1. (30 pts) Children of ages 0-15 were classified according to whether they carried Streptococcus pyogenes and according to the size of their tonsils.

	Tonsil size		
	Normal	Slightly enlarged	Very enlarged
Carrier	19	29	24
Non-carrier	497	560	269

(a) Analyze this table in an appropriate manner assuming that we are dealing with nominal variables, and report relevant statistics (e.g., X^2 and/or G^2) and your conclusion.

Solutions:

R Code/Output:

```
> tonsil=matrix(c(19,497,29,560,24,269),ncol=3,
dimnames=list(Carrier=c('Carrier',"non-carrier"),
Size=c('Normal','Slightly Enlarged','Very Enlarged')))
> tonsil
             Size
Carrier       Normal Slightly Enlarged Very Enlarged
  Carrier         19                29            24
  non-carrier    497               560           269
>
> # Pearson's Chi-squared test with
Yates' continuity correction
> result<-chisq.test(tonsil)
> result

Pearson's Chi-squared test

data:  tonsil
X-squared = 7.8848, df = 2, p-value = 0.0194

>
>
> ###Pearson's Chi-squared test  withOUT Yates' continuity correc
> result<-chisq.test(tonsil, correct=FALSE)
> result

Pearson's Chi-squared test

data:  tonsil
```

```
X-squared = 7.8848, df = 2, p-value = 0.0194


> ### Likelihood Ratio Chi-Squared Statistic
> G2=2*sum(tonsil*log(tonsil/result$expected))
> G2
[1] 7.320928
> pvalue=1-pchisq(2*sum(tonsil*log(tonsil/result$expected)),
df=2)
> pvalue
[1] 0.02572057
```

SAS Code/Output:

```
data tonsil;
input carrier $ tonsil $ count;
datalines;
car norm 19
car slt 29
car very 24
nocar norm 497
nocar slt 560
nocar very 269
;

proc freq;
weight count;
tables carrier*tonsil / chisq expected deviation cmh cellchi2
measures;
run;
```

Statistic	DF	Value	Prob
Chi-Square	2	7.8848	0.0194
Likelihood Ratio Chi-Square	2	7.3209	0.0257
Mantel-Haenszel Chi-Square	1	7.1876	0.0073
Phi Coefficient		0.0751	
Contingency Coefficient		0.0749	
Cramer's V		0.0751	

Let H_0 : be the null hypotheses that the independence model holds and let H_a : be the alternative hypothesis that the independent model is not valid. From the SAS/R output the p-values of the X^2 and G^2 are 0.02 and 0.026 respectively. So, at the 0.05 significant level we reject (p-value$<$ 0.05) the null hypothesis of independence. Thus, carrying Streptococcus pyogene is related with tonsil size.

(b) How many measures of association are needed to describe the table's departure from independence? Provide estimates and intervals for the relative risks, and interpret the results. Why are you able to make inference about relative risks?

Solutions: It makes sense to let the response Y to be the categorical variable for tonsil size and the explanatory variable X be the categorical variable for carrying Streptococcus pyogene. We are interested in studying the effect of carrying Streptococcus pyogene on tonsil size. There are $(I-1)(J-1) = 2$ difference in proportions, relative risks and odds ratios. We are able to make inference about relative risks because the sampling scheme is multinomial; children were classified according to carrier and tonsil size rather than fixing a number with each size tonsil.

The risk of slightly enlarged tonsils versus normal tonsils:

	slightly enlarged	Normal
Carrier	29	19
Non-carrier	560	497

and the relative risk

	estimate	95% interval
relative risk	1.14	(0.90, 1.44)

The risk of very enlarged tonsils versus normal tonsils:

	very enlarged	Normal
Carrier	24	19
Non-carrier	269	497

and the relative risk

	estimate	95% interval
relative risk	1.59	(1.20, 2.11)

The risk of very enlarged tonsils versus slightly enlarged tonsils:

	very enlarged	slightly
Carrier	24	29
Non-carrier	269	560

and the relative risk is

	estimate	95% interval
relative risk	1.40	(1.02, 1.91)

(c) Find an appropriate partitioning of the chi-square for this problem. Give the subtables, and show that your partitioning works. Did you learn anything more, inference wise, in comparison to part (a).

Solutions:

For partitioning of the table, you can do, for example, "Normal/Slightly enlarged" vs "Very enlarged". You can then analyze the relationship between "Normal" and "Slightly enlarged" as a 2 by 2 table. These two individual Chi-squares tests both have degree of freedom one and you will see that these two Chi-squares add up to the original Chi-squares total. You can also analyze "Normal" vs "Enlarged", etc.

