Lab 4 ANOVA and Chi-square

Instructions: Read through and answer or implement the instructions given below. You will submit your answers in a lab report through Canvas. For your report, please answer the questions in narrative form where possible and using screenshots where needed. For instance, any graph needs a screenshot. When in doubt, give a screenshot. The lab report is best submitted in Word® or .pdf® format in Canvas (e.g. Google Docs and Apple Numbers are not permitted).

Goal: Learn how to perform ANOVA and Chi-square tests in R.

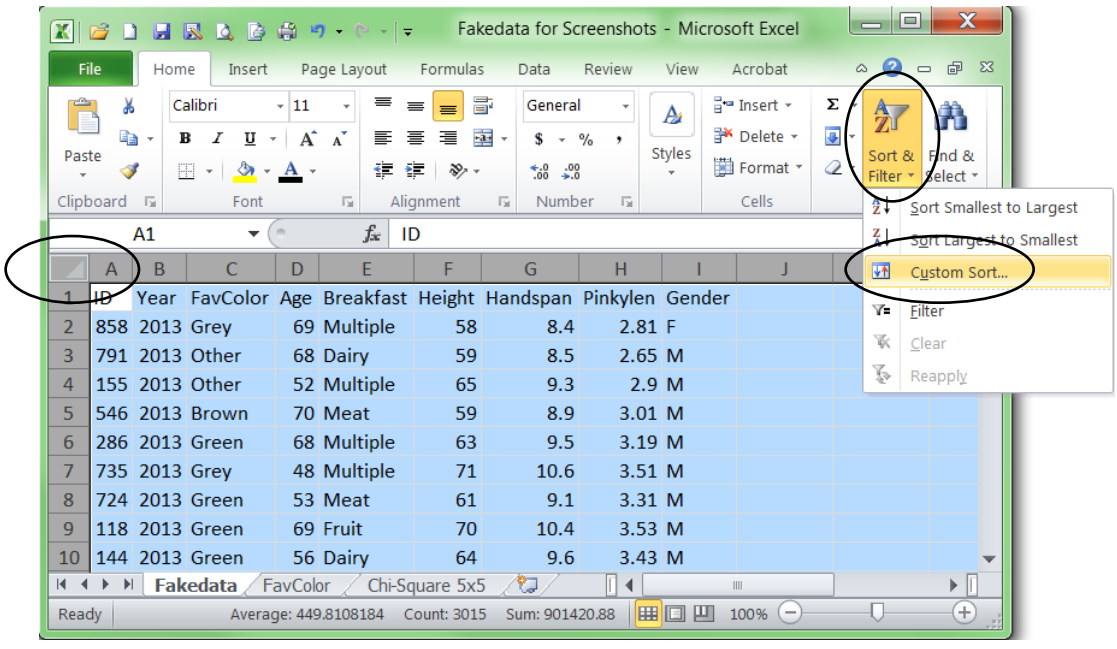
1. **Loading data**
   We begin this lab by first loading our data into Excel® for analysis.
   Start Excel® and open "Fakedata.xlsx" as you did in Lab 1. You need not include anything in your lab report for this step.

2. **ANOVA**
   Suppose that we would like to test whether there is a significant difference in the mean age of those whose favorite color is Brown, Green, Grey or Other (I know, it seems like a strange relationship to test). We will label these groups with the numbers 1, 2, 3 and 4 respectively. Hence the hypothesis test we aim to compute is

   H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$,
   H_a: at least one of the means is different

