Due Sunday January 31 by 11:30pm Eastern
with no penalty grace period to Tuesday, February 2 by 11:30pm Eastern

Directions. This homework pertains to materials covered in the first three weeks, particularly the third week and the analysis of two-way tables. The assignment should be typed, with your name on the document, and with properly labeled computer output. I suggest you attach your R/SAS input code at the end of your file clearly indicating the problem it corresponds to. If you choose to collaborate, the write-up should be your own. Please show your work! Upload the file to the HW3 Dropbox on ANGEL.

1. (5 pts) The probability simplex for the four possible outcomes of two binary random variables is a tetrahedron. Draw the independence surface (the set of probability distributions corresponding to the independence model for a 2×2 table) inside the tetrahedron. What does it mean for the empirical distribution to be above or below this surface?

   Solution: Some nice pictures appeared in the piazza discussion. The key is to draw all the straight lines that come from fixing one probability and letting the other vary. Two line segments which are edges of the tetrahedron are excluded from the surface: the line segment 00 to 11 of perfect correlation and the line segment 01 to 10 of perfect anticorrelation. Points which are not on the surface but closer to 00-11 are postively associated, points which are closer to 01-10 are negatively associated.

2. (5 pts) True or False?

   (a) In 2×2 tables, statistical independence is equivalent to a population odds ratios value of $\theta = 1$ **Solution:** TRUE

   (b) With rare events, when population proportions are very close to zero, difference of proportions is a better measure of association than the relative risk? **Solution:** FALSE. The difference in proportions will be very small but the effect could be strong.

3. (10 pts) Give an example of inferences in a two-way table for which multinomial, product multinomial, and Poisson sampling assumptions are appropriate.

   Solution: The multinomial sampling treats each count in a two-way table n_{ij} as a multinomial with index n and parameter $\pi = \{\pi_{ij}\}$ where the grand total n is fixed and known. Poisson sampling treats n_{ij} as independent Poisson random variables with parameter μ_{ij}, where the overall n is not fixed. Product-multinomial sampling when each row of the table is multinomial, row totals are fixed and rows are independent (similarly for columns). In other words, if rows represent a characteristic product-multinomial sampling occurs when you draw n_{i+} subjects with characteristic i and record the outcome. An

example will be the study the effect of contraceptive use and heart attack incidence. If the total sample size is a random variable and each count is treated as the four combinations of contraceptive use (yes, no) and heart attack (yes, no) then Poisson sampling is appropriate. If n people are randomly selected and classified according to contraceptive use and outcome then multinomial sampling is appropriate. Finally, if the example is a retrospective study then product-multinomial sampling is appropriate.

4. (30 pts) For adults who sailed on the Titanic on it fateful voyage, the odds ratio between gender (female, male) and survival (yes , no) was 11.4.

   (a) What is wrong with the following interpretation: "The probability of survival for females was 11.4 times that for males"? Give the correct interpretation. **Solution:** The above interpretation corresponds to the relative risk, which is the ratio of probability of survival of females to the probability of survival for males. The correct interpretation is the *odds* of survival for females was 11.4 times the *odds* of survival for males.

   (b) The odds of survival for females equaled 2.9. For each gender, find the proportion who survived. **Solution:** Let p_1 denote the probability of survival for females and p_2 the probability of survival for males. The odds of survival for females is $\frac{p_1}{1-p_1} = 2.9$. Solving for p_1 we get that the probability of survival for females is $p_1 = 0.743$. The odds of survival for males is $\frac{p_2}{1-p_2} = 2.9/11.4 = 0.254$. So the probability of survival for males is $p_2 = (1 + 0.254)/0.254 = 0.202$

   (c) Find the value of R in the interpretation, "The probability of survival for females was R times that for males". **Solution:** $p_1 = Rp_2$. Therefore, $R = p_1/p_2 = 0.743/0.202 = 3.68$.

5. (20 pts) A handwriting expert claims to have ability to discern whether a note was written by a man or a woman. An experiment was performed in which five men and five women submitted handwriting samples. The samples were then presented to the judge (who was told their were five men and five women), and he rendered his opinion on which samples were male and which were female.

