HW 2 Spring 2018

1. (2 parts)

(a)For the hair eyes data as given, construct a contingency table and comment on a apparent associations between hair color and eye color.

Problem 1 a

Table of eyecolor by haircolor				
eyecolor	**haircolor**			
Frequency	**fair**	**medium**	**dark**	**Total**
light	688	584	188	1460
medium	343	909	412	1664
dark	98	403	681	1182
Total	1129	1896	1281	4306

As we can see from the contingency table, People have dark eye color tend to have a medium and dark hair color. And people with medium eye color tend to have medium hair color. And people with the light eye color are more likely to have fair and medium hair color. On the other hand, people with dark eye color and fair hair color is quite small compared to other situations. So, the color of people's hair color and eye color tend to be similar(close) to each other.

(b) Perform and comment on appropriate tests of association for the table, and interpret the results.

The FREQ Procedure				
Table of eyecolor by haircolor				
eyecolor	haircolor			
Frequency Expected Cell Chi-Square	fair	medium	dark	Total
light	688 382.8 243.33	584 642.86 5.3894	188 434.34 139.71	1460
medium	343 436.29 19.947	909 732.69 42.429	412 495.03 13.925	1664
dark	98 309.91 144.9	403 520.45 26.506	681 351.64 308.5	1182
Total	1129	1896	1281	4306

Statistics for Table of eyecolor by haircolor

Statistic	DF	Value	Prob
Chi-Square	4	944.6434	<.0001
Likelihood Ratio Chi-Square	4	923.8350	<.0001
Mantel-Haenszel Chi-Square	1	814.7860	<.0001
Phi Coefficient		0.4684	
Contingency Coefficient		0.4242	
Cramer's V		0.3312	

Sample Size = 4306

1b) First, as we can see from the contingency table, the expected cell counts are much greater than 5 and the categories are ordinal, so we can look at the result of Chi-Square, Likelihood Ratio Chi-Square and Mantel-Haenszel Chi-Square. And all tests reject the null hypothesis (P value is less than 0.0001), so there exits some correlation between the hair color and eye color. And Mantel-Haenszel Chi-Square test result tells us that there exits some linear correlation between the hair color and eye color
Second, the number of people with both black eye color and hair color is significant larger than expected; the number of people with black eye color and fair hair color is significant smaller than expected; the number of people with light eye color and fair color is significant larger than expected, so there should be some correlation between these two categorical variables.
The result of cell Chi-Squre tests tells us that the black eye color and hair color contributes the most. And people with black eye color and fair hair color as well as the number of people with light eye color and fair color contribute a lot to the difference between the real number and expected value.

2. (3 parts) Now consider association between the two darker eye colors and hair colors.

(a) Construct a table with the medium values omitted. For the remaining 2x2 table of light and dark eye color and fair and dark hair color, comment on any apparent associations between these eye and hair color groups.

Problem 2 a

Table of eyecolor by haircolor			
eyecolor	**haircolor**		
Frequency **Expected**	**fair**	**dark**	**Total**
light	688 416.03	188 459.97	876
dark	98 369.97	681 409.03	779
Total	786	869	1655

As we can see from the table, every cell value is very different from the expected value. The number of people with light eye color and fair hair color as well as the

number of people with dark eye color and hair color are larger than expected; On the other hand, the number of people with dark eye color and fair hair color as well as the number of people with light eye color and dark hair color are significantly smaller than expected. As a result, there must be come correlation between these two category variables.

(b) Perform and comment on appropriate tests of association for the table, and interpret the results. Compare the results and conclusions for this 2x2 table to those for the full 3x3 table analyzed in part (a).

