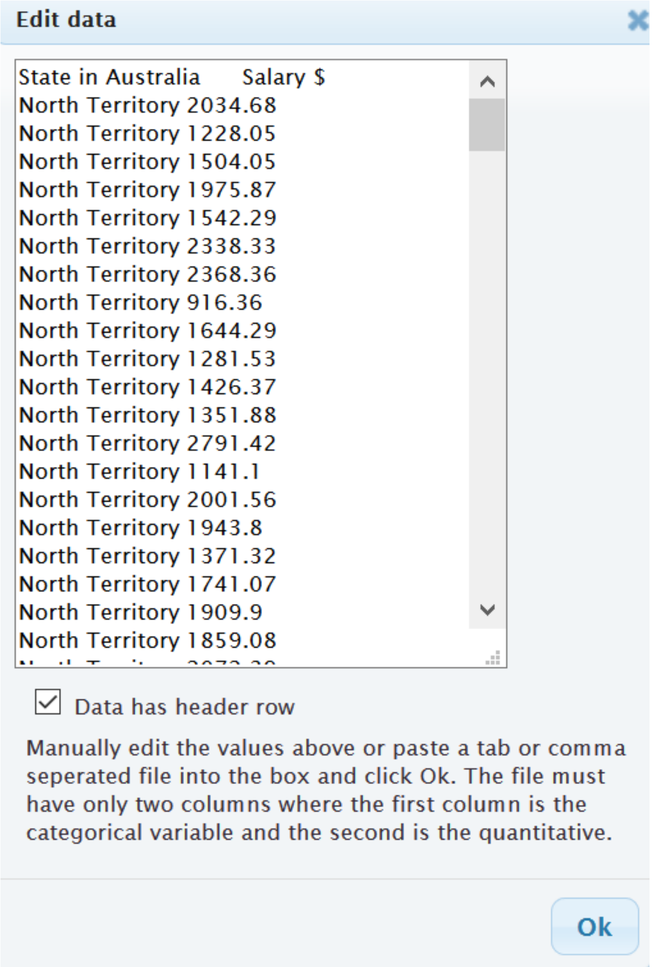
Recall the claim was the alternative hypothesis that where a person lives is related to the salary. To show this we needed to have evidence to support that at least one state was different from the others (alternative hypothesis). Since we rejected the null, we support this claim. Our P-value is very small and our F test statistic very large, so we have significant evidence.

Conclusion: There is significant evidence to support the claim that the mean average salaries of people in Northern Territory, New South Wales, Queensland, Victoria, and Tasmania are different and that the state a person lives in is related to the salary.

Note: Remember "relationship" does not mean "causation" though. This was not an experiment and did not control confounding variables. There are many reasons why a persons' salary is high or low. It would be wrong to say that the place a person lives causes their salary to be low or high.

Simulation

Remember we can also estimate the P-value and determine significance with randomized simulation. Go to www.lock5stat.com and open StatKey. We will need to go to www.matt-teachout.org and open the "Australia Salary Data" in Excel. In Statcato, we needed the quantitative data separated by group, but in StatKey, we need the raw categorical and quantitative data. StatKey will separate the data. In the excel spreadsheet you will see the column that says, "State in Australia" and "Salary". Copy these two data sets together. Under the "More Advanced Randomization Tests" menu click on "ANOVA for Difference in Means". Click on "Edit Data" and paste in the state and salary columns.



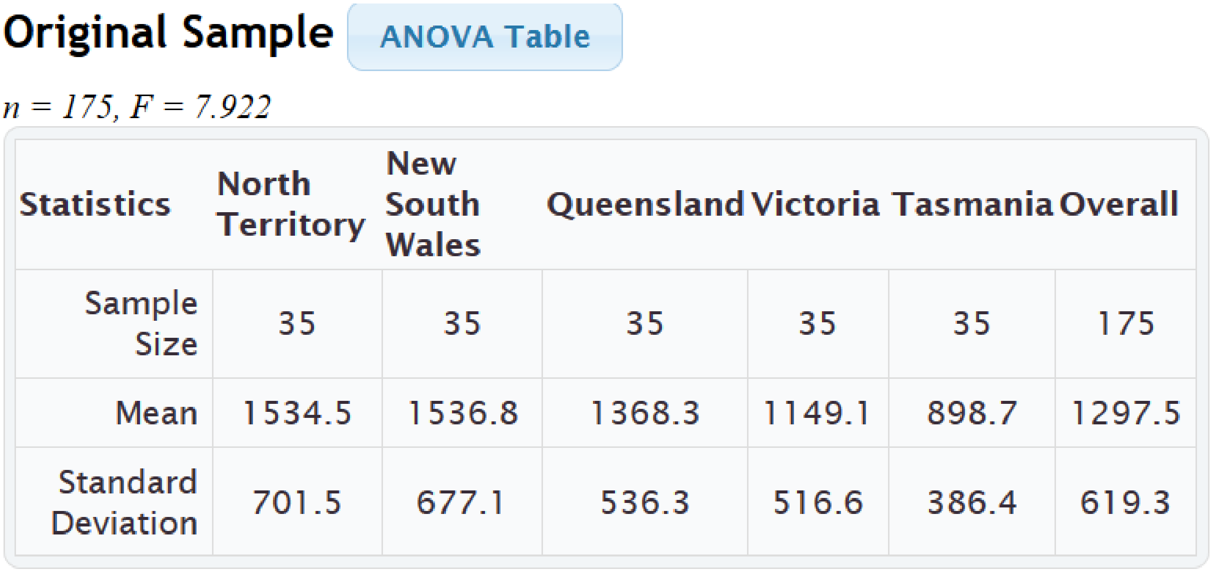
Edit data

```
State in Australia	Salary $
North Territory 2034.68
North Territory 1228.05
North Territory 1504.05
North Territory 1975.87
North Territory 1542.29
North Territory 2338.33
North Territory 2368.36
North Territory 916.36
North Territory 1644.29
North Territory 1281.53
North Territory 1426.37
North Territory 1351.88
North Territory 2791.42
North Territory 1141.1
North Territory 2001.56
North Territory 1943.8
North Territory 1371.32
North Territory 1741.07
North Territory 1909.9
North Territory 1859.08
```