The risk of having enlarged tonsils versus normal tonsils:

	Enlarged	Normal
Carrier	53	19
Non-carrier	829	497

and the relative risk and odds ratio is

	estimate	95% interval
odds ratio	1.67	(0.98, 2.86)
relative risk	1.18	(1.02, 1.36)

The risk of having very enlarged tonsils versus normal or slightly enlarged tonsils:

	very enlarged	Normal/slightly
Carrier	24	48
Non-carrier	269	1057

and the relative risk and odds ratio is

	estimate	95% interval
odds ratio	1.96	(1.18, 3.27)
relative risk	1.64	(1.17, 2.32)

SAS Code

```
data partition1;
input carrier $ tonsil $ count;
datalines;
carrier normal 19
```

```
carrier senlarged 29
non-carrier normal 497
non-carrier senlarged 560
;
run;
proc freq;
weight count;
tables carrier*tonsil/ chisq expected;

data partition2;
input  carrier $ tonsil $ count;
datalines;
carrier Senlarged 29
carrier Venlarged 24
non-carrier Senlarged 560
non-carrier Venlarged 269
;
run;
proc freq;
weight count;
tables carrier*tonsil/ chisq expected;
```

(d) Now consider 'tonsile size' to be an ordinal variable. Re-run the analysis, if necessary, and report the relevant statistics, your conclusions and compare to what you got in parts (a) and (b). If you think it's not necessary to re-run the analysis, then explain why that's the case.

Solutions:

The MH trend statistic with integer scoring mid-rank scoring $M^2 = 6.83$ with $df = 1$ and p-value 0.009. We infer that there is a strong linear association between Streptococcus pyogenes and the size of tonsil. The M^2 test gave smaller p-values than X^2. When a positive or negative trend exist between the two variables M^2 tends to have greater power than X^2 that ignores this trend.

R Code:

```
>
>  # mid-rank scoring
>  pears.cor(tonsil,c(37,735.5),c(258.5,811,1252))
[1] -0.06993871  6.83331896
>
> 1-pchisq(6.8333,1)
[1] 0.008947381
```

SAS Code:

```
data tonsil;
input carrier $ tonsil $ count;
datalines;
car norm 19
car slt 29
car very 24
nocar norm 497
nocar slt 560
nocar very 269
;

proc freq;
weight count;
tables carrier*tonsil / chisq expected deviation cmh cellchi2
measures;
run;
```

2. (25 pts) Get a dataset from
`http://lib.stat.cmu.edu/DASL/Stories/EducationalAttainmentbyAge.html`,
by clicking on the link after **Datafile Name** and input the dataset into SAS or R or other software and perform the appropriate analysis. What interesting conclusions can you derive about relationship between age and educational attainment?

   From the SAS/R output, for testing independence between education and age group the Pearson Chi-squared statistic is $X^2 = 22373.57$ with df=12 (p-value=0.00) and the G^2 statistic is $G^2 = 23365.06$ with df=12 and p-value=0.00 indicating strong association between age group and education. If we consider both categorical variables $Y = levelofeducation$ and X=age to be ordinal variables then we will use the MH test statistic. From the SAS output we see that the linear trend statistic is $M^2 = 11673.54$ with df=1, which along with either the Pearson correlation $r = -0.299$ (or Spearman correlation -0.3035) is indicating that there is a strong negative linear trend between the two variables. This means that as age increases education level decreases.

   Solution: R Code:

```
> # read in data
> edu=read.table("education.txt",header=TRUE)

> # make it a contingence table
> edu2=xtabs(edu$Count~edu$Education+edu$Age_Group)
```

```
> # Chi-square test
> result=chisq.test(edu2) > result
Pearson's Chi-squared test
data: edu2
X-squared = 22373.57, df = 12, p-value < 2.2e-16

> # G2 test
> LR=2*sum(edu2*log(edu2/result$expected))
> LR
[1] 23365.06
> LRchisq=1-pchisq(LR,df=(5-1)*(4-1))
> LRchisq
[1] 0
```

SAS code and Output:

```
data education;
input education $ group $ count;
datalines;
no_high 25_34 5416
no_high 35_44 5030
no_high 45_54 5777
no_high 55_64 7606
no_high >64 13746
high 25_34 16431
high 35_44 1855
high 45_54 9435
high 55_64 8795
high >64 7558
college1 25_34 8555
college1 35_44 5576
college1 45_54 3124
college1 55_64 2524
college1 >64 2503
College4 25_34 9771
College4 35_44 7596
College4 45_54 3904
College4 55_64 3109
College4 >64 2483
;
run;
proc freq order=data;
weight count;
```

```
tables education*group/chisq expected deviation cmh
cellchi2 measures;
run;
```