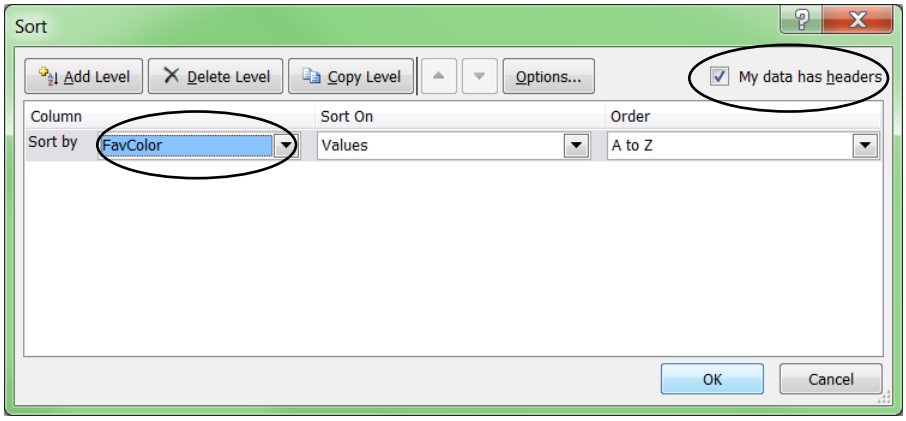
   a. As with most hypothesis tests, you should first construct the appropriate numerical and graphical summaries for each of these groups. In lab 2 you constructed the appropriate graphical summaries for a C-Q relationship. Construct those now and include them in your lab report.

   b. In preparation for running the ANOVA, we need to do some sorting and copying. Select the entire sheet by clicking on the box above the line numbers to the left of the column labels:

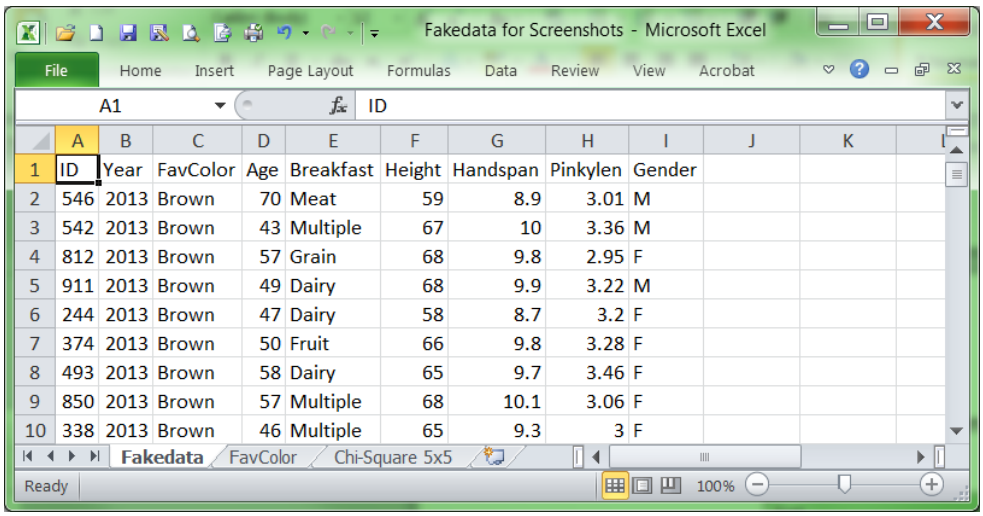


	A	B	C	D	E	F	G	H	I
1	ID	Year	FavColor	Age	Breakfast	Height	Handspan	Pinkylen	Gender
2	858	2013	Grey	69	Multiple	58	8.4	2.81	F
3	791	2013	Other	68	Dairy	59	8.5	2.65	M
4	155	2013	Other	52	Multiple	65	9.3	2.9	M
5	546	2013	Brown	70	Meat	59	8.9	3.01	M
6	286	2013	Green	68	Multiple	63	9.5	3.19	M
7	735	2013	Grey	48	Multiple	71	10.6	3.51	M
8	724	2013	Green	53	Meat	61	9.1	3.31	M
9	118	2013	Green	69	Fruit	70	10.4	3.53	M
10	144	2013	Green	56	Dairy	64	9.6	3.43	M

On the **Home** menu tab, click on the **Sort & Filter** option. Roll down and select **Custom Sort.**

