5. Practice – 2019

Non Parametric hypothesis tests; Binomial distribution

The following non parametric tests are used in this practice:

NAME OF THE TEST	WHAT IS IT GOOD FOR?
Goodness of fit test with Chi-square test	Is the random variable follows a given distribution?
Homogeneity test with Chi-square test.	Are two random variables follow the same distribution?

Goodness of fit test with Chi-square test:

The observed value:

$$\chi^2_{obs} = \sum_{i=1}^{r} \frac{(\nu_i - Np_i)^2}{Np_i},$$

Were:

- r is the number of possible outcomes,
- ν_i is the frequency of the i^{th} outcome
- N is the sample size. (Remember that $\sum_{i=1}^{r} \nu_i = N$),
- p_i is the theoretical probability of the i^{th} outcome.

The critical value can be calculated with a Excel function:

$$\chi^2_{crit} = CHISQ.INV.RT(1-p; f),$$

Where

- p is the significance level,
- r is the number of possible outcomes/ or class divisions
- f is the degree of freedom: $f = r - 1 - k$
    - k is the number of estimated parameters.

Homogeneity test with Chi-square test:

The observed value:

$$\chi^2_{obs} = nm \sum_{i=1}^{r} \frac{\left(\frac{\nu_i}{n} - \frac{\mu_i}{m}\right)^2}{\nu_i + \mu_i},$$

Were:

- r is the number of possible outcomes,
- n is the sample size of the first set of sample,
- ν_i is the frequency of the i^{th} outcome of the first set of sample, hence $\sum \nu_i = n.$
- m is the sample size of the second set of sample,
- μ_i is the frequency of the i^{th} outcome of the second set of sample, hence $\sum \mu_i = m.$

The critical value is the same as in the good test fit.

We accept H0 if $\chi^2_{obs} < \chi^2_{crit}$.

Usually the class intervals are constructed in a way that the frequency of most of the interval is at least 5. The width of the class intervals do not need to be the same.

1. Exercise

   At a telephone exchange experience shows that the time passed between dialling and connection is in between 25 and 85 seconds. For a whole day, the connection time was recorded with seconds' precision. The table below contains the collected data. Can we say that the connection time follows a uniform distribution with a 95% probability?

   Solution:
   Test: Goodness of fit test with Chi-square test.
   H0: The connection time follows a uniform distribution (p=0,95).

   For the solution, we consider the connection time to be a discrete variable. For the observed value we need the possible outcomes and their frequency, so lest make two new columns for them:

   - **D9: Possible outcomes**
   - **E9: frequency**
   - **D10=25, D11=26**, and pull these two down until you reach D70=85
   - **E11=COUNTIF(C$10:C$117;"="&D10)**, and pull down.
   - We see that the frequency of most of the intervals is less than 5. Therefore, we need to merge classes. In the Excel file, you find a possible solution for that.
   - Find the frequency of the new classes.

   The theoretical probabilities are the same for every connection time:

   - The number of different connection times is **m** and the probability of one connection time is **1/m.**

- The theoretical probability of the different classes is **l*1/m**, where **l** is the number of connection times in that class. This is calculated in column **K**.
- Now we can compute the observed value for each class. The sum of these values is the observed value of our test.
- The observed value: 19.36
  The critical value: **CHISQ.INV.RT (1-0.95 ; r-1)** =27.59
  We did not estimate any parameters of the distribution, hence f=r-1.
  Since Chi^2_obs < Chi^2_crit, we accept H0.

2. Exercise

   The numbering of a dice was changed so that one side of the dice has the number 1, two sides have the number 2 and the remaining three sides have 3 written on them. Can we state with a probability of 95% that the probabilities of rolling 1, 2, 3 are 1/6, 2/6 and 3/6, respectively?

   Note: To roll the dice several times, a random number generator was used, the results are in column C. To answer the question it is not necessary to know how the generator works!

   Solution:
   Test: Goodness of fit test with Chi-square test.
   H0: The data follows the given distribution (p=0,95).

   To calculate the observed value, fill out the table.

   - The *possible outcomes* are easy: 1, 2 and 3.
   - The *frequencies*: **F11=COUNTIF($C$11:$C$559;E11)**, and pull it down.
     - Note: Sum up the frequencies in F14 to get the sample size.

- The *relative frequencies* are not needed to calculate the observed value, but to compare it with the *theoretical probabilities*, calculate them in the column G: G11=F11/F$14, and pull down.
    - Note: The sum of the relative frequencies is always 1.
- Fill out the column of the theoretical probabilities: H11=1/6, H12=2/6, H13=3/6.
    - Note: The sum of the theoretical probabilities is always 1.
- For the observed value calculate the fractions behind the summa in column I: I11=(F11-F$14*H11)^2/(F$14*H11), and pull it down.
- Finally in I14 sum up the values in I11-I13: I14=SUM(I11:I13), this is the observed value.
    - Note: With every ENTER (more precisely every time your excel file is refreshed) the dice rolls are regenerated, hence χ^2_{obs} changes every time. It is not a problem at all.

