Project 6: Categorical Data

Name:

The following SAS code may be helpful for this assignment:

Input statement for categorical data (example):

```
data hair;
input sex $ color $ number;
cards;
male black 32
male brown 43
male blonde 16
male red 9
female black 55
female brown 65
female blonde 64
female red 16
;
run;
```

Chi-Square Test for 1xk (uniform):

```
proc freq data=dataset;
weight count;
table var/chisq;
run;
```

Chi-Square Test for 1xk (known p):

```
proc freq data=dataset;
weight count;
table var/ nocum chisq testp=(50 25 10 15);
run;
```

Chi-Square Test for rxc:

```
proc freq data=dataset;
weight count;
table var1*var2/nocol norow chisq;
run;
```

McNemar's Test for Paired Data:

```
proc freq data=dataset;
weight count;
table var1*var2/agree;
run;
```

A genetic researcher is attempting to determine whether a set of flowers is from a particular genetic strain that produces a 75%/25% ratio of green to yellow flowers. Use α=0.05 for all tests. Consider the following data:

Flower Color	
Yellow	Green
16	84

1. Test to determine whether this sample comes from a population with a distribution of 25% yellow and 75% green flowers.
2. Include your hypotheses, SAS output, SAS Code, and conclusions.

H_0: This sample does come from a population with a distribution of 25% yellow and 75% green

H_1: This sample does not come from a population with a distribution of 25% yellow and 75% green

Reject H_0 if p-val < 0.05

p-val: 0.0377

Since the p-val is less 0.05, so we reject the H_0. This means that there is enough evidence to suggest that the sample does not come from a population of 25% yellow and 75% green

Code:
```
data two;
input color $ count @@;
cards;
yellow 16 green 84
;
proc freq data=two;
weight count;
tables color/nocum chisq testp=(75 25);
run;
```

color	Frequency	Percent	Test Percent
green	84	84.00	75.00
yellow	16	16.00	25.00

Chi-Square Test for Specified Proportions	
Chi-Square	4.3200
DF	1
Pr > ChiSq	0.0377

Each of 126 of a certain mammal species was placed in an enclosure with 6 types of food. The frequencies with which the animals chose each of the foods follows.

Food Choice	Count
A	13
B	26
C	31
D	14
E	28
F	14

1. Test to for a food preference among the mammals (uniform distribution).
2. Include your hypotheses, SAS output, SAS Code, and conclusions.

H_0: There is not a food preference among animals (uniform distribution)
H_1: There is a food preference among animals (no uniform distribution)
We reject H_0 if p-val < 0.05
p-val: 0.0068
Since the p-val is less than 0.05, then we reject H_0. This means that there is enough evidence to suggest that there is a food preference among animals – therefore, there is not a uniform distribution.

Code: data food;
input FoodType $ count @@;
cards;
a 13 b 26 c 31 d 14 e 28 f 14

```
;
proc freq data=food;
weight count;
table FoodType/chisq;
run;
```

FoodType	Frequency	Percent	Cumulative Frequency	Cumulative Percent
a	13	10.32	13	10.32
b	26	20.63	39	30.95
c	31	24.60	70	55.56
d	14	11.11	84	66.67
e	28	22.22	112	88.89
f	14	11.11	126	100.00

Chi-Square Test for Equal Proportions	
Chi-Square	16.0000
DF	5
Pr > ChiSq	0.0068

Consider the Fisher's exact test with the “famous” tea tasting example! In a summer tea-party in Cambridge, England, a lady claimed to be able to discern, by taste alone, whether a cup of tea with milk had the tea poured first or the milk poured first.
An experiment was performed by Sir R.A. Fisher himself, then and there, to see if her claim is valid. Eight cups of tea are prepared and presented to her in random order. Four had the milk poured first, and four had the tea poured first. The lady tasted each one and rendered her opinion. The results are summarized in a 2x2 table.