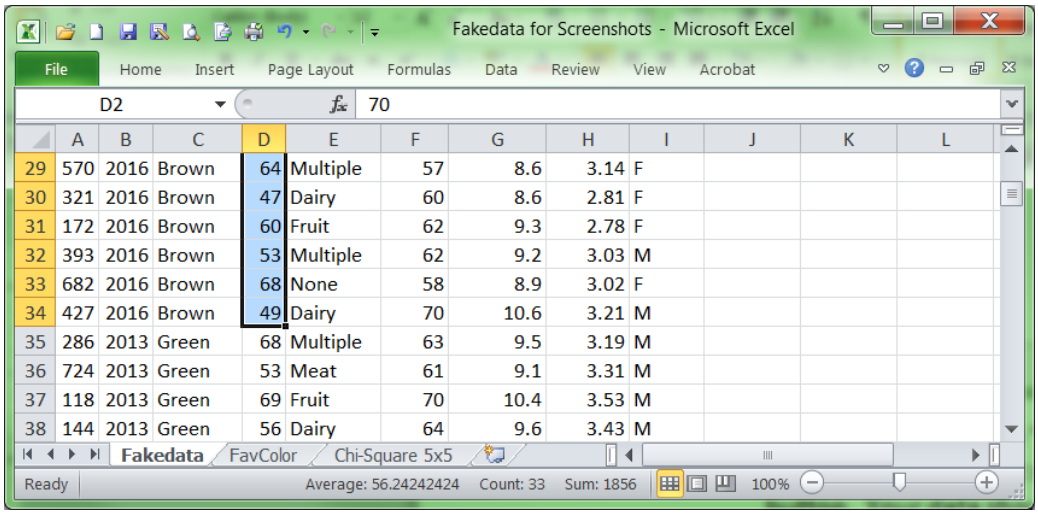


Be sure and check the "My data has headers" if it is not already selected. Click the arrow on the right side of the **Sort by** menu and select FavColor. Click the OK button. Your data should now begin with all of the Browns at the top.



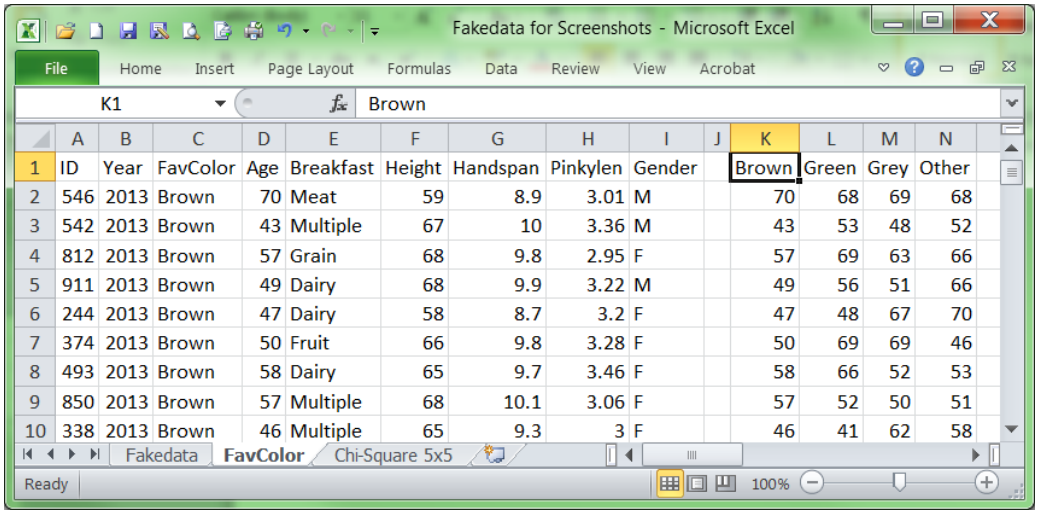
	A	B	C	D	E	F	G	H	I
1	ID	Year	FavColor	Age	Breakfast	Height	Handspan	Pinkylen	Gender
2	546	2013	Brown	70	Meat	59	8.9	3.01	M
3	542	2013	Brown	43	Multiple	67	10	3.36	M
4	812	2013	Brown	57	Grain	68	9.8	2.95	F
5	911	2013	Brown	49	Dairy	68	9.9	3.22	M
6	244	2013	Brown	47	Dairy	58	8.7	3.2	F
7	374	2013	Brown	50	Fruit	66	9.8	3.28	F
8	493	2013	Brown	58	Dairy	65	9.7	3.46	F
9	850	2013	Brown	57	Multiple	68	10.1	3.06	F
10	338	2013	Brown	46	Multiple	65	9.3	3	F

Highlight the Age column from the first Brown to the last brown.