Problem 2 b

Table of eyecolor by haircolor			
eyecolor	**haircolor**		
Frequency **Expected** **Cell Chi-Square**	**fair**	**dark**	**Total**
light	688 416.03 177.79	188 459.97 160.81	876
dark	98 369.97 199.93	681 409.03 180.83	779
Total	786	869	1655

Statistics for Table of eyecolor by haircolor

Statistic	DF	Value	Prob
Chi-Square	1	719.3494	<.0001
Likelihood Ratio Chi-Square	1	789.6714	<.0001
Continuity Adj. Chi-Square	1	716.7068	<.0001
Mantel-Haenszel Chi-Square	1	718.9147	<.0001

Sample Size = 1655

First, as we can see from the contingency table, the expected cell counts are much greater than 5 and the categories are ordinal, so we can look at the result of Chi-Square, Likelihood Ratio Chi-Square and Mantel-Haenszel Chi-Square. And all tests reject the null hypothesis (P value is less than 0.0001), so there exits some correlation between the hair color and eye color. And Mantel-Haenszel Chi-Square test result tells us that there exits some linear correlation between the hair color and eye color. This part is quite similar to the 3x3 table in part(a).
The test result of 2x2 is consist with the test of 3x3 table.

(c) Consider risks for having fair hair. Test and comment on whether those with light eye color are significantly more likely to have fair hair than are those with dark eye color.

Problem 2 c

Table of haircolor by eyecolor			
haircolor	**eyecolor**		
Frequency Expected	**light**	**dark**	**Total**
fair	688 416.03	98 369.97	786
dark	188 459.97	681 409.03	869
Total	876	779	1655

Statistics for Table of haircolor by eyecolor

Column 1 Risk Estimates						
	Risk	**ASE**	**95% Confidence Limits**		**Exact 95% Confidence Limits**	
Row 1	0.8753	0.0118	0.8522	0.8984	0.8502	0.8976
Row 2	0.2163	0.0140	0.1890	0.2437	0.1894	0.2452
Total	0.5293	0.0123	0.5053	0.5534	0.5049	0.5536

A contingency table of blood pressure status and cholesterol status

Column 1 Risk Estimates						
	Risk	ASE	95% Confidence Limits		Exact 95% Confidence Limits	
Difference	0.6590	0.0183	0.6232	0.6948		
Difference is (Row 1 - Row 2)						

Column 2 Risk Estimates						
	Risk	ASE	95% Confidence Limits		Exact 95% Confidence Limits	
Row 1	0.1247	0.0118	0.1016	0.1478	0.1024	0.1498
Row 2	0.7837	0.0140	0.7563	0.8110	0.7548	0.8106
Total	0.4707	0.0123	0.4466	0.4947	0.4464	0.4951
Difference	-0.6590	0.0183	-0.6948	-0.6232		
Difference is (Row 1 - Row 2)						

Sample Size = 1655

完全解释错了

~~According to the test result, people with light hair color are more likely to have a fair hair color compared to dark hair (0.88 vs 0.21); people with dark hair color are more likely to have a dark hair color compared to have fair hair (0.78 vs 0.12).~~

~~The risk of people with light color have a fair hair color is 0.88(Column 1 Risk Estimates on row 1), while the risk of people with dark color have a fair hair color is 0.12(Column 2 Risk Estimates on row 1). So those with light eye color are significantly more likely to have fair hair than are those with dark eye color. Both column 1 and column 2 risk difference's 95% Confidence Limits doesn't contain 0, therefore, the test result is reliable.~~

A contingency table of blood pressure status and cholesterol status

3. (2 parts)

(a) For the heartbpchol data set, construct a contingency table for comparing blood pressure status and cholesterol status and comment on any apparent relationships between blood pressure and cholesterol statuses.

Problem 3 a

Table of BP_Status by Chol_Status				
BP_Status(Blood Pressure Status)	**Chol_Status(Cholesterol Status)**			
Frequency Expected	**Desirable**	**Borderline**	**High**	**Total**
Optimal	20 14.985	25 25.512	22 26.503	67
Normal	65 54.797	92 93.29	88 96.913	245
High	36 51.218	89 87.198	104 90.584	229
Total	121	206	214	541

As we can see from the contingency table, people have high blood pressure tend to have high Cholesterol, so there might be a positive relationship between these two. People with normal blood pressure tend to have borderline and high Cholesterol. There is no significant preference Cholesterol level for people with optimal blood pressure.

