Homework Assignment 2

(Due Monday February 11 at 1PM)

Total Points: 122

Please hand in a print-out of your answer and R code, and also email your R code to me (hoyen@stat.sc.edu). Please use the R markdown Homework template (Stat704_HWtemplate.Rmd) to write your homework solutions.

I. Contingency Tables

1. Download the class simulation data set "task1.csv" from the course web site. Here's the commands that I used to read it in

```
dat <- read.csv("task1.csv", header = FALSE)
dat2 <- dat[,1 : 10]
dat2 <- dat2[complete.cases(dat2),]
vec1 <- as.vector(unlist(dat2))
```

Dat is the original data. Dat2 contains only the data, removing any subjects containing errors. Vec1 is the data disregarding subject level information.

(a) Do the numbers 1-10 appear to be equally likely? Perform the appropriate Chi-squared test. (5 points)

(b) Approximate an exact Chi-squared test by doing the following. Simulate 1,000 random multinomials under the null hypothesis with the command

```
simdat <- t(rmultinom(1000, size = length(vec1), p = rep(.1, 10)))
```

Obtain the chi-squared statistics for each with the command

```
chsqStats <- apply(simdat, 1, function(x) chisq.test(x)$statistic)
```

Calculate the percentage of time that these statistics are greater than the observed statistic. Explain how, provided the Monte Carlo sample is large, this is a p-value. (5 points)

2. The following data show the results of caries surveys in five towns and also the fluoride content of the drinking water.

Area	Essex	Slough	Harwick	Burnham	West Meresa	Total
Fluoride p.p.m	0.15	0.9	2.0	3.5	5.8	
Number children with caries	243	83	60	31	39	456
Number children with caries free teeth	16	36	32	31	12	127
Number examined	259	119	92	62	51	583

The data refer to samples of children aged 12-14 only.

(a) Do a significant test to determine whether the proportions of children caries free various from area to area. What does this test reveal about the effect of the fluoride content of water? (5 points)

(b) Briefly discuss the limitation of this types of study for studying the effect of fluoride. (5 points)

3. Prove that the odds ratio is greater than 1 if and only if the relative risk is greater than 1. (7 points)

4. In a study of the association between cigarette smoking and lung cancer, 1,357 male lung cancer patients were compared with 1,357 controls in terms of their cigarette consumption as follows:

	Cigarette Consumption Daily						
	0	1-	5-	15-	25-	50+	Total
Lung Cancer Patients	7	49	516	445	299	41	1,357
Control	61	91	615	408	162	20	1,357

Compute the odds ratio and log odds ratio in each of the 5 smoking groups compared with non-smokers. Find confidence intervals and graphically

display. Comment and interpret. Can relative risks be estimated? Why or why not? (5 points)

5. Suppose we wish to compare two treatments for breast cancer via, simple mastectomy (S) and radical mastectomy (R). We form matched pairs of women who are within the same decade of age and with the same clinical condition to receive the two treatments and measure their 5-year survival. The results are given (L=live at least 5 year, D=died within 5 years) below. Perform an analysis of this data and interpret your results. (10 points)

Pair	Treatment S person	Treatment R person	Pair	Treatment S person	Treatment R person
1	L	L	11	D	D
2	L	D	12	L	D
3	L	L	13	L	L
4	L	L	14	L	L
5	L	L	15	L	D
6	D	L	16	L	L
7	L	L	17	L	D
8	L	D	18	L	D
9	L	D	19	L	L
10	L	L	20	L	D

II. Logistic Regression

Instructions: feel free to discuss the homework with other students. However, each student must conduct their own analyses and write-up their own solutions. Write as if for a scientific journal. Be brief and accurate.

Case study: The Health Care Cost for Smoking

1. **Analysis Goals**

The analysis goal of this project is to estimate the fraction of total medical expenditures among smokers with coronary hear disease, stroke, and lung cancer that can be attributed to their having smoked.

The medical model that underlines this question can be expressed as:

Smoking -> major diseases -> medical expenditure.

The analyses consist of two parts:

(1) Using MEPS, to estimate the difference in average expenditures between the diseased and non-diseased subgroups.

(2) Using MEPS, to estimate the fraction of total medical expenditures among smokers with CHD and other smoking-related diseases that can be attributed to their having smoked.

Reference: Johnson E., Dominici F., Griswold M., Zeger SL. (2003) "Disease cases and their medical cost attributable to smoking: an analysis of the national medical expenditure survey." Journal of Econometrics, 112: 135-151.

(2) Dataset

Medical Expenditure Panel Survey (MEPS) is a set of large-scale surveys of families and individuals, their medical providers (doctors, hospitals, pharmacies, etc.) and employers across the United States. MEPS collects data on the specific health services that Americans use, how frequently they use them, the cost of these services, and how they are paid for, as well as data on the cost, scope, and breadth of health insurance held by and available to U.S. workers. We will use the MEPS data collected during 2009 for the following analysis.

The MEPS data is available at

http://people.stat.sc.edu/hoyen/Stat704/Data/h129.RData

(3) Logistic Regression

Now use the binary disease indicator for disease status (CHD or stroke or lung cancer) and building regression models to examine the effect of being a current smoker.

1. This is a continuation of last semester's final project. Use the following regression model and complete tables below for male and female participants respectively. (10 points)

$$E(lexp) = \beta_0 + \beta_1 Age + \beta_2 Gender + \beta_3 MSCD + \beta_6 MSCD \times Age$$

where $lep = log_{10}(exp + 100)$, MSCD=1 if major smoking-caused-disease, 0 otherwise.

Age	Estimated **difference** of Median expend(**$**): Diseased -Non-diseased	Std Error (Delta Method)	Std Error (bootstrap)	95% CI
20				
40				
65				
80				

2. Use model building techniques, building a logistic regression model. (5 points)
3. Complete the table below using the model obtained from (12). (15 points)

Age	Odd ratio of Current Smoker (Female)	95% CI	Odd ratio of Current Smoker (Male)	95% CI
20				
40				
65				
80				

4. Estimate the attributable risk of being a current smoker. Attributable risk can be calculated as follows:

$$\frac{\Pr(CHD = 1|Current\ Smoke = 1) - \Pr\ (CHD = 0|Current\ Smoke = 1)}{\Pr\ (CHD = 1|Current\ Smoke = 1)}$$

And complete the table below. (15 points)

Table: Attributable risk of current smoking to CHD

Age	Female (%)	95% CI	Male (%)	95% CI
20				
40				
65				
80				

5. Estimate medical expenditure attributable due to current smoke. Combine the numbers obtain from Tables in Question 1, 3, 4. Calculate how much smoking could cost. Recall that the medical model underline this question can be expressed as

   Smoking -> major diseases -> medical expenditure

   Calculate smoking attributable medical expenditure as follows and complete the table below.

   Attributable medical expenditures=Attributable risk ×expected difference

Table: Attributable medical expenditures (USD) (20 points)

Age	Female ($)	95% CI	Male ($)	95% CI
20				
40				
65				
80				

6. Summarize your regression findings in a brief paragraph as if for a public health journal. Use coefficient estimates and confidence intervals in the text. (15 points)