STAT448 - Advanced Data Analysis

Homework 2

Name: Xinyan Yang

Exercise 1 Solution:

(a).

Table of agegrp by response			
agegrp	**response**		
Frequency Expected	**no**	**yes**	**Total**
older	225 197.99	253 280.01	478
younger	434 461.01	679 651.99	1113
Total	659	932	1591

According to the contingency table above, we can find that the counts of the response "yes" are greater than the counts of the "no" no matter in older group or younger group.

However, the younger group are more likely to answer "yes" since the "yes" counts are largely greater than the "no" counts for the younger group while the "yes" counts only slightly greater than the "no" counts for the older group.

(b)

Statistics for Table of agegrp by response

Statistic	DF	Value	Prob
Chi-Square	1	8.9916	0.0027
Likelihood Ratio Chi-Square	1	8.9394	0.0028
Continuity Adj. Chi-Square	1	8.6618	0.0032
Mantel-Haenszel Chi-Square	1	8.9860	0.0027
Phi Coefficient		0.0752	
Contingency Coefficient		0.0750	
Cramer's V		0.0752	

According to the association test results, since the expected frequency for each cell are all greater than 5, we can use the Chi-Square association test. The P value is 0.0027, therefore under the significance level of 5%, we should reject the null hypothesis and conclude that ***the agegroup and the response is not independent****,* there exists some association between them.

Exercise 2 Solution:

(a)

Table of Greater100 by South			
Greater100	**South**		
Frequency Expected	**no**	**yes**	**Total**
no	20 21.766	13 11.234	33
yes	11 9.234	3 4.766	14
Total	31	16	47

From the contingency table above, we can find that within the group with crime rates greater than 100 crimes per million population, the *no-south cell has apparently more counts than the yes-south cell*.And the real count of yes-south cell is smaller than the expected frequency of this cell while the real count of no-south cell is greater than the expected frequency.

(b)

Statistics for Table of Greater100 by South

Statistic	DF	Value	Prob
Chi-Square	1	1.4130	0.2346
Likelihood Ratio Chi-Square	1	1.4841	0.2231
Continuity Adj. Chi-Square	1	0.7261	0.3941
Mantel-Haenszel Chi-Square	1	1.3829	0.2396
Phi Coefficient		-0.1734	
Contingency Coefficient		0.1708	
Cramer's V		-0.1734	
WARNING: 25% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

Fisher's Exact Test	
Cell (1,1) Frequency (F)	20
Left-sided Pr <= F	0.1988
Right-sided Pr >= F	0.9400
Table Probability (P)	0.1388
Two-sided Pr <= P	0.3211

Conclusion: Since the expected frequency for the crime rates greater than 100 in the south is smaller than 5, we should *use the Fisher's Exact Test* to perform the association test. The P-value of this test is largely greater than 0.05, therefore we should accept the null hypothesis under the significance of 5% ***and conclude that the crime rates greater than 100 and whether the state is in south is independent***.

(c)

Table of South by Greater100			
South	**Greater100**		
Frequency	**no**	**yes**	**Total**
no	20	11	31
yes	13	3	16
Total	33	14	47

Column 2 Risk Estimates						
	Risk	**ASE**	**(Asymptotic) 95% Confidence Limits**		**(Exact) 95% Confidence Limits**	
Row 1	0.3548	0.0859	0.1864	0.5233	0.1923	0.5463
Row 2	0.1875	0.0976	0.0000	0.3787	0.0405	0.4565
Total	0.2979	0.0667	0.1671	0.4286	0.1734	0.4489
Difference	0.1673	0.1300	-0.0875	0.4222		
Difference is (Row 1 - Row 2)						

With respect to the tables above, the row is whether the state is in south and the column is whether the crime rates are greater than 100. To see if Southern states have a significantly higher probability of crime rates greater than 100 than other states do, *we only have to see the risk estimates of column 2 on the rows*. Since the risk difference confidence interval contains 0, we cannot reject the null hypothesis ***and conclude that there is no significant difference*** between the crime rates greater than 100 of southern states and that of non-southern states.

Exercise 3 Solution:

(a)

Table of mcvgrp by drinkgroup						
mcvgrp	**drinkgroup**					
Frequency Expected	**1**	**2**	**3**	**4**	**5**	**Total**
0	65 48.835	23 21.704	41 36.73	10 27.965	5 8.7652	144
1	52 68.165	29 30.296	47 51.27	57 39.035	16 12.235	201
Total	117	52	88	67	21	345

Statistics for Table of mcvgrp by drinkgroup

Statistic	DF	Value	Prob
Chi-Square	4	32.7546	<.0001
Likelihood Ratio Chi-Square	4	35.5619	<.0001
Mantel-Haenszel Chi-Square	1	24.7026	<.0001
Phi Coefficient		0.3081	
Contingency Coefficient		0.2945	
Cramer's V		0.3081	

From the test results above, we should choose the Chi-Square Test since the expected frequencies are all greater than 5. The P-value of Chi-Square test is <0.0001 so that we should reject the null hypothesis under the significance level of 5% and conclude that the mcv and drinkgroup is not independent, which means that they have some association.