(b) Perform and comment on appropriate tests of association for the table, and interpret the results. What can be said about cholesterol statuses of patients with higher blood pressure statuses?

Problem 3 b

A contingency table of blood pressure status and cholesterol status

Table of BP_Status by Chol_Status

BP_Status(Blood Pressure Status)	Chol_Status(Cholesterol Status)			
Frequency Expected Cell Chi-Square	Desirable	Borderline	High	Total
Optimal	20 14.985 1.6782	25 25.512 0.0103	22 26.503 0.765	67
Normal	65 54.797 1.8999	92 93.29 0.0178	88 96.913 0.8197	245
High	36 51.218 4.5217	89 87.198 0.0372	104 90.584 1.987	229
Total	121	206	214	541

Statistics for Table of BP_Status by Chol_Status

Statistic	DF	Value	Prob
Chi-Square	4	11.7368	0.0194
Likelihood Ratio Chi-Square	4	11.9792	0.0175
Mantel-Haenszel Chi-Square	1	9.9553	0.0016
Phi Coefficient		0.1473	
Contingency Coefficient		0.1457	
Cramer's V		0.1042	

Sample Size = 541

First, as we can see from the contingency table, the expected cell counts are much greater than 5 and the categories are ordinal, so we should so we can look at the result of Chi-Square, Likelihood Ratio Chi-Square and Mantel-Haenszel Chi-Square. And all tests reject the null hypothesis (P value is less than 0.0001), so there exits some correlation between the hair color and eye color. Mantel-Haenszel Chi-Square test result tells us that there exits some linear correlation between the blood pressure and Cholesterol status.

A contingency table of blood pressure status and cholesterol status

Second, we can also tell from the cell Chi-Square that the number of people with high blood pressure and desirable Cholesterol is smaller than expected and it contributes most the difference.

People have high blood pressure tend to have high Cholesterol. Majority of people have high blood pressure have either borderline or high Cholesterol.

4. (2 parts) For the heartbpchol data set, consider a one-way ANOVA model to identify differences between group cholesterol means. The normality assumption is reasonable, so you can proceed without testing normality.

(a) Perform a one-way ANOVA for Cholesterol level with BP Status as the categorical predictor, test any assumptions of the model that should be tested (aside from normality, which you do not need to test), comment on the significance of the model, and the variation described by the model.

Problem 4 a

	Cholesterol		
	Mean	**Std**	**N**
Blood Pressure Status			
Optimal	221.93	39.75	67
Normal	229.03	43.24	245
High	240.57	44.73	229

As we can see from the table, the data is unbalanced data. But it is ok to do one way ANOVA.