	Experts classification	
True gender	Male	Female
Male	4	1
Female	1	4

Perform both approximate and exact tests. What is your conclusion regarding the expert's claim?

Solution: Since the experiment fixed both marginal totals we will use Fisher's exact test to test H_0: the expert has no discerning ability ($\theta = 1$) against the one-sided alternative hypothesis H_1: $\theta > 1$ and the two-sided H_1: $\theta = 1$. From SAS/R output we can conclude that for both alternatives, we do not have enough evidence against the null with the Fisher's exact test.

R Code:

```
> writing=matrix(c(4,1,1,4),ncol=2,dimnames
=list(Truth=c("M","F"), Expert=c("M","F")))
> writing
     Expert
Truth M F
    M 4 1
    F 1 4
>
> fisher_greater=fisher.test(writing,alternative="greater")
> fisher_greater

        Fisher's Exact Test for Count Data

data:  writing
p-value = 0.1032
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 0.6498735       Inf
sample estimates:
odds ratio
   10.9072

>
> result=chisq.test(writing)
Warning message:
In chisq.test(writing) : Chi-squared approximation may be incorrect
> result

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

data:  writing
X-squared = 1.6, df = 1, p-value = 0.2059

>
> LR=2*sum(writing*log(writing/result$expected))
> LR
```

```
[1] 3.854895
>
> LRchisq=1-pchisq(LR,df=1)
> LRchisq
[1] 0.04960103
>
>
```

SAS Code:

```
data writing;
input truth $ expert $ count; datalines;
male male 4
male female 1
female male 1
female female 4
;
proc freq data=writing order=data;
weight count;
tables truth*expert/ chisq relrisk riskdiff expected; exact fisher chisq or;
run;
```

Pearson Chi-Square Test

Chi-Square	3.6000
DF	1
Asymptotic Pr > ChiSq	0.0578
Exact Pr >= ChiSq	0.2063

The SAS System

The FREQ Procedure

Statistics for Table of truth by expert

The FREQ Procedure

Statistics for Table of truth by expert

Likelihood Ratio Chi-Square Test

Chi-Square	3.8549
DF	1
Asymptotic Pr > ChiSq	0.0496
Exact Pr >= ChiSq	0.2063

Fisher's Exact Test

Cell (1,1) Frequency (F)	4
Left-sided Pr <= F	0.9960

```
Right-sided Pr >= F          0.1032

Table Probability (P)        0.0992
Two-sided Pr <= P            0.2063
```

6. (30 pts) *Sensitivity* and *specificity* are measures often calculated for 2×2 tables. These measures are of particular interest when you want to determine the efficacy of a screening test (e.g. for a disease, lie detection, etc.). To learn a bit about these measures, you should read the "AccuracyHandout.pdf" (see online Lesson 3.1.7: Difference in proportions `https://onlinecourses.science.psu.edu/stat504/node/77`). In Agresti(2013) you can find some information in Sec. 2.1.2. Of course, you can always search the web, or use other textbooks.

   (a) (10 pts) a) How do sensitivity and specificity relate to conditional probabilities, relative risk and odds ratios?

   Solution:

Test Result	**True Condition**		
	Positive Z=1	Negative Z=2	Total
Positive Y=1	a	b	a+b
Negative Y=2	c	d	c+d
Total	a+c	b+d	N=a+b+c+d

$$\text{Sensitivity} = P(\text{fail to reject null}|\text{null is true}) = \frac{a}{a+c}$$

$$\text{Specificity} = P(\text{reject null}|\text{alternative is true}) = \frac{d}{b+d}$$

$$\text{Relative risk} = \rho = \frac{P(Z=1|Y=1)}{P(Z=1|Y=2)} = \frac{a}{c}$$

$$\text{odds ratio} = \theta = \frac{P(Z=1|Y=1)P(Z=2|Y=1)}{P(Z=1|Y=2)P(Z=2|Y=2)} = \frac{a/b}{c/d}$$

Thus, the sensitivity is equal to $\frac{\rho}{1-\rho}$ and the odds ratio can be written as θ =(sensitivity)(1-sensitivity)/ (specificity)(1-specificity).