	Lady Says		
Poured First	Tea First	Milk First	Total
Tea	3	1	4
Milk	1	3	4
Total	4	4	8

1. Use Fisher's exact test to determine if there is a difference between the two proportions.
2. Include your hypotheses, SAS output, SAS Code, and conclusions.
   H_0: There is no difference between the two proportions
   H_1: There is a difference between the two proportions
   We reject H_0 if p-val is less than 0.05

p-val: $\frac{4!4!4!4!}{8!3!3!} = 0.2286$

Since the p-val is not less than 0.05, then we fail to reject the H_0. This means that there is not enough evidence to suggest a difference between the two proportions.

This is code that I did first, but realized that I had to do something different. Here it is just in case

Code: data three;
input First $ Drink $ Count @@;
cards;
tea tea 3
tea milk 1
milk tea 1
milk milk 3
;
proc freq data=three;
weight count;
table First*Drink/ nocol norow chisq;
run;

Statistic	DF	Value	Prob
Chi-Square	1	2.0000	0.1573
Likelihood Ratio Chi-Square	1	2.0930	0.1480
Continuity Adj. Chi-Square	1	0.5000	0.4795
Mantel-Haenszel Chi-Square	1	1.7500	0.1859
Phi Coefficient		0.5000	
Contingency Coefficient		0.4472	
Cramer's V		0.5000	
WARNING: 100% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

The following data are frequencies of skunks found with and without rabies in two different geographic locations.

	Rabies		
Area	With	Without	Total
East	14	29	43
West	12	38	50
Total	26	67	93

1. Test whether the instances of rabies are different across the areas.
2. Include your hypotheses, SAS output, SAS Code, and conclusions.

H_0: There is no difference between the instance of rabies
H_1: There is a difference between the instance of rabies
We reject H_0 if p-val < 0.05
p-val: 0.3592
Since our p-val is not less than 0.05, we fail to reject the H_0. This means that there is no enough evidence to suggest a difference between the instance of rabies.

Code:
```
data four;
input area $ with $ count @@;
cards;
east yes 14
east no 29
west yes 12
west no 38
;
proc freq data=four;
weight count;
table area*with/ nocol norow chisq;
run;
```

Statistic	DF	Value	Prob
Chi-Square	1	0.8407	0.3592
Likelihood Ratio Chi-Square	1	0.8392	0.3596
Continuity Adj. Chi-Square	1	0.4695	0.4932
Mantel-Haenszel Chi-Square	1	0.8316	0.3618
Phi Coefficient		-0.0951	
Contingency Coefficient		0.0946	
Cramer's V		-0.0951	

Researchers examined 150 men with each of three types of cancer and recorded their blood types. The data are below:

	Blood Type			
Cancer	**O**	**A**	**B**	**AB**
Colon	61	65	18	6
Lung	69	57	15	9
Prostate	73	60	12	5

1. Test for an association between cancer and blood type.
2. Include your hypotheses, SAS output, SAS Code, and conclusions.

H_0: There is an association between cancer and blood type
H_1: There is not an association between cancer and blood type
We reject H_0 if p-val < 0.05
p-val: 0.6575
Since our p-val is more than 0.05, then we fail to reject the H_0. This means that there is evidence of an association between cancer and blood type.

Code:
```
data five;
input cancer $ BloodType $ count @@;
cards;
colon O 61
colon A 65
colon B 18
colon AB 6
lung O 69
lung A 57
lung B 15
lung AB 9
prostate O 73
prostate A 60
prostate B 12
prostate AB 5
;
proc freq data=five;
weight count;
table cancer*BloodType/ nocol norow chisq; run;
```

Statistic	DF	Value	Prob
Chi-Square	6	4.1419	0.6575
Likelihood Ratio Chi-Square	6	4.1202	0.6604
Mantel-Haenszel Chi-Square	1	0.9015	0.3424
Phi Coefficient		0.0959	
Contingency Coefficient		0.0955	
Cramer's V		0.0678	