Dependent Variable: Cholesterol

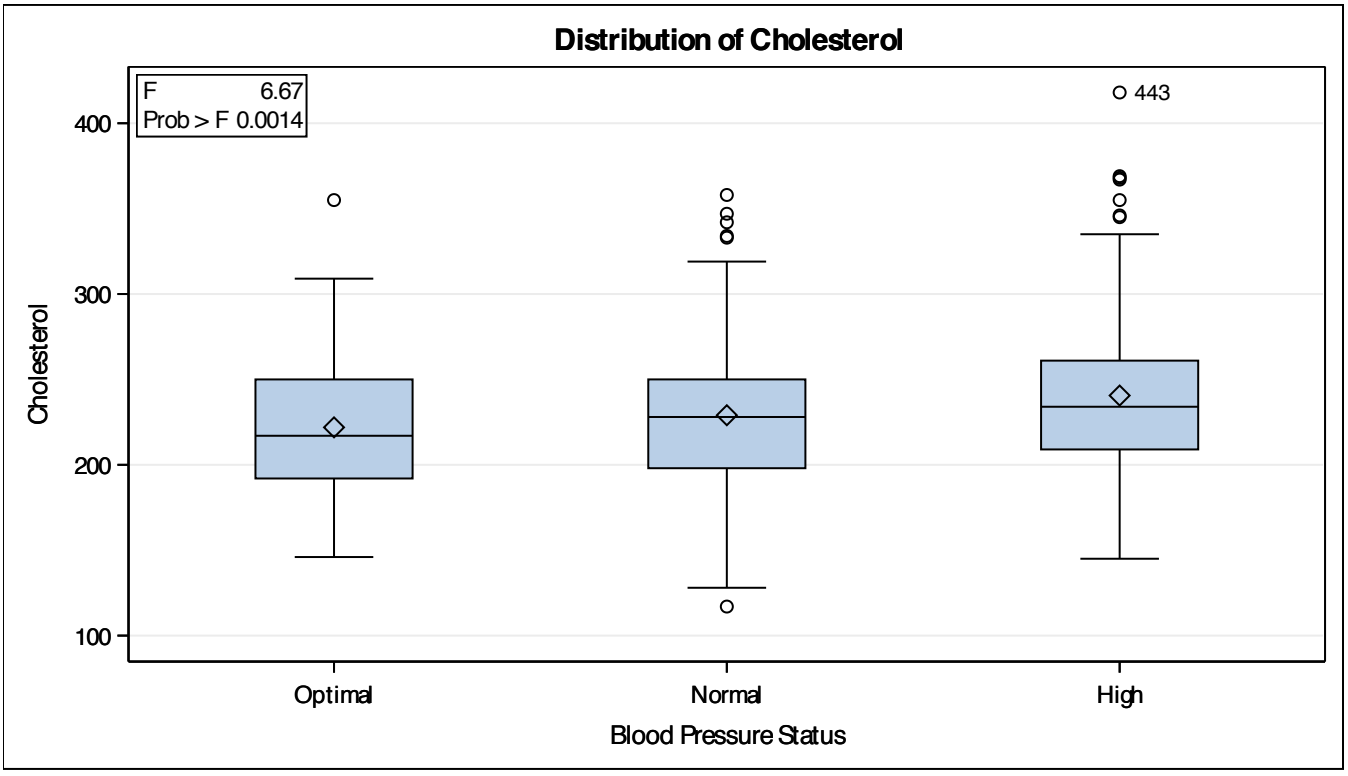
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	25210.845	12605.422	6.67	0.0014
Error	538	1016631.488	1889.650		
Corrected Total	540	1041842.333			

As we can see from the table, the P value is 0.0014 (less than 0.05), so this model is significant.

R-Square	Coeff Var	Root MSE	Cholesterol Mean
0.024198	18.65388	43.47010	233.0351

As we can see from the table, the value of R square is quite small so that this module is not good.

Source	DF	Anova SS	Mean Square	F Value	Pr > F
BP_Status	2	25210.84472	12605.42236	6.67	0.0014

