Exercise 1

a) The following contingency table shows the frequencies for crime rates greater than 100 crimes known to police per million population. We see the observed percentage of states with crime levels above that level outside the South is almost twice that for states in the South, so there may be an association of higher crime rates with states outside the South, but hypothesis tests will be needed to see if that association is significant.

Table of South by Greater100			
South	**Greater100**		
Frequency Expected Row Pct	**no**	**yes**	**Total**
no	20 21.766 64.52	11 9.234 35.48	31
yes	13 11.234 81.25	3 4.766 18.75	16
Total	33	14	47

b) The expected counts are almost large enough to trust the asymptotic chi-square tests, but we should use Fisher's exact test to be safe here. None of the p-values for Fisher's exact test are less than .05, so we see no significant association between whether or not a state is in the South and whether or not that state has a crime rate greater than 100 known crimes per million population.

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

c) Column 1 will give us the risk difference between the non-South states and the South states for having a crime rate less than 100. This will be the same as the risk difference between the South and the non-South for having a crime rate greater than 100. We see that the estimated difference is -.1673 with a confidence interval of (-.4222, .0875). Thus the South is estimated to have a lower risk of having a crime rate over 100, but the estimate is not significantly different from 0, so we conclude there is no significant difference in the risk of having a crime rate over 100 in the South vs. not being in the South.

Column 1 Risk Estimates						
	Risk	**ASE**	**(Asymptotic) 95% Confidence Limits**		**(Exact) 95% Confidence Limits**	
Row 1	0.6452	0.0859	0.4767	0.8136	0.4537	0.8077
Row 2	0.8125	0.0976	0.6213	1.0000	0.5435	0.9595
Total	0.7021	0.0667	0.5714	0.8329	0.5511	0.8266
Difference	-0.1673	0.1300	-0.4222	0.0875		
Difference is (Row 1 - Row 2)						

Exercise 2

a) The following contingency table shows the frequencies for eye and hair colors in order of increasing darkness. We see that the percentages are largest on the diagonal and seem to decrease away from the diagonal, so there appears to be an association and there may also be an ordinal association between the categories with the lightness of eye color largely matching up with lightness of hair color.

Table of eyecolor by haircolor				
eyecolor	**haircolor**			
Frequency Expected Row Pct	**fair**	**medium**	**dark**	**Total**
light	688 382.8 47.12	584 642.86 40.00	188 434.34 12.88	1460
medium	343 436.29 20.61	909 732.69 54.63	412 495.03 24.76	1664
dark	98 309.91 8.29	403 520.45 34.09	681 351.64 57.61	1182
Total	1129	1896	1281	4306

The sample size is large enough to trust chi-square tests. We should not use Fisher's exact test here because it is unnecessary and would take a lot of computing time and memory to calculate. Pearson's chi-square and the likelihood ratio chi-square are both extremely significant indicating a significant association between eye color and hair color. The Mantel-Haenzel test is appropriate here because the variables are ordinal. It is highly significant indicating that there is a significant linear association between eye and hair color.

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	

b) The following contingency table shows the frequencies for eye and hair colors in order of increasing darkness for just the lighter two levels. We see that the percentages again are largest on the diagonal and seem to decrease away from the diagonal, so again there appears to be an association and that association appears to be ordinal with lightness of eye color largely matching up with lightness of hair color.

Table of eyecolor by haircolor			
eyecolor	**haircolor**		
Frequency **Expected** **Row Pct**	**fair**	**medium**	**Total**
light	688 519.58 54.09	584 752.42 45.91	1272
medium	343 511.42 27.40	909 740.58 72.60	1252
Total	1031	1493	2524

The hypothesis tests agree. The Pearson and likelihood ratio tests are highly significant indicating a statistically significant association between eye color and hair color. Mantel-Haenszel is also highly significant indicating a linear association. The conclusions are pretty much the same. The only real difference is a slight difference in the magnitude of the association. Cramer's V was slightly higher in the 3x3 case than in the 2x2 case.

Statistic	DF	Value	Prob
Chi-Square	1	186.0462	<.0001
Likelihood Ratio Chi-Square	1	188.8613	<.0001
Continuity Adj. Chi-Square	1	184.9431	<.0001
Mantel-Haenszel Chi-Square	1	185.9725	<.0001
Phi Coefficient		0.2715	
Contingency Coefficient		0.2620	
Cramer's V		0.2715	

c) To compare risks of fair hair given eye color, we can use the column 1 risk differences directly. The estimated difference for risk of fair hair between light and medium eye color is .2669 with a confidence interval of (.2300, .3038). The confidence interval does not contain 0, so the difference is significant and we conclude that when comparing just the two lightest levels of eye and hair color, those with light eye colors have a greater probability of having fair hair than those with medium eye colors.

Column 1 Risk Estimates						
	Risk	ASE	(Asymptotic) 95% Confidence Limits		(Exact) 95% Confidence Limits	
Row 1	0.5409	0.0140	0.5135	0.5683	0.5130	0.5685
Row 2	0.2740	0.0126	0.2493	0.2987	0.2494	0.2996
Total	0.4085	0.0098	0.3893	0.4277	0.3892	0.4280
Difference	0.2669	0.0188	0.2300	0.3038		
Difference is (Row 1 - Row 2)						

Exercise 3

a) The following are the ANOVA tables and goodness of fit statistics for modeling cholesterol as a function of blood pressure level.

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			

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

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

b) Levene's test is highly insignificant with a p-value of .5719, so we have no reason to reject the hypothesis of equal variance. From the ANOVA tables in part a, we can see that overall model is statistically significantly better than an error only model and that blood pressure status explains more variation than we would expect due solely to random normal error. However, the R-Square value of .024 indicates that only 2.4% of the variation in cholesterol is described by this model. This model will not be very useful if we want to predict cholesterol level.

Note: The quality of the model should not be too surprising because the blood pressure status variable bins a continuous value so a lot of the variation in blood pressure is washed out by the binning. For instance, if blood pressures just above or just below a bin cutoff have very similar cholesterol levels, there would be less group difference in general because there are more similar values in multiple bins.

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		

c) Looking at Tukey tests to compare cholesterol levels between blood pressure groups, we see significant differences between the high blood pressure group and the other two groups, but no significant difference between the normal and optimal blood pressure groups. We estimate that the high blood pressure group on average has cholesterol 11.543 higher than the normal group with a confidence interval of (2.153, 20.934) and the high group has cholesterol on average 18.647 higher than the optimal blood pressure group with a confidence interval of (4.456, 32.837).

Tukey's Studentized Range (HSD) Test for Cholesterol

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	