Table of education by Age_Group

education Age_Group

Frequency / Expected / Percent / Row Pct / Col Pct	25-34	35-44	45-54	55-64	>64	Total
College1	8555	5576	3124	2524	2503	22282
	6843.9	3416.9	3788.8	3753.7	4478.8	
	6.54	4.26	2.39	1.93	1.91	17.04
	38.39	25.02	14.02	11.33	11.23	
	21.30	27.80	14.05	11.46	9.52	
College4	9771	7596	3904	3109	2483	26863
	8250.9	4119.4	4567.7	4525.4	5399.5	
	7.47	5.81	2.98	2.38	1.90	20.54
	36.37	28.28	14.53	11.57	9.24	
	24.32	37.87	17.55	14.11	9.44	
HS	16431	1855	9435	8795	7558	44074
	13537	6758.7	7494.3	7424.9	8859	
	12.56	1.42	7.21	6.72	5.78	33.70
	37.28	4.21	21.41	19.96	17.15	
	40.90	9.25	42.42	39.92	28.75	
NOHS	5416	5030	5777	7606	13746	37575
	11541	5762.1	6389.2	6330	7552.7	
	4.14	3.85	4.42	5.82	10.51	28.73
	14.41	13.39	15.37	20.24	36.58	
	13.48	25.08	25.98	34.52	52.29	
Total	40173	20057	22240	22034	26290	130794
	30.71	15.33	17.00	16.85	20.10	100.00

Chi-Square	12	22373.5656	<.0001
Likelihood Ratio Chi-Square	12	23365.0598	<.0001
Mantel-Haenszel Chi-Square	1	11440.7382	<.0001
Phi Coefficient		0.4136	
Contingency Coefficient		0.3822	
Cramer's V		0.2388	

Sample Size = 130794

3. (30 pts) In 1972, a sample of 1,524 adults reported both their current religious affiliation and their religious affiliation at age 16.

   (a) Is there any evidence of a change in the rate of Catholic affiliation over time? Find a confidence interval for the rate of change.

	Current affiliation	
Affiliation at age 16	Catholic	Non-Catholic
Catholic	351	67
Non-Catholic	33	1073

Solutions:

Let X and Y categorical variables for Catholic affiliation with categories Catholic=1 and non-Catholic=2. Since both X and Y are measured on the same set of individuals at two different point in time and because we are interested in testing if there is a change of religious affiliation over time we will use the McNemar's test of marginal homogeneity. Thus we are interested in testing:

$H_0 : \pi_{1+} = \pi_{+1}$

The rate of change is the difference in proportions: $\delta = \pi_{1+} - \pi_{+1}$ and its estimate is given by $\hat{\delta} = \frac{n_{12}}{n} - \frac{n_{21}}{n} = 0.02$. The estimated variance is:

$$\begin{aligned}\hat{V}(\hat{\delta}) &= \frac{1}{n}[\frac{n_{12}}{n}(1 - \frac{n_{12}}{n}) + \frac{n_{21}}{n}(1 - \frac{n_{21}}{n}) + 2\frac{n_{12}n_{21}}{n^2}] \\ &= \frac{1}{1524}(\frac{67}{1524}(1 - \frac{67}{1524}) + \frac{33}{1524}(1 - \frac{33}{1524}) + 2\frac{67(33)}{1524^2}) = 0.0000393\end{aligned}$$

Hence a 95% CI for δ is $\hat{\delta} \pm z_{\alpha/2}\sqrt{\hat{V}(\hat{\delta})}$ or $\delta \in (0.008, 0.0032)$

Moreover the Mc Nemars test statistic is

$$z = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}} = 34/10 = 3.4$$

and the corresponding p-value is $2P(Z > 3.4) = 0.0006 < 0.05$. We reject the null hypothesis of marginal homogeiniety and hence we conclude that there is a change of Catholic affiliation over time.

R Code/Output

```
> religion=matrix(c(351,33,67,1073),ncol=2,
+ dimnames=list(Time=c("Catholic","Non-Catholic"),
+ Religion=c('Catholic','Non-Catholic')))
> religion
             Religion
Time          Catholic Non-Catholic
  Catholic          351           67
  Non-Catholic       33         1073
> mcnemar.test(religion,correct=F)

McNemar's Chi-squared test

data:  religion
McNemar's chi-squared = 11.56, df = 1, p-value = 0.0006739

>
> mcnemar.test(religion,correct=T)

McNemar's Chi-squared test with continuity correction

data:  religion
McNemar's chi-squared = 10.89, df = 1, p-value = 0.0009668

>
```

SAS Code/Output

```
data matched;
input first second count;
datalines;
1 1 351
1 2 67
2 1 33
2 2 1073
;
proc freq; weight count;
tables first*second / agree; exact mcnem;
run;
```

McNemar's Test	
Statistic (S)	11.5600
DF	1
Asymptotic Pr > S	0.0007
Exact Pr >= S	8.737E-04

(b) For these data, was it beneficial to record religious affiliation for the same individuals at both points in time? In other words, could we have done just as well if we had recorded current religious affiliation for 1,524 individuals, and religious affiliation at age 16 for a separate independent sample of 1,524 other individuals?