For the critical value: H20=CHISQ.INV.RT(1-0,95;3-1)=5,99.
Again, we did not estimate any parameters of the distribution, hence f=r-1.

We accept H0 if $\chi^2_{obs} < \chi^2_{crit}$.

3. Exercise

A company packages a particular product in cans of three different sizes, each one using a different production line. Most cans conform to specifications, but a quality control engineer has identified the following reasons for non-conformance:

1. Blemish on can
2. Crack in can
3. Improper pull tab location
4. Pull tab missing
5. Other

A sample of nonconforming units is selected from each of the three lines, and each unit is categorized according to reason for nonconformity, resulting in the following contingency table data. Does the data suggest that the proportions falling in the various non-conformance categories are the same for the three lines with a probability of 95%?

Solution:
Test: Homogeneity test with Chi-square test.
H0: The proportions falling in the various nonconformance categories are the same for the three lines with a probability of 95%. (Test this for the pairs of lines.)

For the computation see the Excel file.

4. Exercise

A highchair manufacturing company follows a design guideline where they assume that

a) the height of the 10-12 age old children follows a normal distribution (p=95%), and
b) the average height of the girls and the boys are not significantly different (p=98%).

A research group examined 40 boys and 40 girls from the relevant age group. Their result can be found below. Are their findings suggests that the company's assumption is correct?

Solution:

a) Test: Goodness of fit test with Chi-square test.
H0: The height of the children follows a normal distribution (p=0,95).

For the observed value we have everything but p_i, that is the probabilities that the normal distributed random variable falls into the i^{th} interval. These probabilities can be determined from the distribution function:

$$p_i = P(x_{i-1} \leq \xi \leq x_i) = F(x_i) - F(x_{i-1}).$$

where $F(x) = P(\xi < x)$ is the distribution function of ξ.
Let's calculate the value of $F$ in both boundaries of each interval with the built in function **NORM.DIST(x, expected value, standard deviation,1)**. In the case when the mean and the standard deviation are not known, an unbiased estimation of them is the average and the sample standard deviation, respectively. Not that these are two estimated parameters, hence the degree of freedom will decrease with 2.

- For the lower boundaries: **G8=NORM.DIST(D8;D$18;D$19;1)**, and pull down.
- For the upper boundaries: **H8=NORM.DIST(E8;D$18;D$19;1)**, and pull down.
- Then p_i: **I8=H8-G8**, and pull down.

Now if we sum up the values of p_i we do not get 1, and that is a problem. To overcome this, we write **0 in G8**, and **1 in H16**. Why? Because the probability of having a child shorter than 130 m is 0, and all of the children are shorter than 161 cm, hence the probability of having a child shorter than 161 cm is 1.
Now we can calculate the observed value:

- In F18 calculate how many kids were measured (N): **F18=SUM(F8:F16)** =80
- **J8=(F8-$F$18*I8)^2/$F$18/I8**, and pull down.
- Then χ^2_{obs}: **J18=SUM(J8:J16)** =5,31.

The critical value, χ^2_{crit}: **J20=CHISQ.INV.RT(1-0,95;9-1-2) =**12.59
We estimated two parameters, namely the mean and the standard deviation of the normal distribution, hence the degree of freedom is f=r-1-2
Since $\chi^2_{obs} < \chi^2_{crit}$, we accept H0.

b) Test: Welch-test
H0: The mean of the heights of the boys and the girls are the same, p=99%.

5. Exercise
A company's deliverers compete with each other. The company wants to know if the failure susceptibility of the delivered parts are the same in the case of the two deliverers, hence they took a sample from both of the delivered parts and examined the failure susceptibilities. Can they say that failure susceptibilities of the delivered parts of the two deliverers are the same with a probability of 99%?
The table below shows the lifetime of the parts in the two samples.

Solution:
Test: Homogeneity test with Chi-square test.
H0: The failure susceptibilities are the same (p=0,99).

To calculate the observed value, we first have to determine the possible outcomes, and then the frequencies of them. Since the failure susceptibility is a continuous variable, the outcomes will be intervals, and the goal is to create such intervals that the joint frequencies would be about the same in each interval.