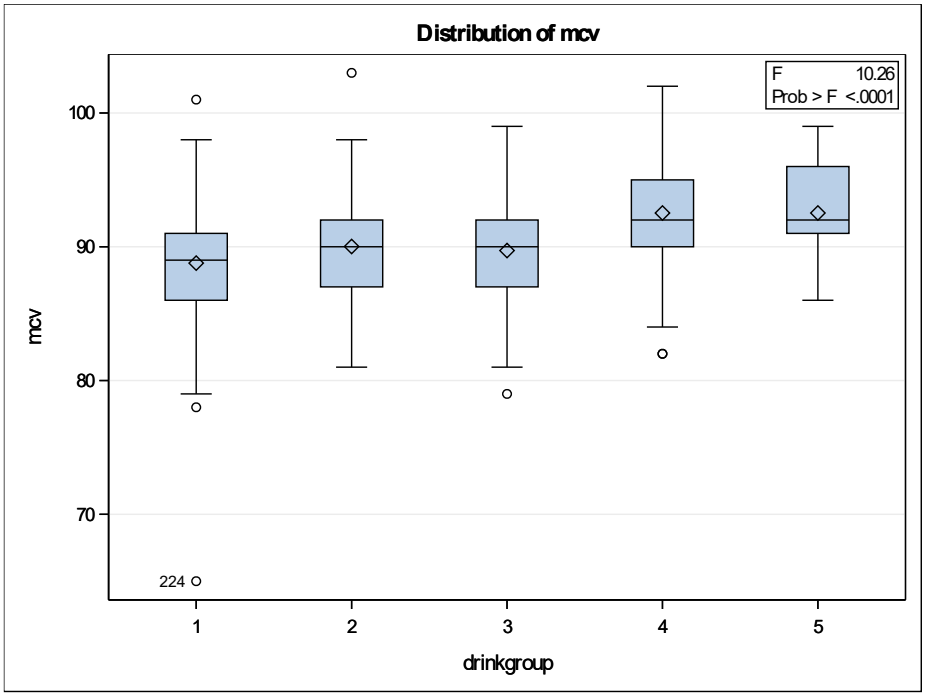
(b)

Dependent Variable: mcv

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	733.176652	183.294163	10.26	<.0001
Error	340	6073.055232	17.861927		
Corrected Total	344	6806.231884			

R-Square	Coeff Var	Root MSE	mcv Mean
0.107721	4.687627	4.226337	90.15942

Source	DF	Anova SS	Mean Square	F Value	Pr > F
drinkgroup	4	733.1766522	183.2941630	10.26	<.0001



Levene's Test for Homogeneity of mcv Variance ANOVA of Squared Deviations from Group Means					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
drinkgroup	4	1940.3	485.1	0.33	0.8587
Error	340	501897	1476.2		

From the one-way ANOVA model result above, we can notice that the P-Value for the model is <0.001, therefore we reject the null hypothesis at the significance level of 5% and think the model is statically significant, which means that the model is useful.

The R-Square for the model is 0.107722, therefore about 10.77% of the variation can be described by the model.

According to the hov test result, since the p-value= 0.8687 is not significant, we cannot reject the null hypothesis, which means that the equal variance assumption can be trusted.

(c)

Two model results both indicate that there are some association between the mcv and drinkgroup. They give the same conclusion.

(d)

Tukey's Studentized Range (HSD) Test for mcv

Alpha	0.05
Error Degrees of Freedom	340
Error Mean Square	17.86193
Critical Value of Studentized Range	3.87844

Comparisons significant at the 0.05 level are indicated by *.**				
drinkgroup Comparison	**Difference Between Means**	**Simultaneous 95% Confidence Limits**		
5 - 4	0.0014	-2.8973	2.9001	
5 - 2	2.5046	-0.4922	5.5014	
5 - 3	2.8079	-0.0070	5.6228	
5 - 1	3.7460	0.9991	6.4929	***
4 - 5	-0.0014	-2.9001	2.8973	
4 - 2	2.5032	0.3611	4.6453	***
4 - 3	2.8065	0.9272	4.6858	***
4 - 1	3.7446	1.9688	5.5204	***
2 - 5	-2.5046	-5.5014	0.4922	
2 - 4	-2.5032	-4.6453	-0.3611	***
2 - 3	0.3033	-1.7240	2.3307	
2 - 1	1.2415	-0.6903	3.1732	
3 - 5	-2.8079	-5.6228	0.0070	
3 - 4	-2.8065	-4.6858	-0.9272	***
3 - 2	-0.3033	-2.3307	1.7240	
3 - 1	0.9381	-0.6974	2.5736	
1 - 5	-3.7460	-6.4929	-0.9991	***
1 - 4	-3.7446	-5.5204	-1.9688	***
1 - 2	-1.2415	-3.1732	0.6903	
1 - 3	-0.9381	-2.5736	0.6974	