☑ Data has header row

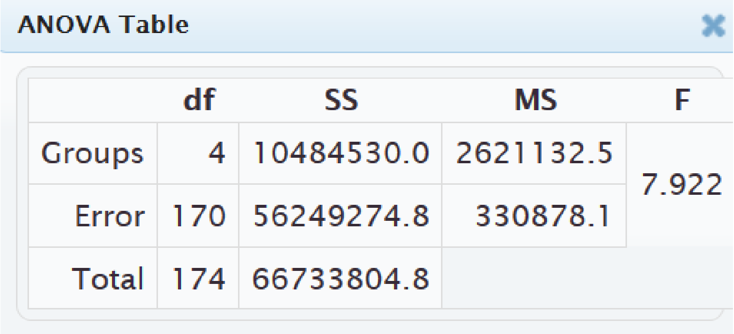
Manually edit the values above or paste a tab or comma seperated file into the box and click Ok. The file must have only two columns where the first column is the categorical variable and the second is the quantitative.

Notice under “Original Sample”, StatKey has calculated the F-test statistic for you along with the sample means, sample sizes and sample standard deviations. If you wish to see the variance between and the variance within calculations click on “ANOVA Table”. It looks similar to the Statcato printout.



$n = 175, F = 7.922$

Statistics	North Territory	New South Wales	Queensland	Victoria	Tasmania	Overall
Sample Size	35	35	35	35	35	175
Mean	1534.5	1536.8	1368.3	1149.1	898.7	1297.5
Standard Deviation	701.5	677.1	536.3	516.6	386.4	619.3