	A	B	C	D	E	F	G	H	I
29	570	2016	Brown	64	Multiple	57	8.6	3.14	F
30	321	2016	Brown	47	Dairy	60	8.6	2.81	F
31	172	2016	Brown	60	Fruit	62	9.3	2.78	F
32	393	2016	Brown	53	Multiple	62	9.2	3.03	M
33	682	2016	Brown	68	None	58	8.9	3.02	F
34	427	2016	Brown	49	Dairy	70	10.6	3.21	M
35	286	2013	Green	68	Multiple	63	9.5	3.19	M
36	724	2013	Green	53	Meat	61	9.1	3.31	M
37	118	2013	Green	69	Fruit	70	10.4	3.53	M
38	144	2013	Green	56	Dairy	64	9.6	3.43	M

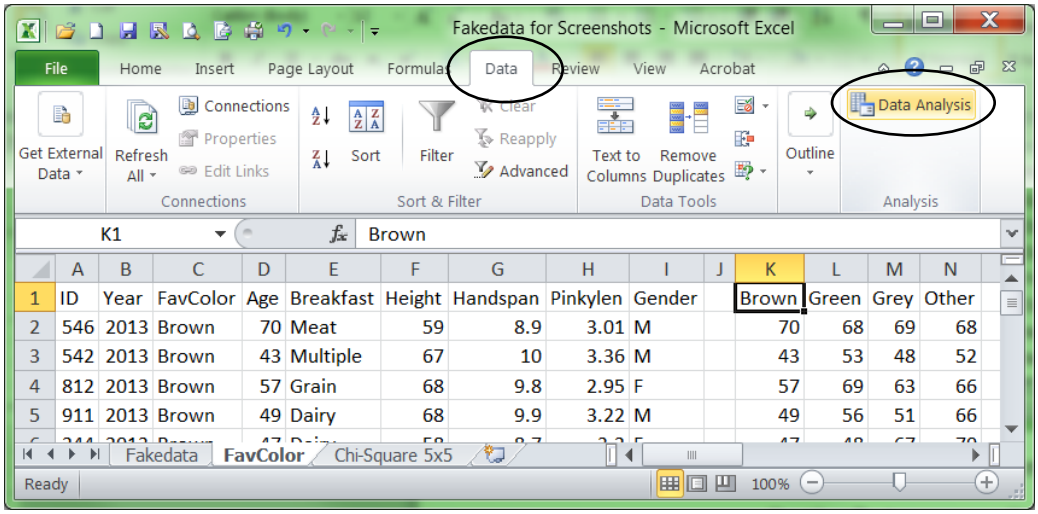
Type the word Brown in the top cell of column K and paste the Brown ages below that. Repeat this for the other 3 colors.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	ID	Year	FavColor	Age	Breakfast	Height	Handspan	Pinkylen	Gender		Brown	Green	Grey	Other
2	546	2013	Brown	70	Meat	59	8.9	3.01	M		70	68	69	68
3	542	2013	Brown	43	Multiple	67	10	3.36	M		43	53	48	52
4	812	2013	Brown	57	Grain	68	9.8	2.95	F		57	69	63	66
5	911	2013	Brown	49	Dairy	68	9.9	3.22	M		49	56	51	66
6	244	2013	Brown	47	Dairy	58	8.7	3.2	F		47	48	67	70
7	374	2013	Brown	50	Fruit	66	9.8	3.28	F		50	69	69	46
8	493	2013	Brown	58	Dairy	65	9.7	3.46	F		58	66	52	53
9	850	2013	Brown	57	Multiple	68	10.1	3.06	F		57	52	50	51
10	338	2013	Brown	46	Multiple	65	9.3	3	F		46	41	62	58

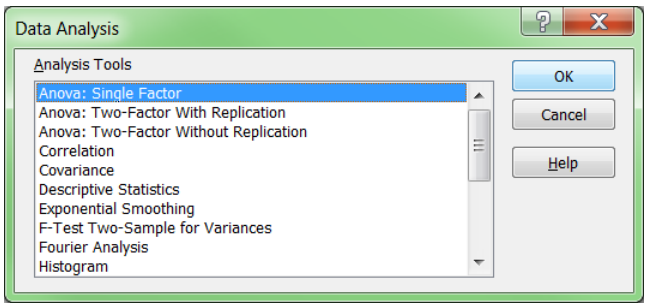
We are now set up to run the ANOVA.

c. On the **Data** menu, click the **Data Analysis** option.



