Homework # 1

Due Wednesday 1/31

1. (15 points) *Poisson Distribution*

   (a) Derive the mean and variance of a Poisson random variable to show the equidispersion property of the Poisson distribution.

   (b) Suppose that $y_1, ..., y_n$ are independent from a Poisson distribution with parameter μ. Find the likelihood function and the MLE for μ.

   (c) Construct a large-sample test statistic for $H_0 : \mu = \mu_0$ using:

   i. the Wald method,
   ii. the score method, and
   iii. the likelihood-ratio method.

   (d) For large-sample $(1 - \alpha)\%$ confidence intervals for μ:

   i. find the Wald interval,
   ii. set-up the score interval (*Hint: the set of μ_0 such that ...)*, and
   iii. set-up the likelihood-ratio interval.

2. (10 points) *Multinomial Marginal Distribution*: Suppose that $(n_1, n_2, ..., n_c)$ follows a multinomial distribution with parameters n and $\pi = (\pi_1, ..., \pi_c)$. Show that the marginal distribution of n_1 is binomial with parameters n and π_1.

3. (10 points) *Multinomial Distribution Correlation*: For the multinomial distribution, show that

$$corr(n_j, n_k) = -\pi_j \pi_k / \sqrt{\pi_j(1 - \pi_j)\pi_k(1 - \pi_k)}.$$

   When $c = 2$, show that this simplifies to $corr(n_1, n_2) = -1$. Explain why this special case makes intuitive sense.

4. (15 points) *Invariance and Association Measures*: For a 2x2 table of counts $\{n_{ij}\}$:

   (a) Show that the odds ratio is invariant to interchanging rows with columns, and multiplication of cell counts within rows or columns by $c \neq 0$.

   (b) Why is invariance a desirable property?

   (c) Show that the difference of proportions and the relative risk do not have these properties.

Show all calculations. Turn in R code if applicable.

5. (10 points) Each of 100 multiple-choice questions on an exam has four possible answers, one of which is correct. For each question, a student guesses by selecting an answer randomly.

   (a) What is the distribution of the number of correct guesses?

   (b) Find the mean and standard deviation of the distribution in part (a). Would it be surprising if the student made at least 50 correct responses? Why?

   (c) What is the distribution of (n_1, n_2, n_3, n_4) where n_j is the number of times the student picked choice j.

(d) Find $E(n_j)$, $var(n_j)$, and $corr(n_j, n_k)$. Interpret the correlation.

6. (15 points) In a clinical trial comparing a new drug to a standard drug, let π denote the probability that the new drug is better. Consider estimating π and testing $H_0 : \pi = .5$. In 20 independent observations, the new drug is better just one time.

   (a) Find the maximum likelihood estimate of π.

   (b) Conduct a Wald test and construct a 95% Wald confidence interval for π. Interpret the result.

   (c) Conduct a score test and construct a 95% score confidence interval for π. Interpret the result. *Hint: For the CI you can use equation 1.14 in the book or numerically/computationally solve. If you numerically solve then turn in your R code.*

   (d) Conduct a likelihood ratio test and construct a 95% likelihood-based confidence interval for π. Interpret the result. *Hint: Solve for the CI numerically/computationally. Turn in your R code.*

   (e) Construct an exact binomial test. *Hint: See Section 1.4.4 of the textbook. Note that the exact test involves calculating the probability of observing more extreme events based on the null binomial distribution.*

   (f) Comment on how the results from the different tests and confidence intervals compare.

7. (10 points) *2x2 Tables and Diagnostic Testing*: Let π_1 be the probability that a diagnostic test is positive given that a subject has the disease, and π_2 be the probability that a test is positive given that a subject does not have it. Let ρ be the probability that a subject has the disease.

   (a) Using this notation, state the sensitivity (also called the true positive rate), the specificity and the false positive rate. What is the relationship between specificity and false positive rate?

   (b) Derive the formula for the positive predictive value (the probability that a subject has the disease given that a test is positive).

   (c) Suppose that a diagnostic test has both sensitivity and specificity equal to .95 and $\rho = 0.005$. What is the probability that a subject has the disease given that the test is positive?

   (d) Repeat part (c) using $\rho = .10$. Describe how the probability that a subject has the disease given that the test is positive depends on the prevalence ρ.

8. (15 points) 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.

	Final Grade	
Study Time	A	B
≥ 10 hours/week	70	10
< 10 hours/week	2	12

   (a) Estimate the probability of a grade of A, conditional on study amount in category (i) ≥ 10 hours and (ii) < 10 hours.

   (b) Estimate the probability of studying ≥ 10 hours, conditional on final grade in category (i) A and (ii) B.

(c) What would be an appropriate sampling model for this example? What is the most natural choice of response variable?

(d) For the most appropriate choice of response variable, find and interpret the difference of proportions, relative risk and odds ratio. Is there an association between final grade and study amount? Justify your answer.

(e) Now consider a third "Final Grade" category of C/D/F, where 2 students studying ≥ 10 hours/week are in this grade category, and 4 students studying < 10 hours/week are in this grade category. Estimate the two local odds ratios θ_{11} and θ_{12}. Interpret these results.

(f) Still considering the $2x3$ table from part (e), find the third odds ratio θ_{13} using only the odds ratio results from part (e). Is θ_{13} necessary to fully describe the association in the $2x3$ table?

(g) Still considering the $2x3$ table from part (e), find the Goodman and Kruskal γ measure for association of ordinal variables. Interpret the result.

(h) Now consider only the students who studied ≥ 10 hours/week. Assuming the multinomial distribution, what are the maximum likelihood estimates of $\{\pi_1, \pi_2, \pi_3\}$?

(i) Still considering only the students who studied ≥ 10 hours/week, carry out (i) the Pearson chi-squared test and (ii) the likelihood-ratio chi-square test for $H_0 : \{\pi_1, \pi_2, \pi_3\} = \{.80, .19, .01\}$. Interpret the conclusions of these tests.