Problem Set 3 (two pages)

Statistics 24510-30040 (W19)

Due Tuesday, January 29 at the beginning of class.

Requirements Provide detailed derivations. Select only the relevant part of the output to be inserted. Attach your code or output as an appendix if necessary. Discussions allowed, the assignment should be devised and written by yourself completely.

Problem assignments (Related to Rice chapter 9 on likelihood ratio and chapter 11 on comparing two samples)

1. (*Geometric property of standardized bivariate normal*)

   Let (U, V) be a standardized bivariate normal random vector.
   Thus $E(U) = E(V) = 0$, $Var(U) = Var(V) = 1$, and $corr(U, V) = \rho$.

   (a) Prove that on the (u, v) plane, the contours or the level curves of the density $\{(u, v) : f(u, v) = constant\}$ are ellipses, and $u = v$ is the direction of the major axis.

   (b) Find the ratio of the lengths of major axis and minor axis of the elliptical contours.

2. (*Regression fallacy on test scores*)

   Let X and Y be the scores of a typical student on midterm and final exam in a class.
   We model these scores as

   $$X = S + E_1, \qquad Y = S + E_2$$

   where S, E_1, E_2 are independent random variables distributed as $S \sim N(70, 49)$, and $E_1, E_2 \sim N(0, 16)$. Let's think of S as a *skill* part of the score and E_1, E_2 as *luck* components. In the following, formulate your approach precisely and show your work.

   (a) What is the joint distribution of (X, Y)? Show your steps.

   (b) If a student received a midterm score of 60, what do you expect his/her final score to be?

   (c) If a student received a midterm score of 90, what do you expect his/her final score to be?

3. Suppose random variables X_1 and X_2 are independent with $X_i \sim Bin(n_i, p_i)$.

   We wish to test the hypothesis $H_o : p_1 = p_2$ versus the alternative hypothesis $H_a : p_1 \neq p_2$.
   (Assume the n_i's are known).

   (a) Derive the likelihood ratio test LR for this hypothesis.

   (b) What is the approximate distribution of $-2\log(LR)$ under the null hypothesis assuming that n_1 and n_2 are both large?

   (c) Now consider applying this test to the ups and downs of the Dow Jones Industrial Average in the 20th century. During 1900-1949, there were 7810 up (including unchanged) days out of 14,736, and during 1950-1999, there were 6688 up days out of 12,673.

   Under what assumptions on the sequence of up and down days would the null hypothesis in part (a) hold?

   (d) Using your result in part (a), find the numerical value for the likelihood ratio statistic for the hypothesis that the probability of an up day is equal for the two halves of the 20th century. Find the p-value for this test. What do you conclude?

   (Note: you may want to use the R command `pchisq` to find the p-value accurately.)

4. X and Y are independent Poisson random variables with $E(X) = \lambda_X$ and $E(Y) = \lambda_Y$. Consider the two sample problem for testing $\lambda_X = \lambda_Y$ versus $\lambda_X \neq \lambda_Y$.

   (a) Find the likelihood ratio test for this hypothesis.

   (b) When should you expect $-2 \log LR$ to follow approximately a χ^2 distribution? In that situation, what should be the degrees of freedom for this distribution?

   (c) Now suppose $Z \sim Binom(n, p)$. Find the likelihood ratio for testing $p = \frac{1}{2}$ vs. $p \neq \frac{1}{2}$.

   (d) Is there a relationship between the test in part (c) and the LRT in part (a)? Explain in terms of conditional distributions.

5. Let us assume that the number of occurrences of major earthquakes (magnitude at least 7 on the Richter scale) in the Northern and Southern hemispheres can be modeled as independent Poisson random variables. From 2001-2005, discounting obvious aftershocks, there were 34 major earthquakes in the northern hemisphere and 31 in the southern hemisphere.

   (a) For testing the hypothesis that these two Poisson random variables have the same mean versus the alternative of different means, find the numerical value for the likelihood ratio test statistic you found in 4(a), i.e., part (a) in Problem 4 in this assignment.

   (b) Using $n = 65 = 31 + 34$ and supposing that $Z = 34$ is the number of quakes in the northern hemisphere, find the value for the likelihood ratio test statistic you found in 4(c). How does this compare to the LR you found in part (a) in this problem?

   (c) Find the p-value for the test in part (a) using the asymptotic result in 4(b).

   (d) Fine the exact p-value for the test in part (b).

   (e) Which p-value do you think is more relevant for this problem?

   (f) Based on what you know or can easily look up about earthquakes, discuss briefly whether the Poisson and independence assumptions underlying the analysis in part (a) are plausible for these data.

6. (*Confidence intervals: Pooled versus paired*)

   Suppose $X_1, \cdots, X_n$ are $N(\mu_1, \sigma^2)$ random variables, and Suppose $Y_1, \cdots, Y_n$ are $N(\mu_2, \sigma^2)$ random variables. Note that only the means are different. Furthermore, assume that the pairs $(X_1, Y_1), \cdots, (X_n, Y_n)$ are independent. This problem compares two confidence intervals for $\mu_1 - \mu_2$, namely the *pooled interval* (ref. pp. 422-423 in Rice) and the *paired interval* (ref. pp. 446-447 in Rice).

   (a) Show that if the true correlation $\rho = Corr(X_i, Y_i) = 1$, then with probability one a *pooled interval* has length zero.

   (b) Assume now that the observations are such that the sample correlation coefficient is exactly equal to zero. Which confidence interval would be shorter, the *paired interval* or the *pooled interval* ? In your answer determine the ratio of the two interval lengths.