Homework # 2

Due Wednesday 2/14

Show all calculations. Turn in R code if applicable.

1. *Estimating binomial varaince*: Suppose $Y \sim bin(n,p)$. Use the delta method to find the standard error of the estimated variance of $\hat{p}$.

2. *HW #1, Problem 8 continued*: The following table shows final grades for students enrolled in a mathematical statistics class, according to whether the student spent 10+ hours a week studying or not. Assume "Final Grade" as the response variable.

	Final Grade	
Study Time	A	B
$\geq$ 10 hours/week	70	10
$<$ 10 hours/week	2	12

   (a) Find and interpret 95% Wald confidence interval for the difference of proportions.
   (b) Find and interpret 95% Wald confidence interval for the relative risk.
   (c) Find and interpret 95% Wald confidence interval for the odds ratio.
   (d) Is there an association between final grade and study amount? Justify your answer.
   (e) *Simulated coverage probability for Wald CI of the odds ratio.*
      i. Consider the estimated joint probabilities $\{\hat{p}_{ij}\}$ of this contingency table to be the "true" distribution. Simulate 100 2x2 contingency tables with this "true" distribution and $n = 500$. *Hint: Compare the four cumulative probabilities* $(p_{11}, p_{11} + p_{12}, p_{11} + p_{12} + p_{21}, p_{11} + p_{12} + p_{21} + p_{22})$ *to a random draw from the uniform(0,1) distribution. Repeat this* 500 *times and count how many times the uniform random variable falls in each range to determine the count in each cell.*
      ii. Compute the 95% Wald confidence interval for the odds ratio for each of the 100 simulated tables.
      iii. Find the proportion of times that the 95% Wald confidence interval contains the true odds ratio. This is the estimated coverage probability. How does it compare to the nominal coverage probability?
      iv. Now repeat the simulation with $n = 100$. What went wrong?

3. *Testing for independence and identical distributions.*

   (a) In professional basketball games during the 2009-2010 NBA regular season, when LeBron James shot a pair of free throws, 8 times he missed both, 152 times he made both, 33 times he made only the first, and 37 times he made only the second. Is it plausible that the successive free throws are independent?
   (b) For a 2x2 table consider $H_0 : \pi_{11} = \theta^2, \pi_{12} = \pi_{21} = \theta(1-\theta), \pi_{22} = (1-\theta)^2$. Show that the marginal distributions are identical and that independence holds.
   (c) For a multinomial sample, under H_0 from part (b), show that $\hat{\theta} = (p_{1+} + p_{+1})/2$.
   (d) Explain how to test H_0. Show that $df = 2$ for the test statistic.

(e) Are LeBron James' pairs of free throws plausibly independent *and* identically distributed?

4. Using the following data, consider studying the association of taking (and passing!) Stat 6231 and job/study status following graduation from the MS program:

Did the student take Stat 6231?	Status after graduation		
	PhD	Employed	Unemployed
Yes	192	75	8
No	459	586	471

(a) Use X^2 and G^2 to test the hypothesis of independence between status after graduation and Stat 6231. Make sure to report the p-values and interpret.

(b) Use Pearson standardized residuals to describe the evidence of association.

(c) Partition the χ^2 into components between going on to PhD studies and being employed, and between these two combined and being unemployed. Interpret.

5. Consider the University of California Berkeley admissions data from Lecture 4 (see Blackboard for the R code to obtain this dataset).

(a) Write out all possible (hierarchical) loglinear models for the three categorical variables and describe the association/independence structure assumed in each one. *Hint: There are 9.*

(b) Find the likelihood ratio test statistic G^2, associated degrees of freedom and p-value for each of the possible loglinear models. Present the results in a table.

(c) Which model would you choose based on the overall fit statistic?

(d) Consider the loglinear model that only excludes the 3-way interaction term - sometimes called the *homogeneous model*. Find the likelihood ratio test statistic G^2, associated degrees of freedom and p-value for testing each of the two-way interaction terms. Present the results in a table.

6. Consider the loglinear model

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}.$$

(a) Describe in words the type of independence assumed in this model.

(b) Write the loglikelihood for this model.

(c) Find the set of likelihood equations for this model.

(d) Show that the likelihood equations derived from derivatives with respect to the interaction terms determine the MLEs. That is, show that they imply the other equations.

(e) For fixed k, show that $\{\hat{\mu}_{ijk}\}$ equal the fitted values for testing independence between X and Y within level k of Z.

(f) Show that the LRT statistic G^2 for testing this model's fit has the form $G^2 = \sum_k G_k^2$ where G_k^2 test independence between X and Y at level k of Z.

7. *Understanding Constraints.*

(a) Show that the following are equivalent ways to write the loglinear independence model for a 2-way table:

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y \text{ subject to } \sum_{i=1}^{I} \lambda_i^X = \sum_{j=1}^{J} \lambda_j^Y = 0 \tag{1}$$

$$\log \mu_{ij} = \gamma + \gamma_i^X + \gamma_j^Y \text{ where } \gamma = \lambda + \frac{\sum_{h=1}^{I} \lambda_h^X}{I} + \frac{\sum_{h=1}^{J} \lambda_h^Y}{J}, \tag{2}$$

$$\gamma_i^X = \lambda_i^X - \frac{\sum_{h=1}^{I} \lambda_h^X}{I} \text{ and } \gamma_j^Y = \lambda_j^Y - \frac{\sum_{h=1}^{J} \lambda_h^Y}{J}$$

(b) How would you similarly reparameterize the following loglinear model to incorporate the constraint?

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y \text{ subject to } \lambda_1^X = \lambda_1^Y = 0.$$

Group Request Form

Please list the students that would like in your group for the group lecture project.

All students should turn in this form.

Groups should be 2 or 3 students.

Your name:

Other group members names:

1.
2.