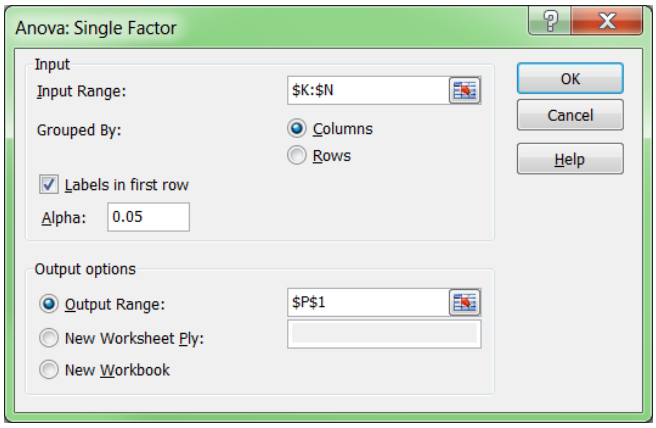
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	ID	Year	FavColor	Age	Breakfast	Height	Handspan	Pinkylen	Gender		Brown	Green	Grey	Other
2	546	2013	Brown	70	Meat	59	8.9	3.01	M		70	68	69	68
3	542	2013	Brown	43	Multiple	67	10	3.36	M		43	53	48	52
4	812	2013	Brown	57	Grain	68	9.8	2.95	F		57	69	63	66
5	911	2013	Brown	49	Dairy	68	9.9	3.22	M		49	56	51	66

The various Analysis Tools will be listed in the **Data Analysis Dialog** box. Scroll to **Anova: Single Factor** and click the OK button.

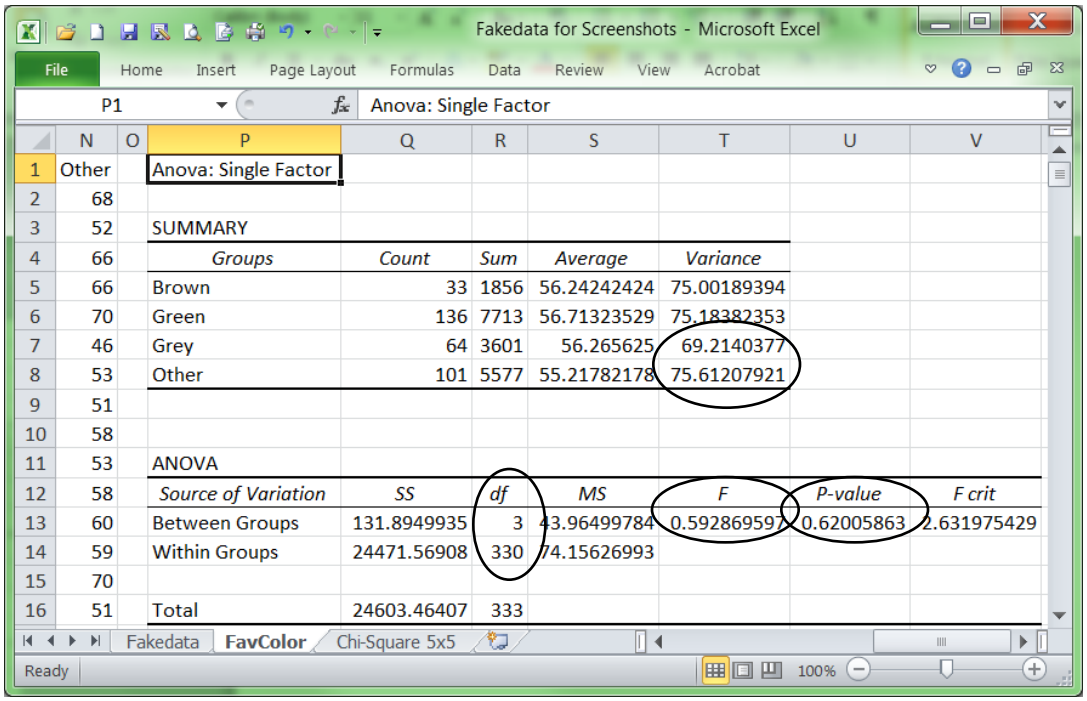


Click the **Input Range:** box and then highlight the four columns with the ages by FavColor. Click the **Labels in first row** box. Click the **Output Range** radio button, and then click in the box to the right of the **Output Range** and click cell P1 on your spreadsheet. Click the **OK** button.