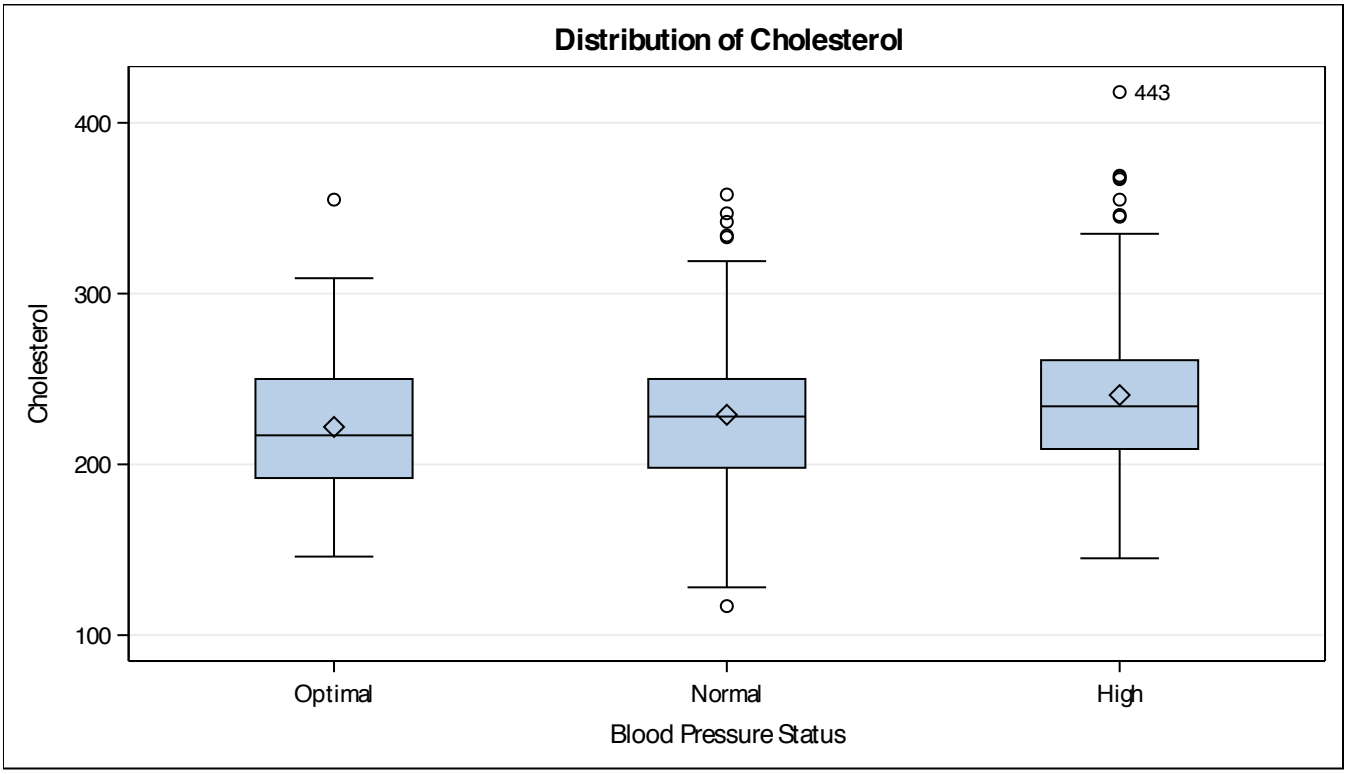


Levene's Test for Homogeneity of Cholesterol Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
BP_Status	2	9964224	4982112	0.56	0.5719
Error	538	4.7912E9	8905491		

As we can see from the table, the P value is 0.5719 (much larger than 0.05), so the variance of Cholesterol in three group is equal.

So, the result of one-way ANOVA is reliable for this dataset.

(b) Comment on any significantly different cholesterol means as determined by the best test for comparing all pairwise differences. Explain what that tells us about differences in cholesterol levels across blood pressure status groups, and comment on how these results compare with the results from part (b) of Exercise 3.



Tukey's Studentized Range (HSD) Test for Cholesterol

Note: This test controls the Type I experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	538
Error Mean Square	1889.65
Critical Value of Studentized Range	3.32371

Comparisons significant at the 0.05 level are indicated by *.**				
BP_Status Comparison	**Difference Between Means**	**Simultaneous 95% Confidence Limits**		
High - Normal	11.543	2.153	20.934	***
High - Optimal	18.647	4.456	32.837	***
Normal - High	-11.543	-20.934	-2.153	***
Normal - Optimal	7.103	-6.982	21.188	
Optimal - High	-18.647	-32.837	-4.456	***
Optimal - Normal	-7.103	-21.188	6.982	

As we can see from the table, the difference means Cholesterol between people with high blood pressure and normal blood pressure is significant; the difference means Cholesterol value of people with high blood pressure and optimal blood pressure is significant. Besides, Simultaneous 95% Confidence Limits doesn't contain 0, therefore, this test result is reliable.

In 3(b), we find people have high blood pressure tend to have high Cholesterol. Majority of people have high blood pressure have either borderline or high Cholesterol. Test in 4(b) specify this correlation. We can find in 4(b) that people with high blood pressure usually have higher level of Cholesterol compared to the people with normal or optimal blood pressure. Therefore, there exist some positive relationship between the blood pressure and Cholesterol level.