Lab session week 18 : Exercises Chapter 11

May 11, 2016

We observe the following $n = 20$ sized vector sample of a variable $Y \in \{y_1, \cdots, y_k\}$

$$(Y_1, Y_2, \cdots, Y_{20}) = (16,\ 8,\ 13,\ 14,\ 11,\ 13,\ 11,\ 19,\ 11,\ 5,\ 14,\ 13,\ 4,\ 8,\ 12,\ 4,\ 13,\ 9,\ 7,\ 9\,).$$

We want to test the hypothesis H_0 that the sample comes from an uniform distribution between its minimum and maximum values observed.

Exercise 1 - Chi-Square Goodness of Fit Test

Recall that the statistic of test T is defined as

$$T := \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i},$$

where N_i denotes the number of the Y_j 's that equal y_i and p_i the probability that Y equals y_i, e.i.

$$p_i = \mathbb{P}_{H_0}(Y = y_i).$$

Call t the observed value of T, the *p-value* is defined as

$$p - value = \mathbb{P}_{H_0}(\chi^2_{k-1} > t),$$

where χ^2_{k-1} is a chi-square random variable with $k - 1$ degrees of freedom.

- Describe the set $\{y_1, \cdots, y_k\}$
- What the values of p_i?
- Approximate the p-value by using the chi-square approximation.
- What this value means ? What can we conclude ?

Hint: you can use the `hist` function to avoid the counting part. The distribution for a chi-square distribution is `pchisq`.

Exercise 2 - The Kolmogorov-Smirnov Test

Let F_e be the empirical distribution function defined by

$$F_e(x) = \frac{\#i : Y_i \leq x}{n}.$$

Assuming that F is the distribution of Y under the hypothesis H_0, we should expect that $F_e(x)$ is close enough to $F(x)$ for any x. Thus, a natural quantitie underlying the goodness of fit is

$$D := \sup_x |F_e(x) - F(x)| = \max\left\{\frac{j}{n} - F(Y_{j,n}), F(Y_{j,n}) - \frac{j-1}{n}; j = 1 \cdots, n\right\},$$

where $Y_{j,n}$ is the j-smallest value among $\{Y_1, \cdots, Y_n\}$.

Call d the observed value of D, it follows that the p-value for this test is given by

$$p - value = \mathbb{P}_{H_0}(D \geq d).$$

One useful result state that the distribution F has no impact on the p-value, e.i.

$$\mathbb{P}_{H_0}(D \geq d) = \mathbb{P}\left(\max_{x \in [0,1]} \left| \frac{\# i : U_i \leq x}{n} - x \right| \geq d \right),$$

where $U_1, \cdots, U_n$ are i.i.d uniform variables on $[0, 1]$. Using the relation

$$\max_{x \in [0,1]} \left| \frac{\# i : U_i \leq x}{n} - x \right| = \max \left\{ \frac{j}{n} - U_{j,n}, U_{j,n} - \frac{j-1}{n}; j = 1, \cdots, n \right\},$$

where $U_{j,n}$ is the j-smallest value among $\{U_1, \cdots, U_n\}$, we want to estimate the p-value in the context of the previous exercise.

- What is F under H_0 ?
- Compute the value $D = d$ for the obseved vector $(Y_1, Y_2, \cdots, Y_{20})$.
- Generate $N = 500$ samples of a n-sized sample with uniform distribution. How can you use this latter to estimate $\mathbb{P}_{H_0}(D \geq d)$?
- Compare it with the previous p-value.
- Graphically compare the functions F and F_e.