- Let’s copy the lifetimes one under the other in column G, this is the joint data.
- Since there are 35 lifetimes, we will create $\sqrt{35} \approx 6$ intervals (just as we learned at the histogram).
- Make a table similar to this:

Interval	Lower boundary	Upper boundary	frequency of the 1st deliverer	frequency of the 2nd deliverer	Chi^2
1					
2					
3					
4					
5					
6					
		Total:			

- The boundaries will be determined so that each interval would contain about 1/6th of the 35 lifetimes:
  - For the lower boundaries: J15=PERCENTILE.INC(G$11:G$45;(I15-1)/6), and pull down.
  - For the lower boundaries: K15=PERCENTILE.INC(G$11:G$45;I15/6), and pull down.
- Now we have the possible outcomes. Count for both of the deliverers how many measurements fall in each interval:
  - For the first deliverer: L15=COUNTIF(C$11:C$25;”<”&K15)- COUNTIF(C$11:C$25;”<”&J15), and pull down.
  - For the second deliverer: M15=COUNTIF(E$11:E$30;”<”&K15)- COUNTIF(C$11:C$25;”<”&J15), and pull down.
  - Sum these frequencies up to get the original sample sizes (15 and 20). Pay attention that with the formulas above we do not count those measurements that fall on the upper boundary of the last interval!
    L20=COUNTIF(C$11:C$25;”<=”&K20)- COUNTIF(C$11:C$25;”<”&J20)
    M20=COUNTIF(D$11:D$30;”<=”&K20)- COUNTIF(D$11:D$30;”<”&J20)
- Now we can calculate the fractions behind the summa for each outcome:
  - N15=(L15/L$21-M15/M$21)^2/(L15+M15), and pull down.
- We have everything for the observed value:
  - M23=Chi^2_obs
  - N23= L21*M21* SUM(N15:N20) = 1,788
- For the critical value:
  - M25=Chi^2_crit
  - N25= CHISQ.INV.RT(1-0,99;6-1) = 15,086
- Since Chi^2_obs < Chi^2_crit, H0 is accepted.

6. Exercise

A concern of car manufacturers carried out a lifetime examination on a sample of 200 H7 halogen headlights. During the examination all the headlights were placed on a big panel, where they could be switched on and off the same time. After every 250. time they counted the defected headlights. The test ended after 3750 switching, when 6 headlights were still working.

Determine if the lifetime-distribution follows an exponential distribution with a probability of 90%! The distribution function is $F(x) = 1 - e^{-x/1200}$ where x denotes the number of switching-ons.

Solution:
Test: Goodness of fit test with Chi-square test.
H0: The lifetime-distribution follows an exponential distribution with a probability of 90%

This exercise is similar to Exercise 4/a, but now the distribution function is $F(x) = 1 - e^{-x/1200}$ and not normal distribution. When calculating the theoretical probabilities for each interval, one has to take into account that the distribution is a continuous distribution, hence we have to create subintervals so that their union would be the whole interval. So instead of [1,250], [251,500]… use the intervals (0,250], (250,500],…

7. Exercise

120-person grade is before the exam term. Everybody has to take 8 oral exams, where they get one question to talk about. The grade finds the material for the exams too much, so they decided to only study 80% of the material for each of the exams, and if they got asked about the rest of the questions they fail.
Let's simulate the results of the exams with a random number generator, and determine how many student pass all the exams and how many fail 1, 2,..8 exams. From these frequencies calculate the relative frequencies and compare the results with the theoretical probabilities calculated from the binomial distribution.

Solution:

This exercise is about the binomial distribution and not a hypothesis test.
When a random experiment is independently repeated several (n) times, one may ask how many times (k) did we get the same result (successful result). We can give an answer to the question with the binomial distribution.
Let ξ denote the number of successful result. When we repeat the experiment n times and the probability of a successful outcome is p, the probability of having k successful outcomes out of n:

$$P(\xi = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Where

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

In Excel:

- $n! = \boldsymbol{FAKT}(n)$, this only works when $n < 170$.
- $\binom{n}{k} = \boldsymbol{COMBIN}(n; k)$
- $P(\xi = k) = \boldsymbol{BINOM.DIST}(k; n; p; 0)$, this cannot be used in the test
- $P(\xi < k) = \boldsymbol{BINOM.DIST}(k; n; p; 1)$, this cannot be used in the test

1. First let’s calculate the theoretical probabilities of k successful exams of a student when only 80% of the material is studied in:
    - R14= COMBIN(8;O14)*$B$10^O14*(1-$B$10)^(8-O14), and pull down.
    - Check that the sum of the probabilities. It must be 1.

2. To simulate the exams of the 120 students we will use the function RAND(). This function gives a random number in the interval [0,1) and the probability of every number is the same. So a student passes an exam if RAND()<p, and fails otherwise. When an exam is passed we will put ☺ in the corresponding cell, and when it is failed we will put ●:
    - C14= IF(RAND()<$B$10;$E$10;$F$10), pull this formula to the right and then down.
3. We are interested in the number of passed exams, so in column K count the number of ☺ for every student:
    - K14= COUNTIF(C14:J14;$E$10), and pull down.
4. Now we can count the frequencies of the number of passed exams in column P, and then the relative frequencies:
    - P14= COUNTIF($K$14:$K$133;O14), and pull down.
        - The sum of the frequencies must be 120.
    - Q14=P14/120, and pull down.
        - The sum of the relative frequencies must be 1.
5. When pressing the button F9 we can generate the results again and again. Doing so watch the diagram how the relative frequencies fluctuate around the theoretical probabilities.