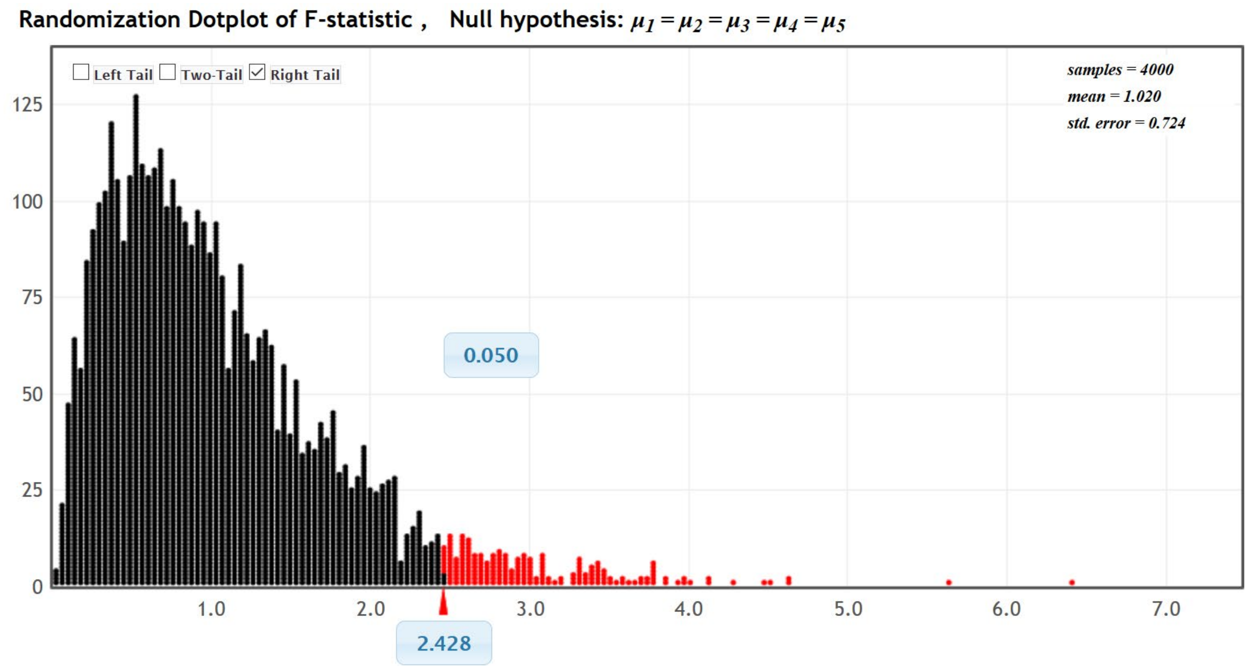


ANOVA Table

	df	SS	MS	F
Groups	4	10484530.0	2621132.5	7.922
Error	170	56249274.8	330878.1	
Total	174	66733804.8		

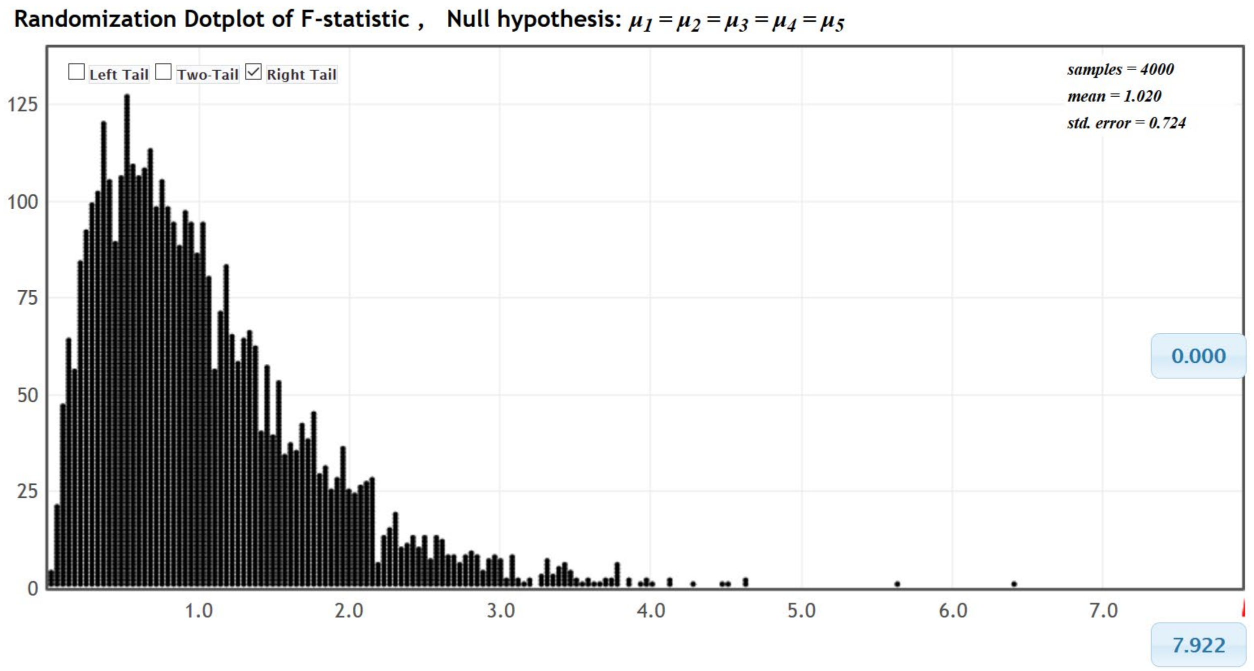
Notice the null hypothesis is that all five population means are equal. To simulate the null hypothesis click “Generate 1000 Samples” a few times. In simulations with only one or two groups, we usually use the sample mean, sample mean difference, sample proportion, or sample proportion difference. In tests with more than two groups, we cannot use that approach. When a test involves three or more groups, we will resort to using the test statistic to summarize the sample data.

In this simulation, the computer has randomly collected thousands of samples and calculated thousands of F test statistics. Remember the real test statistic can be found under “Original Sample”. ANOVA is a right tailed test so we will click on “Right-Tail”. If we put in the 5% significance level in the proportion box in the right tail, we will have the critical value. Because of sampling variability, you will get slightly different answers, but this simulation gave a critical value of 2.428, which is not far from the theoretical critical value calculated by Statcato earlier. We can now use this graph to determine if the test statistic falls in the tail. Notice our F-test statistic of 7.922 does fall in the tail, so our sample data significantly disagrees with the null hypothesis and our variance between the groups is significantly higher than the variance within the groups.



↑
F =7.922

We can also calculate the P-value by putting the test statistic in the bottom box of the simulation. Notice our P-value came out to be about zero. So this sample data is unlikely to occur because of sampling variability if the null hypothesis was true. We would reject the null hypothesis and get the same conclusion as we did with the traditional approach.


Randomization Dotplot of F-statistic , Null hypothesis: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$
□ Left Tail □ Two-Tail ☑ Right Tail
samples = 4000
mean = 1.020
std. error = 0.724