Note: If you do not click in that Output Range box first, then you will have to reselect the data in the Input Range box (order matters here).



The following output should appear.



d. We have not received numerical summaries by group for free (the Summary section of the output). Based on this output, it appears that the mean height is lower for the group whose favorite color is "Other," ANOVA will tell us if this is a significant difference.

e. ANOVA assumes that the sample is large enough and random, and that the standard deviations are nearly equal. We will assume the first two are the case,

but what about the standard deviations? A rule of thumb for checking the standard deviation is that the largest standard deviation is less than twice the smallest. For variances, the rule is less than four times as much. In this case the smallest variance is from Grey group and is 69.21, while the largest is 75.61. In this case inspection tells us that clearly the largest is less than twice the smallest, but we should be prepared to do a calculation for this in the future. We can do this by picking a blank cell and entering the command

= 4*T7
(You can get T7 by clicking on the cell of the lowest variance).

This will return 276.86. Include a screenshot your output at this point in your report.

Now that we have checked the assumptions and the summary statistics, we are ready to interpret the ANOVA output. Note where the F statistic and degrees of freedom are located in the output. In this case $F_{3,330} = 0.593$ and the p-value is 0.62. What is the conclusion of this hypothesis test? Include your output and justify your answer.

f. Finally, we will learn the appropriate way to summarize the results of ANOVA or any hypothesis test in a report. Should you ever need to do this, you may want to look up APA format for statistical output, where you will find many example of how you should appropriately report statistical output. We will not get into all of the details here on significant figures, but will practice the standard format. The output will be some variation of the following, where the blanks are filled in within the context of your test. Note that when using a test that has a degrees of freedom, you should include those with your output.

A(n) _____name of test_____ was run comparing ____verbal statement of H_o____ and __there was or was not__ significant evidence that ____verbal statement of Ha___, __z, t(df), F(df1, df1), or χ^2(df)= insert value__, p-value = __value__.

So in our case the output would be:

An ANOVA was run comparing the difference between the mean ages of those whose favorite colors were brown, green, grey or other, and there was not significant evidence that at least one mean was different, F(3,330) = 0.593, p-value = 0.62.

You do not need to include anything for this step in your lab report as you will be practicing it later in the lab.

3. **Chi-Square**

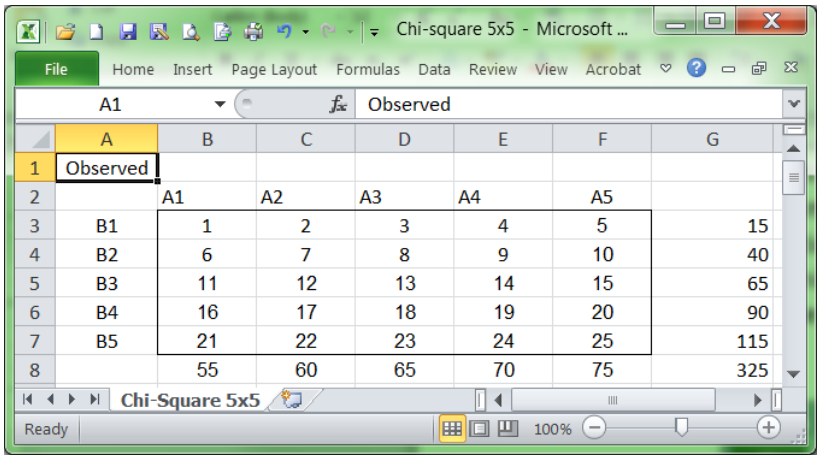
Suppose we would like to know if there is a relationship between Gender and Favorite Color. That is we aim to test the hypotheses:

H_0: There is no relationship between Gender and FavColor,
H_a: There is a relationship between Gender and FavColor.

a. Again, you should first construct the appropriate numerical and graphical summaries for each of these groups. In lab 2 you constructed the appropriate summaries for a C-C relationship. Construct those now and include them in your lab report.

