MATHEMATICS E-156, SPRING 2014
MATHEMATICAL FOUNDATIONS OF STATISTICAL SOFTWARE
Module #4 (Hypothesis Testing with Chi Square)

Last modified: February 23, 2014 (fixed dog-bite section problem)

Reading from Chihara and Hesterberg

- Sections 3.4.2 through 3.8
- Appedix B.6 (the Poisson distribution)
- Appendix B.3 (the multinomial distribution)

Optional Reading from Haigh

- Page 53 derives the Poisson distribution as the limit of the binomial distribution.

Proof of the Week

- Prove that the Poisson distribution with parameter λ has mean and variance both equal to λ. Prove that if X_1 and X_2 are independent Poisson random variables with parameters λ_1 and λ_2 respectively, then $X_1 + X_2$ is Poisson with parameter $\lambda_1 + \lambda_2$.

  (This is theorems B.5 and B.6 on page 380)

R scripts

- 4A-ChiSquare.R
  Topic 1 – when should we expect a chi-square distribution?
  Topic 2 – why do we sum over all the cells in the table?

- 4B-Homogeneity.R

  Topic 1 – creating a contingency table and doing a chi-square test
  Topic 2 – doing a permutation test with a chi-square statistic

- 4C-GoodnessFit.R

  Topic 1 - comparing data with a probability model

- 4D-GoodnessFitDistrib.R

  Topic 1 - hypothesized distribution has all parameters specified
  Topic 2 - hypothesized distribution has a parameter to be estimated from the data

- 4E-Multinomial.R

  Topic 1 - installing and using the multinomial test
  Topic 2 – using the multinomial test on examples from the textbook

- 4P-Proof 4.R

  Topic 1 - Poisson distribution as the limit of the binomial distribution
  Topic 2 – the sum of two Poisson distributions

- 4X-Hypergeometric.R (optional)

  Topic 1 - some fabricated examples witth small numbers
  Topic 2 – some real data: do Bush voters own guns?

Mathematical notes

1. Poisson distribution as the limit of a binomial distribution

   Random variable X_n has a binomial distribution with parameters n and p, expectation $\lambda = np$. So $p = \lambda/n$.

   Its mass function is

   $$P(X_n = x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} \frac{\lambda^x}{n^x} (1 - \frac{\lambda}{n})^{n-x}.$$

   Take the limit as $n \to \infty$ to get the mass function for a Poisson random variable X:

   $$P(X = x) = e^{-\lambda} \frac{\lambda^x}{x!}.$$

2. Proof of the week

   (a) Prove that the Poisson distribution with parameter λ has mean and variance both equal to λ.

   (b) Prove that if X_1 and X_2 are independent Poisson random variables with parameters λ_1 and λ_2 respectively, then $X_1 + X_2$ is Poisson with parameter $\lambda_1 + \lambda_2$.

3. Multinomial distribution

   (a) Suppose that we have n objects of r different types. Show that the number of ways of selecting x_1 objects of type 1, x_2 objects of type 2, $\cdots$ x_r objects of type r is

$$\frac{n}{x_1!x_2!\cdots x_r!}.$$

   (b) Prove that if the probability that an object is of type i is p_i, then the probabillity that x_1 objects are of type 1, $\cdots$ x_r objects are of type r is

$$P(X_1 = x_1, X_2 = x_2, \cdots X_r = x_r) = \frac{n!}{x_1!x_2!\cdots x_r!} p_1^{x_1} p_2^{x_2} \cdots p_r^{x_r}.$$

Section problems

1. Exercise 15 on page 72:

   For the Flight Delays Case Study in Section 1.1, conduct a test of homogeneity to determine if there is a relationship between carrier and the number of flights delayed more than 30 minutes (`Delayed30`). The answer is on page 401.

2. According to Stirzaker, *Elementary Probability*:
   "During 1979, in Bristol, 1103 postmen sustained 215 dog bites. A total of 191 postmen were bitten, of whom 145 were bitten just once."

   A member of the class has pointed out that this is impossible, since the other 46 postmen were bitten at least twice, making the total bites at least 237.

   So change the data to

   "During 1979, in Bristol, 1103 postmen sustained 315 dog bites. A total of 191 postmen were bitten, of whom 145 were bitten just once."

   Let random variable X be the number of dog bites sustained by an individual postman. If a dog, having decided to bite, chooses its victim at random, X will have a binomial distribution that is very well approximated by a Poisson distribution with $\lambda = 315/1103$. Carry out a goodness-of-fit test for this model.

3. The file PopDensity.csv contains the population density (per square mile) for every city in the U.S. whose poplulation is 100000 or more.

   (a) For the last digit, count the number of occurrences of each of the five pairs of digits 0-1, 2-3, 4-5, 6-7, 8-9 to get a table like the one at the top of page 64, then carry out a test of the hypothesis that the last digits are distributed uniformly. Hint: `234%%10` is equal to 4.

   (b) For the first digit, count the number of occurrences of each of the three triples of digits 1-2-3, 4-5-6, 7-8-9 to get a table like the one at the top of page 64, then carry out a test of the hypothesis that the first digits are distributed uniformly. Hint: `substr(as.character(234),1,1)` is the character "2".

Homework assignment This assignment should be submitted as a single R script. Include enough comments so that it is clear what you are doing and where each problem begins. You can upload it to the dropbox on the Class 4 page of the Web site.

It is OK to paste and edit lines from the scripts on the course Web site. It is not OK to paste lines from your classmates' solutions!

The first two problems have odd numbers. Feel free to check your answers against page 401.

1. Exercise 25 on page 73 of the textbook. Solve the problem by doing a chi-square goodness-of-fit test, then solve it by assuming all numbers equally likely and drawing lots of random samples of 500 numbers.

2. Exercise 21 on page 72.

   For the given pdf, it is easy to integrate the density function by standard calculus techniques to find the probabilities for each interval, as was done on page 65. Or you can have R do numerical integration.

   However, the given distribution is a special case of the "Pareto distribution," which is supported by R. If you execute these three lines:

   ```
   install.packages("actuar") #comment this out after the first time

   library(actuar)

   help("ppareto")
   ```

   you will be able to have R compute the probabilities for the different intervals by subtraction.

3. Exercise 14 on page 71 Do this two ways:

   (a) Replicate the contingency table (including the row and column names), then use the built-in chi-square test.

   (b) Create a data frame with two columns, then do a permutation test by permuting the Sex column, using chi square as a measure of distance between observed and expected values.

4. Exercise 22 on page 73. The fifty numbers are in the file Exercise22.csv.

5. For the Red Sox data in RedSox2013.csv, consider the number of runs per game scored by the Red Sox. Model this using a Poisson distribution, and perform a goodness-of-fit test to compare the model with the empirical data.