Statistics and Data Analysis

Lecture 8

Lecture by Dr. Zohar Yakhini
Typeset by Steven Karas

2016-12-27
Last edited 17:59:26 2016-12-28

1 Review: Correlations

Mathematical measure of the strength of a relationship between two variables. We discussed assessing statistical significance. A relationship can be strong, but not significant. A relationship can be weak, yet significant. The key factor is the size of the sample. Small samples can produce strong correlation by chance. Large samples can easily show significance, but we need to evaluate this together with relationship strength.

2 Spearman Correlation - Derivation

Convert the u_is to x_is, and the r_is to y_is (ranked).

$Sp(u,r) = Pearson(x,y)$

$$\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\left(\sum(x_i-\bar{x})\sum(y_i-\bar{y})\right)^{1/2}}$$

$$Sp(\rho) = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

where $d_i = x_i - y_i$

$$\frac{1}{2}\sum d_i^2 = \frac{1}{2}\sum(x_i-y_i)^2 = \frac{1}{2}\left(\sum x_i^2 - 2\sum x_i y_i + \sum y_i^2\right)$$

$$= \sum x_i^2 - \sum x_i y_i$$

$$\frac{\sum x_i y_i}{\sum x_i^2} = 1 - \frac{\frac{1}{2}\sum d_i^2}{\sum x_i^2}$$

$$x_i \to a_i = x_i - \frac{n+1}{2}$$

$$y_i \to b_i = y_i - \frac{n+1}{2}$$

Proof of this by induction is left as an exercise:

$$\sum_{i=1}^{k} i^2 = \frac{k(k+1)(2k+1)}{6}$$

$$\sum_{i=-\frac{n-1}{2}}^{\frac{n-1}{2}} = \sum_{i=1}^{\frac{n-1}{2}} 2i^2 = 2 \cdot \frac{1}{6}\frac{n-1}{2}\frac{n+1}{2} \cdot n$$

$$= \frac{1}{12}n(n^2-1)$$

3 Convolution of two distributions

An example of the function ConvDbns and it's implementation in matlab was given on the board.

$$c = conv(u, r)$$

$$c_i = \sum_j u_j r_{i-j}$$

$$P(X+Y=1) = \sum_j P(X=j)P(Y=i-j)$$

3.1 N-Fold convolution

trivial to implement using a generic convolution.

4 Hypergeometric distribution

Given a class of 40 students, of whom 10 got an A on the exam. The last names of 8 of them begin with a letter in the range A-M.

4.1 Example: seating

$N = 40$ seats $B = 30$ students $n = 20$ students on right half of class $b \geq 20$ sitting on the right

$$\frac{X \sim Binom(40, 1/2)}{P(X=30)}$$

$$HG(N,B,n,b) = \frac{\binom{B}{b}\binom{N-B}{n-b}}{\binom{N}{n}}$$

4.2 Example: cold treatment

	yes	no	Total
1-3 days	86	19	105
4-7 days	16	79	95
Total	102	98	200

We want to evaluate this treatment via the HG null model:

$N = 200$ $B = 105$ people whose cold lasted < 3 days $n = 102$ people who took the treatment $b = 86$

$$HGT(N, B, n, b) = \sum_{t=b}^{\min(n,b)} HG(N, B, n, b)$$

For this example, the p value is 10^{-15}

5 Contingency tables with finer partitions

5.1 Example: Lotto

6 of 42 numbered balls are drawn at random without replacement. You wrote 6 on your card ahead of time. How many match?

$$HG(42, 6, 6, 0) = \frac{\binom{6}{0}\binom{36}{6}}{\binom{42}{6}} \approx 0.37$$

$$HG(42, 6, 6, 1) \approx 0.43$$

$$HG(42, 6, 6, 5) \approx 0.00004$$

$$HG(42, 6, 6, 6) \approx 0.00000002$$

$$HG(42, 7, 6, 6) = 7 \cdot HG(42, 6, 6, 6)$$

6 Multi-Hypergeometric

k different types of items total of m items m_i items of type i We select n out of m with all possible such selections being equiprobable

Example: shapes 3 circles, 5 squares, 9 triangles, choosing 3 items:

$m_1 = 3 \quad m_2 = 5 \quad m_3 = 9 \quad k = 3$

$$P(Y_i = j_i \mid \forall i \in [1, k])$$

7 Confidence intervals

Reminder: Wald confidence interval:

$$p^* \pm \Phi^{-1}(0.95)\sqrt{p^* \cdot q^*/n}$$

Example: cold treatment continued The confidence interval is around 0.055, but we observed 0.64. Therefore, we can reject this with a p-value of virtually 0:

8 Comparing two samples with student's t-test

Matlab function ttest2.

9 HW3

9.1 Haberman data

Data from chicago hospitals for breast cancer surgeries.

Age, year of operation, Number of positive axillary nodes detected, Survival status (1=survived $>$5 years, 2=died within 5 years)

Do patients with more than 10 positive nodes have lesser survival?