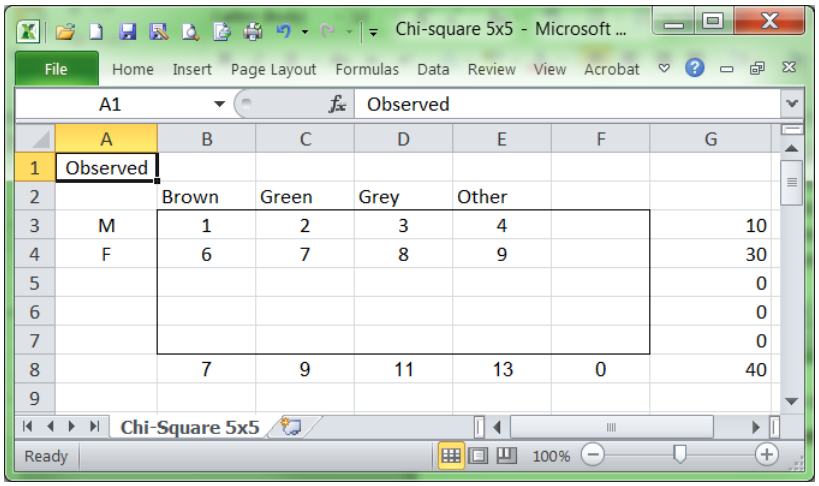
b. Once you have constructed the appropriate summaries, you may notice that it appears that there may be differences in the proportions in each row. We would like to implement a χ^2-test to determine if this is a significant difference.

c. Open the file Chi-square 5x5.xlsx.



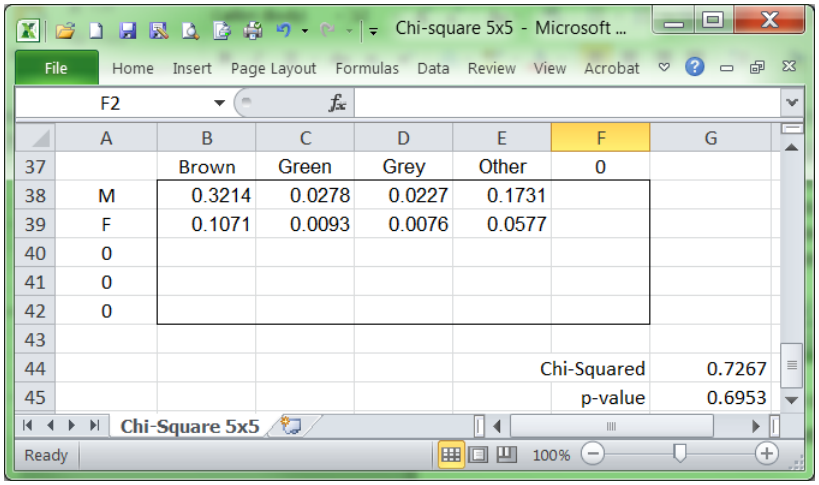
	A	B	C	D	E	F	G
1	Observed						
2		A1	A2	A3	A4	A5	
3	B1	1	2	3	4	5	15
4	B2	6	7	8	9	10	40
5	B3	11	12	13	14	15	65
6	B4	16	17	18	19	20	90
7	B5	21	22	23	24	25	115
8		55	60	65	70	75	325

Type in or paste the Gender by FavColor table values into the Observed section of the sheet. Be sure to delete the entries which you will not be using in that section.



	A	B	C	D	E	F	G
1	Observed						
2		Brown	Green	Grey	Other		
3	M	1	2	3	4		10
4	F	6	7	8	9		30
5							0
6							0
7							0
8		7	9	11	13	0	40
9							

Scroll down to the end of the calculation to find the χ^2 and p-values.



d. Finally, we will practice the formal report format of this output for this test. Use the format specified earlier in this lab to report the conclusions of this hypothesis test.

Application Questions

For this lab you will continue to use the Fakedata set. Answer each of the following questions. Include the appropriate numerical and graphical summary for each and write the results of the test in standard report format.

A1. Is there a significant relationship between gender and the preferred breakfast?

A2. Are there significant differences between the mean pinky lengths of these who prefer the different breakfasts?