(b) (20 pts) PET scans can be used in diagnosing a certain type of cancer, such as lung cancer. Consider following hypothetical data on 150 adults in a mining town:

Table 1: default

	Test result	
Cancer	positive	negative
yes	65	5
no	3	77

Are the test results and lung cancer status independent? Report the results of chi-squared test of independence, and interpret.

Describe the association of the test results and cancer using any of the measures of associations you deem appropriate (e.g., difference in proportion, the relative risk and the odds ratio) and interpret your findings. Also report the sensitivity, specificity, false positive and false negative rates?

Solution:

From the result of the Pearson's chi-square test and likelihood ratio chi-square test, we can reject null hypothesis that there is no relationship between cancer and test result, that is, the test result is associated with cancer.

Difference in proportion is 0.89, indicating that the risk of getting positive test result given the person truly has cancer is 0.89 larger than that of the patient truly does not have cancer.

Relative risk is 24.76, indicating that the risk of getting positive test result given the person truly has cancer is 24.76 times that of the patient truly does not have cancer.

Odds ratio is 333.67, meaning that the odds of having positive test result given the patient truly has cancer is 333.67 times that of the patient truly does not have cancer.

Further,

$$\text{Sensitivity} = 65/70 = 0.93$$
$$\text{Specificity} = 77/80 = 0.94$$
$$\text{False positive rate} = 3/80 = 0.04$$
$$\text{False negative rate} = 5/70 = 0.07$$

R Code:

```
> cancer=matrix(c(65,3,5,77),ncol=2,dimnames
=list(Cancer=c("Y","N"),Result=c("P","N")))
> cancer
      Result
Cancer  P  N
     Y 65  5
     N  3 77
>
> ### Pearson's Chi-squared test with Yates' continuity correction
> result=chisq.test(cancer)
> result

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

data:  cancer
X-squared = 116.0453, df = 1, p-value < 2.2e-16

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

        Pearson's Chi-squared test

data:  cancer
X-squared = 119.6139, df = 1, p-value < 2.2e-16

>
```

```
> ### Likelihood Ratio Chi-Squared Statistic
> G2=2*sum(cancer*log(cancer/result$expected))
> G2
[1] 145.0244
> pvalue=1-pchisq(2*sum(cancer*log(cancer/result$expected)),df=1)
> pvalue
[1] 0

>
> RowSums=rowSums(cancer)
> ColSums=colSums(cancer)
>
> ### Column 1 Risk Estmates
> risk1_col1=cancer[1,1]/RowSums[1]
> risk2_col1=cancer[2,1]/RowSums[2]
> rho1=risk1_col1/risk2_col1
> total1=ColSums[1]/sum(RowSums)
> diff1=risk2_col1-risk1_col1
> rbind(risk1_col1,risk2_col1,total1,diff1)
                    Y
risk1_col1  0.9285714
risk2_col1  0.0375000
total1      0.4533333
diff1      -0.8910714
> ### The confidence interval for the difference in proportions for column 1
> SE_diff1=sqrt(risk1_col1*(1-risk1_col1)/RowSums[1]
+risk2_col1*(1-risk2_col1)/RowSums[2])
> CI_diff1=cbind(diff1-qnorm(0.975)*SE_diff1,diff1+qnorm(0.975)*SE_diff1)
> SE_diff1
         Y
0.03739911
> CI_diff1
        [,1]       [,2]
N -0.9643723 -0.8177705


> ### Column 2 Risk Estmates
> risk1_col2=cancer[1,2]/RowSums[1]
> risk2_col2=cancer[2,2]/RowSums[2]
> total2=ColSums[2]/sum(RowSums)
> diff2=risk2_col2-risk1_col2
> rbind(risk1_col2,risk2_col2,total2,diff2)
                    Y
risk1_col2 0.07142857
risk2_col2 0.96250000
total2     0.54666667
diff2      0.89107143
> ### The confidence interval for the difference in proportions for column 2
> SE_diff2=sqrt(risk1_col2*(1-risk1_col2)/RowSums[1]
+risk2_col2*(1-risk2_col2)/RowSums[2])
> CI_diff2=cbind(diff2-qnorm(0.975)*SE_diff2,diff2+qnorm(0.975)*SE_diff2)
```

```
> SE_diff2
         Y
0.03739911
> CI_diff2
       [,1]      [,2]
N 0.8177705 0.9643723



> ### Estimate of the Odds of the two rows
> odds1=(cancer[1,1]/RowSums[1])/(cancer[1,2]/RowSums[1])
>
>
> odds2=(cancer[2,1]/RowSums[2])/(cancer[2,2]/RowSums[2])
> ### Odds Ratio
> oddsratio=odds1/odds2
> oddsratio
       Y
333.6667


> ### Confidence Interval of the odds ratio
> log_CI=cbind(log(oddsratio)-qnorm(0.975)*sqrt(sum(1/cancer)),
log(oddsratio)+qnorm(0.975)*sqrt(sum(1/cancer)))
> CI_oddsratio=exp(log_CI)
> CI_oddsratio
      [,1]     [,2]
Y 76.80031 1449.648
```