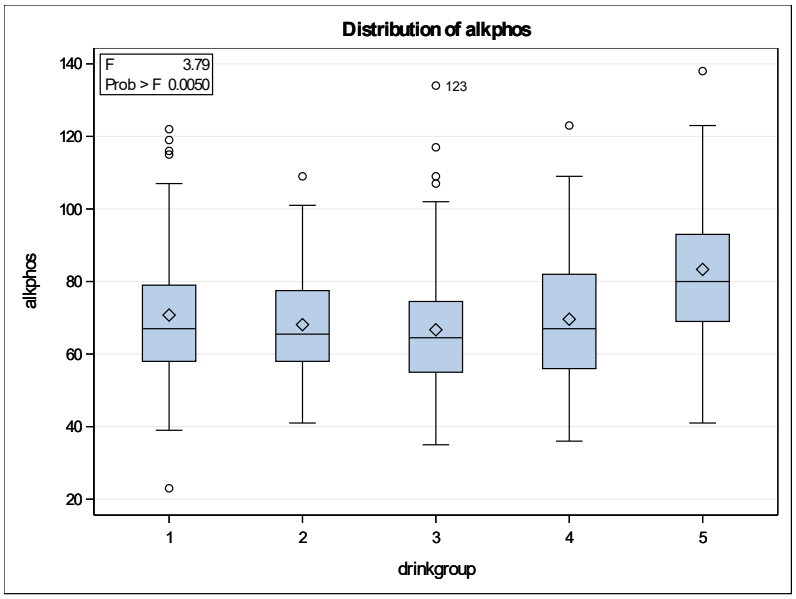
From the tukey’s test results, we can find that the mcv of group 4 is significantly greater than the group 1, 2 and 3 and the mcv of group 5 is significantly greater than group1.

(e)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	4945.6283	1236.4071	3.79	0.0050
Error	340	110857.5021	326.0515		
Corrected Total	344	115803.1304			

R-Square	Coeff Var	Root MSE	alkphos Mean
0.042707	25.84372	18.05690	69.86957

Source	DF	Anova SS	Mean Square	F Value	Pr > F
drinkgroup	4	4945.628301	1236.407075	3.79	0.0050



Levene's Test for Homogeneity of alkphos Variance ANOVA of Squared Deviations from Group Means					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
drinkgroup	4	1065624	266406	0.96	0.4293
Error	340	94305238	277368		

From the one-way ANOVA model result above, we can notice that the P-Value for the model is 0.005, therefore we reject the null hypothesis at the significance level of 5% and think the model is statically significant, which means that the model is useful.

The R-Square for the model is 0.042707, therefore about 4.27% of the variation can be described by the model.

According to the hov test result, since the p-value= 0.4293 is not significant, we cannot reject the null hypothesis, which means that the equal variance assumption can be trusted.

(f)

Alpha	0.05
Error Degrees of Freedom	340
Error Mean Square	326.0515
Critical Value of Studentized Range	3.87844

Comparisons significant at the 0.05 level are indicated by *.**				
drinkgroup Comparison	**Difference Between Means**	**Simultaneous 95% Confidence Limits**		
5 - 1	12.573	0.837	24.309	***
5 - 4	13.721	1.337	26.106	***
5 - 2	15.218	2.414	28.022	***
5 - 3	16.629	4.602	28.656	***
1 - 5	-12.573	-24.309	-0.837	***
1 - 4	1.149	-6.438	8.736	
1 - 2	2.645	-5.608	10.899	
1 - 3	4.056	-2.931	11.044	
4 - 5	-13.721	-26.106	-1.337	***
4 - 1	-1.149	-8.736	6.438	
4 - 2	1.497	-7.656	10.649	
4 - 3	2.907	-5.122	10.937	
2 - 5	-15.218	-28.022	-2.414	***
2 - 1	-2.645	-10.899	5.608	
2 - 4	-1.497	-10.649	7.656	
2 - 3	1.411	-7.251	10.073	
3 - 5	-16.629	-28.656	-4.602	***
3 - 1	-4.056	-11.044	2.931	
3 - 4	-2.907	-10.937	5.122	
3 - 2	-1.411	-10.073	7.251	

According to the Tukey's test results for alkphos across different groups, we can find that the alkphos mean of group 5 is significantly higher than that of the other 4 groups, which is different from the mcv case.