Solutions:

Suppose we recorded religious affiliation for 1524 individuals, and religious affiliation at age 16 for a separate independent sample of 1524. The new table from two independent, cross-sectional samples would have looked like this:

```
>  religion=matrix(c(384,418,1140,1106),ncol=2,
dimnames=list(Time=c("Time 1","Time 2"),
Religion=c('Catholic','Non-Catholic')))
> religion
        Religion
Time     Catholic Non-Catholic
  Time 1      384         1140
  Time 2      418         1106
>
> chisq.test(religion)

Pearson's Chi-squared test with Yates' continuity correction

data:  religion
X-squared = 1.8427, df = 1, p-value = 0.1746

>
```

	Catholic	Non-Catholic
Age 16	418	1106
Current	384	1140

Now the test for change over time is a test of column and row independence. The standard error for the difference in proportion of this new table is

$$\hat{V}(d) = \sqrt{\frac{1}{1524}(\frac{418}{1524}(1-\frac{418}{1524}) + \frac{384}{1524}(1-\frac{384}{1524}))} = 0.0159.$$

which is much greater than the standard error we had previously (0.0000309). Thus for these data it was beneficial to record religious affiliation for the same individuals at two points in time since the confidence interval is more precise (smaller standard error). Using a longitudinal sample apparently produced a much more precise estimate of the rate change than cross-sectional samples would have produced. Moreover, The chi-square test here yields an insignificant result (chi-square-1.84 with p- value=0.17), suggesting that the independent model does fit. This contrasts to McNemar's test result.

SAS Code/Output

```
data matched;
input first second count;
datalines;
1 1 351
1 2 67
2 1 33
2 2 1073
;
proc freq; weight count;
tables first*second / chisq
run;
```

4. (15 pts) Calculate *kappa* for a 4×4 table having $n_{ii} = 5$ for all i, $n_{i,i+1} = 15, i = 1, 2, 3, n_{41} = 15$, and all other $n_{ij} = 0$. Explain why strong association does not imply strong agreement.

Solutions: R code/Output:

```
> ratings=matrix(c(5,0,0,15,15,5,0,0,0,15,5,0,0,0,15,5),ncol=4)
> ratings
     [,1] [,2] [,3] [,4]
[1,]    5   15    0    0
[2,]    0    5   15    0
[3,]    0    0    5   15
[4,]   15    0    0    5
>  library(vcd)
```

```
Loading required package: MASS
Loading required package: grid
Loading required package: colorspace
> kappa=Kappa(ratings)
> kappa
           value        ASE
Unweighted   0.0 0.06454972
Weighted     0.1 0.12677314
>
```

SAS code:

```
data kappa;
input i $ j $ count ;
datalines;
a a 5
 a b 15
 a c 0
 a d 0
 b a 0
 b b 5
 b c 15
 b d 0
 c a 0
 c b 0
 c c 5
 c d 15
 d a 15
 d b 0
 d c 0
 d d 5

 ;
proc freq data=kappa; weight count;
    tables i*j /agree chisq measures;
run;
```

Kappa Statistics

Statistic	Value	ASE	95% Confidence Limits	
Simple Kappa	0.0000	0.0645	-0.1265	0.1265
Weighted Kappa	0.1000	0.0787	-0.0542	0.2542

Sample Size = 80

	B1	B2	B3	B4	Total
A1	5	15	0	0	20
A2	0	5	15	0	20
A3	0	0	5	15	20
A4	15	0	0	5	20
Total	20	20	20	20	80

Statstic	Value	ASE	95% Lower Limits	95% Upper Limits
Simple Kappa	0	0.0645	-0.1265	0.1265
Weighted Kappa	0.1	0.0787	-0.0542	0.2542

From the SAS/R output the unweighted Cohen's kappa statistic is close to 0.0 indicating that there is no aggrement between the two ratings. Indeed, from the table we observe that there is a tendency of opposite ratings. So strong association does not imply strong agreement for example two critics give opposite ratings (strong disagreement). Or if observer A consistently rates subjects one category higher than observer B, strength of agreement is poor even though the association is strong. (Agresti, page 432)