SAS Code:

```
data cancer;
input cancer $ result $ count ;
datalines;
Y P 65
N P 3
Y N 5
N P 77
;
proc FREQ order=data;
weight count;
tables cancer*result/chisq relrisk riskdiff expected;
/*tables race*party/all expected relrisk riskdiff */
exact or;
run;
```

```
                                The SAS System

                              The FREQ Procedure

                          Table of cancer by result

                    cancer     result
```

Frequency Expected Percent Row Pct Col Pct	P	N	Total
Y	65 31.733 43.33 92.86 95.59	5 38.267 3.33 7.14 6.10	70 46.67
N	3 36.267 2.00 3.75 4.41	77 43.733 51.33 96.25 93.90	80 53.33
Total	68 45.33	82 54.67	150 100.00

The FREQ Procedure

Statistics for Table of cancer by result

Statistic	DF	Value	Prob
Chi-Square	1	119.6139	<.0001
Likelihood Ratio Chi-Square	1	145.0244	<.0001
Continuity Adj. Chi-Square	1	116.0453	<.0001
Mantel-Haenszel Chi-Square	1	118.8164	<.0001
Phi Coefficient		0.8930	
Contingency Coefficient		0.6661	
Cramer's V		0.8930	

Fisher's Exact Test

Cell (1,1) Frequency (F)	65
Left-sided Pr <= F	1.0000
Right-sided Pr >= F	2.058E-32
Table Probability (P)	4.279E-30
Two-sided Pr <= P	2.058E-32

Column 1 Risk Estimates

	Risk	ASE	(Asymptotic) 95% Confidence Limits		(Exact) 95% Confidence Limits	
Row 1	0.9286	0.0308	0.8682	0.9889	0.8411	0.9764
Row 2	0.0375	0.0212	0.0000	0.0791	0.0078	0.1057
Total	0.4533	0.0406	0.3737	0.5330	0.3720	0.5366
Difference	0.8911	0.0374	0.8178	0.9644		

Difference is (Row 1 - Row 2)

Column 2 Risk Estimates

	Risk	ASE	(Asymptotic) 95% Confidence Limits		(Exact) 95% Confidence Limits	
Row 1	0.0714	0.0308	0.0111	0.1318	0.0236	0.1589
Row 2	0.9625	0.0212	0.9209	1.0000	0.8943	0.9922
Total	0.5467	0.0406	0.4670	0.6263	0.4634	0.6280
Difference	-0.8911	0.0374	-0.9644	-0.8178		

Difference is (Row 1 - Row 2)

The FREQ Procedure

Statistics for Table of cancer by result

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	333.6667	76.8003	1449.6483
Cohort (Col1 Risk)	24.7619	8.1437	75.2918
Cohort (Col2 Risk)	0.0742	0.0319	0.1729

Odds Ratio (Case-Control Study)

Odds Ratio	333.6667
Asymptotic Conf Limits	
95% Lower Conf Limit	76.8003
95% Upper Conf Limit	1449.6483
Exact Conf Limits	
95% Lower Conf Limit	67.2709
95% Upper Conf Limit	2009.4138

Sample Size = 150