BST 140.652
Problem Set 6

Problem 1. 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 2. 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, 20 of 150 patients died within 12 months. In the control group, 40 of 190 died with 12 months.

a. Test equivalence of the two proportions.

b. Give confidence intervals for the absolute change in proportions, the relative risk and odds ratio.

c. Create Bayesian credible intervals for the risk difference, risk ratio and odds ratio. Plot the posterior for each and interpret the results.

Problem 3. Researchers are interested in estimating the natural log of the proportion of people in the population with hypertension. In a random sample of n subjects, let X be the number with hypertension. Derive a confidence interval for the natural log of the proportion of people with hypertension. Assume that n is large.

Problem 4. This problem considers the delta method.

a. Derive the asymptotic standard error for $\sqrt{\hat{p}}$ where $\hat{p}$ is a binomial sample proportion.

b. Assume that $n = 200$ and $p = .5$. Implement a simulation study to verify that the delta method results in approximately normally distributed variables.

Problem 5. In this homework, we will evaluate the performance of the log odds ratio interval estimate

$$\log \hat{OR} \pm 1.96\sqrt{1/n_{11} + 1/n_{12} + 1/n_{21} + 1/n_{22}}.$$

Use R to generate $1,000$ random binomials with `n1` trials and `p1` success probability; call this vector `x`. Use R to generate $1,000$ random binomials with `n2` trials and `p2` success probability; call this vector `y`. Squash these to vectors together with the command `z = cbind(x, y)`. Now, create $1,000$ sample odds ratios with the command

```
OR = apply(z, 1,
           function(x) x[1] * (n2 - x[2]) / (x[2] * (n1 - x[1]))
         )
```

Log these odds ratios to obtain $1,000$ sample log odds ratios. Now obtain $1,000$ standard errors with the command

```
SELOGOR = apply(z, 1,
          function(x) sqrt(1 / x[1] + 1 / (n1 - x[1]) +
                           1 / x[2] + 1 / (n2 - x[2])
                          )
         )
```

Now, see how often the interval for the log odds ratio contains the true log odds ratio. Repeat this process for all of the following combinations

```
p1 = .1; p2 = .1; n1 = 100; n2 = 100
p1 = .1; p2 = .5; n1 = 100; n2 = 100
p1 = .1; p2 = .9; n1 = 100; n2 = 100
p1 = .5; p2 = .5; n1 = 100; n2 = 100
p1 = .5; p2 = .9; n1 = 100; n2 = 100
p1 = .9; p2 = .9; n1 = 100; n2 = 100
```

Summarize your findings.

Problem 6. In this homework, we will also evaluate the performance of the log relative risk interval estimate

$$\log \hat{RR} \pm 1.96\sqrt{(1-\hat{p}_1)/(\hat{p}_1 n_1) + (1-\hat{p}_2)/(\hat{p}_2 n_2)}$$

Use R to generate $1,000$ random binomials with `n1` trials and `p1` success probability; call this vector `x`. Use R to generate $1,000$ random binomials with `n2` trials and `p2` success probability; call this vector `y`. Squash these to vectors together with the command `z = cbind(x, y)`. Now, create $1,000$ sample risk ratios with the command

```
RR = apply(z, 1,
          function(x) (x[1] / n1) / (x[2] / n2)
         )
```

Log these risk ratios to obtain $1,000$ sample log risk ratios. Now obtain $1,000$ standard errors with the command

```
SELOGRR = apply(z, 1,
            function(x) {
               phat1 <- x[1] / n1
               phat2 <- x[2] / n2
               sqrt((1 - phat1) / phat1 / n1 + (1 - phat2) / phat2 / n2)
            }
         )
```

Now, see how often the interval for the log relative risk contains the true log relative risk. Repeat this process for the following combinations

```
p1 = .1; p2 = .1; n1 = 100; n2 = 100
p1 = .1; p2 = .5; n1 = 100; n2 = 100
p1 = .1; p2 = .9; n1 = 100; n2 = 100
p1 = .5; p2 = .5; n1 = 100; n2 = 100
p1 = .5; p2 = .9; n1 = 100; n2 = 100
p1 = .9; p2 = .9; n1 = 100; n2 = 100
```

Summarize your findings.

Problem 7. The following data show the results of caries surveys in five towns and also the fluoride content of the drinking water.

Area Fluoride p.p.m.	Surrey and Essex 0.15	Slough 0.9	Harwick 2.0	Burnham 3.5	West Meresa 5.8	Total
Number children with 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. Estimate the odds ratio for the propobability of caries for lowest and highest (.15 to 5.8) categories of flouride. Interpret these results.
b. Estimate the relative risk for the propobability of caries for lowest and highest (.15 to 5.8) categories of flouride. Interpret these results.
c. Estimate the risk difference for the propobability of caries for lowest and highest (.15 to 5.8) categories of flouride. Interpret these results.
d. Test the equivalence for the probability of caries between the lowest and highest flouride concentrations (.15 to 5.8) give a P-value and interpret your results.

Problem 8. 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.

Age 35-44

	Alcohol H	Alcohol L
case	8	5
control	52	164

Age 45-54

	Alcohol H	Alcohol L
case	25	21
control	29	138

Age 55-64

	Alcohol H	Alcohol L
case	50	61
control	27	208

Give a confidence interval estimate of the odds ratio in the Oldest age group.

Problem 9. In a study of aquaporins, 120 frog eggs were randomized, 60 to receive a protein treatment and 60 controls. If the treatment of the protein is effective, the frog eggs would implode. The resulting data was

	Imploded	Did not	Total
Treated	50	10	60
Control	20	40	60
Totals	70	50	120

State the appropriate hypotheses and report and interpret a P-value.

Problem 10. Let $\hat{p}$ be the sample proportion from a binomial experiment with n trials. Recall that the standard error of $\hat{p}$ is $\sqrt{\hat{p}(1-\hat{p})/n}$. Define $f(\hat{p}) = \log\{\hat{p}/(1-\hat{p})\}$ as the sample log odds. Note, the following fact might be useful:

$$f'(x) = \frac{1}{x(1-x)}$$

Use the delta method to create a confidence interval for the sample log odds.

Problem 11. Refer to the previous problem. Let $\hat{p}_1$ be the sample proportion from one binomial experiment with n_1 trials and $\hat{p}_2$ be the sample proportion from a second with n_2 trials. Define the log odds ratio to be $f(\hat{p}_1) - f(\hat{p}_2)$. Use your answer to part 2 to derive a confidence interval for the log odds ratio.

Problem 12. Two drugs, A and B , are being investigated in a randomized trial with the data are given below. Investigators would like to know if the Drug A has a greater probability of side effects than drug B.

	None	Side effects	N
Drug A	10	30	40
Drug B	30	10	40

a. State relevant null and alternative hypotheses and perform the relevant test.

b. Estimate a confidence interval for the odds ratio, relative risk and risk difference. Interpret these reslts.

Problem 13. You are flipping a coin and would like to test if it is fair. You flip it 10 times and get 8 heads. Specify relevant hypotheses and report and interpret an exact P-value.

```
pbinom(7, 10, 0.5, lower.tail = FALSE) + pbinom(2, 10, 0.5)
```