BST 140.652
Problem Set

Problem 1. The table below gives the children's genders in a random sample of 1,000 two children families.

	First child				
	Male		Female		
Second child	Male	Female	Male	Female	Total
Count	218	227	278	277	1,000

a. It is typically thought that the gender of offspring within a family are independent and identically distributed with males and females being equally likely. Is this hypothesis supported by the data above?

b. Specifically test independence of the gender of the first child to the second.

Problem 2. Consider the hypothesis testing problem of comparing two binomial probabilities H_0 : $p_1 = p_2$. Show that the square of statistic $(\hat{p}_1 - \hat{p}_2)/\mathrm{SE}_{\hat{p}_1 - \hat{p}_2}$ is the same as the χ^2 statistic. Here, the standard error in the denominator is calculated under the null hypothesis. (Clearly define any notation you introduce.)

Problem 3. A study of the effectiveness of *streptokinase* in the treatment of patients who have been hospitalized after myocardial infarction involves a treated and control group. In the streptokinase group, 2 of 15 patients died within 12 months. In the control group, 4 of 19 died with 12 months.

a. Use Fisher's exact test to test for a difference in mortality rates. Do this by hand by writing down all possible tables with fixed marginal totals. You may confirm your results with a computer.

b. Compare your results using the test statistics based on the normal and χ^2 approximations.

Problem 4. 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 orginal 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.

b. Approximate an exact Chi-squared test by doing the following. Simulate 1,000 random nultinomials 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.

Problem 5. A case-control study of esophageal cancer was performed. Daily alcohol consumption was ascertained ($80+$ gm = high, $0-79$ gm = low). The data was stratified by 3 age groups.

	Alcohol	
	H	L
case	8	5
control	52	164

Age 35-44

	Alcohol	
	H	L
case	25	21
control	29	138

Age 45-54

	Alcohol	
	H	L
case	50	61
control	27	208

Age 55-64

Assuming a constant odds ratio across age-strata, test to see if the odds ratio is 1. If not, estimate it.

Problem 6. Retinitis pigmentosa is a disease which manifests itself via different genetic modes of inheritance. Cases have been documented with a dominant, recessive, and sex-linked form of inheritance. It has been conjectured that the form of inheritance is related to the ethnic origin of the individual. Cases of the disease have been surveyed in an English and Swiss population with the following results: out of 125 English cases, 46 had sex-linked disease, 25 had recessive disease and 54 had dominant disease; out of the 110 Swiss cases, one had sex-linked disease, 99 had recessive disease, and 10 had dominant disease. Based on these data is there a significant association between ethnic origin and genetic type? Analyze and interpret (in words) this data. (10 points)

Problem 7. 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
Controls	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.

Problem 8. In a retrospective study of the possible effect of blood group on the incidence of peptic ulcers, Woolf (1955) obtained data from three cities. The table gives for each city data for blood groups 0 and A only. In each city, blood group is recorded for peptic ulcer subjects and for a control series of individuals not having peptic ulcer.

	Peptic Ulcer		Control	
	Group 0	Group A	Group 0	Group A
London	911	579	4578	4219
Manchester	361	246	4532	3775
Newcastle	396	219	6598	5261

a. Compute the odds ratio for each city with a confidence interval. Interpret.

b. Suppose that it is required to estimate P(ulcer|A) − P(ulcer|0). What further information is needed to do this from the current data?

Problem 9. Suppose we wish to compare two treatments for breast cancer, viz., 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=lived at least 5 years, D=died within 5 years) below. Perform an analysis of this data, and interpret your results.

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

Problem 10. Suppose we are interested in comparing the effectiveness of 2 different antibiotics A and B in treating gonorrhea. We match each person receiving antibiotic A with an equivalent person (age within 5 years, same sex) to whom we give antibiotic B and we ask that these persons return to the clinic within 1 week to see if the gonorrhea has been eliminated.
Suppose the results are as follows:

For 40 pairs of people both antibiotics are successful.

For 20 pairs of people antibiotic A is effective while antibiotic B is not.

For 16 pairs of people antibiotic B is effective while antibiotic A is not.

For 3 pairs of people neither antibiotic is effective.

Perform an analysis to compare the relative effectiveness of the two antibiotics. Interpret your results.

Problem 11. Consider a retrospective study with matched pairs. Show that McNemar's test statistic is equivalent to performing a Mantel-Haenszel test for all 2×2 tables (with one table for each pair).

Problem 12. A researcher is studying migration patterns. She collected the location of the current and previous homes for subjects who moved across regions. She recorded the following:

	Previous home		
Current home	Northeast	Southeast	West
Northeast	-	267	255
Southeast	135	-	139
West	240	234	-

Here the diagonals are not included since she only studied subjects who moved between regions. She would like to know if the probability of moving from region a to b is the same as the probability of moving from region b to a for all regions a and b.

a. Mathematically state her null and alternative hypotheses defining any notation you use.

b. Calculate the expected counts under the null hypothesis.

c. Perform the Chi-squared test and state your conclusions in the language of the problem. (Hint the df is 3.)