PC9 – Chi-square tests

November 5 2018

Exercise 1: Chi square goodness of fit test for cross-bredding plants

Two different types of plants are cross-bred. They differ by two traits ; the first trait can be A or a, the second trait can be B or b. The first generation of cross-bredding plants is homogeneous: all the plants have the following genotype $AaBb$. We question the following model:

- A is dominant and a is recessive,
- B is dominant and b is recessive.

By the Mendelian inheritance, this model would lead to a second generation for which the 4 phenotypes AB (genotype $AABB$, $AaBB$, $AABB$ or $AaBb$), Ab (genotype $AAbb$ or $Aabb$), aB (genotype $aaBB$ or $aaBb$) and ab (genotype $aabb$) would have the following respective frequencies 9/16 , 3/16 , 3/16 and 1/16.

Yet, with a sample of 160 plants, we observe 100 phenotypes AB, 18 phenotypes Ab, 24 phenotypes aB, Ab and 18 phenotypes ab.

1. Give the statistical model.
2. Write the likelihood in this model.
3. Propose an estimator for the parameter of your model using the method of moments.
4. Test the considered model at level $\alpha = 0.05$.
5. What can you say about the p-value associated with the observed result? In other words below which level, this result does not lead to reject the considered level? You can use the R function qchisq to obtain the χ^2 distribution quantiles.
6. Check that you obtain the same result using the function **chisq.test** of R.
7. Answer Question 2 in the case where the sample has 80 plants, and we observe 50 phenotypes AB, 9 phenotypes Ab, 12 phenotypes aB and 9 phenotypes ab (i.e. with the same proportion as in the previous data).

Exercise 2: Chi square independence test for burgers

The Mac Burger's society launches its new burger FolBurger in the US and in Europe. It does a survey by asking for a feedback (bad, correct and good) to customers living in four cities. They obain the following answers:

	Bad	Correct	Good
Los Angeles	29	124	87
Chicago	74	278	208
Madrid	114	277	87
Paris	182	417	123

The society wants to test the dependence of feedbacks with the place of residence of the customer.

1. Give the statistical model.
2. Enter the data in a matrix called "tab" and give names to rows and columns. Compute the totals by rows and culumns and merge it with the previous matrix in a new matrix called tab2.
3. Explain what is displayed in R when compiling the following instructions:
   - "prop.table(tab)",

- "prop.table(tab,margin=1)",
- "prop.table(tab,margin=2)".

4. Display the six barplots of the feedbacks of customers living in Los Angeles, Chicago, the US, Madrid, Paris and Europe.
5. Comment these histograms.
6. Compute test at level $\alpha = 0.01$ to answer the following questions.
   (a) Do the feedbacks on FolBurger depend on the place of residence of the customer?
   (b) Are the feedbacks of customers living in Los Angeles different from those living in Chicago?
   (c) Are the feedbacks of customers living in Madrid different from those living in Paris?
7. Do the same tests using the R function chisq.test.
8. What do you conclude?

Exercise 3: About the asymptotic distribution of the test statistic in chi square goodness of fit tests

Let X_i be i.i.d. random variables with values in $\{1, 2, \cdots, k\}$. Under our model, the distribution of X_i depends on the parameter $\theta = p \in \Delta_{k-1} := \{p \in (\mathbb{R}_+)^k : \ p_1 + p_2 + \cdots + p_k = 1\}$. The distribution of X_i is denoted P_p. Under P_p, the X_i are i.i.d. and $P_p(X_i = c) = p_c$ for $c \in \{1, 2, \cdots, k\}$. So that, the statistical model is

$$\left(\{1, 2, \cdots, k\}^n, \mathcal{P}(\{1, 2, \cdots, k\}^n), \{P_p : p \in (\mathbb{R}_+)^k : \ p_1 + p_2 + \cdots + p_k = 1\}\right).$$

Let p_0 be in Δ_{k-1}. We want to test

$$H_0 : \ p = p_0 \ \text{ against } \ H_1 : \ p \neq p_0 \, .$$

It is a test of goodness of fit. In the lecture, the following test statistic is proposed:

$$S(X_1, \cdots, X_n) = \sum_{j=1}^{4} \frac{(N_j - e_j)^2}{e_j},$$

where $e_j = (p_0)_j n$ and $N_j(X_1, \cdots, X_n) = \#\{i \in \{1, \ldots, n\} : \ X_i = j\}$, $j \in \{1, \cdots, k\}$.

1. Under H_0, S is said to be asymptotically distributed from a $\chi^2(k-1)$. Check numerically this property with R, when $p_0 = (0.2, 0.3, 0.5)$ and $n = 100$. First create a function in R with inputs p_0, x and n and output $S(x)$.
2. What happens when n is not large enough. Check the ditsribution of S when $p_0 = (0.002, 0.3, 0.5, 0.198)$ and $n = 100$.