chi2\_test.R

pranav

2020-10-09

CHI SQUARE TESTING

#

Chi square test can be applied to check if two attributes are independent or not. The data here is the count of the observations with presence of both, either one or none of the attributes.

#  
#

The hypotheses for such a test are as follows…

# H\_0: The attributes are independent / not associated  
# H\_1: The attributes are dependent / associated  
# (Unlike correlation, we do not check if the association is negative or positive)  
#

Chi square test can also be applied to check if a sample can be said to be drawn from a population that follows a certain probability distribution. The data here is the observed frequencies, from which we derive expected frequencies (assuming a certain population proportion or probability distribution of the population).

#  
#

The hypotheses for such a test are as follows…

# H\_0: The sample's population follows the hypothesised proportion or probability distribution  
# H\_1: Opposite of H\_0, lol  
# (The conclusion states whether the assumed probability distribution can apply to the sample or not)

QUESTION 1

#

Here, we test whether family type and anxiety level are associated. The data matrix for the given sample is as follows…

q1\_data = matrix(data = c(35, 42, 61, 48, 51, 68), ncol = 3, byrow = TRUE)  
q1\_data

## [,1] [,2] [,3]  
## [1,] 35 42 61  
## [2,] 48 51 68

The rows represent family type (joint, nuclear) and columns represent anxiety level (low, normal, high). Performing the Chi square test on the matrix…

chisq.test(q1\_data)

##   
## Pearson's Chi-squared test  
##   
## data: q1\_data  
## X-squared = 0.53441, df = 2, p-value = 0.7655

We get p = 0.7655 > 0.05. Hence, H\_0 may be accepted. In other words, we may say that family type and anxiety level are not associated.

#  
#

QUESTION 2

#

Here, we test whether gender is associated with ice-cream preference. The data matrix for the given sample is as follows…

q2\_data = matrix(data = c(100, 120, 60, 350, 320, 150), ncol = 3, byrow = TRUE)  
q2\_data

## [,1] [,2] [,3]  
## [1,] 100 120 60  
## [2,] 350 320 150

The rows represent the gender (men, women) and the columns represent the ice-cream flavours (chocolate, vanilla, strawberry). Performing the Chi square test on the matrix…

chisq.test(q2\_data)

##   
## Pearson's Chi-squared test  
##   
## data: q2\_data  
## X-squared = 4.3195, df = 2, p-value = 0.1154

We get p = 0.1154 < 0.05. Hence, H\_0 may be rejected. In other words, we may reject the hypothesis that gender and ice-cream preference are not associated, and accept that they are associated, at least in the sample.

#  
#

QUESTION 3

#

Here, we check, in a sample of 5 children families, whether the probability of male births equals that of female births. This question is to do with fitting of a probability distribution to a sample. Here, we hypothesise that the probability of male births equals the that of female births. In other words, we hypothesise that the probability of success (which could mean either male or female birth) is 0.5.

#  
#

Hence, our H\_0 and H\_1 are

#  
# H\_0: The probability of male birthe quals 0.5  
# H\_1: The probability of male births is not equal to 0.5  
# (Note that we assume that there are only male or female births)  
#

Observed frequencies for male births out of 5 births in the following order: 5, 4, 3, 2, 1, 0

q3\_obf = c(14, 56, 110, 88, 40, 12)  
q3\_obf

## [1] 14 56 110 88 40 12

For expected frequencies, we use binomial distribution, because each trial (birth) has two possible outcomes (here, we classify male birth as success, female birth as failure, purely in a statistical context), with a fixed probability of success (at least, it is assumed so). Hence, we have the following paramters…

q3\_p = 0.5 # Probability of success  
q3\_N = sum(q3\_obf) # Total trials

Expected frequencies for male births out of 5 births in the following order: 5, 4, 3, 2, 1, 0

q3\_exf = dbinom(c(5, 4, 3, 2, 1, 0), size = 5, prob = q3\_p) \* q3\_N  
q3\_exf

## [1] 10 50 100 100 50 10

Checking of the sums of observed and expected frequencies are equal…

sum(q3\_obf)

## [1] 320

sum(q3\_exf)

## [1] 320

Calculated Chi square value is given by

# X^2 = sum((obf - exf)^2 / exf)  
q3\_X2 = sum((q3\_obf - q3\_exf)^2 / q3\_exf)  
q3\_X2

## [1] 7.16

To check the table X^2 value, we need the degrees of freedom, and significance level…

df = 6 - 1 # Degrees of freedom  
a = 0.05 # SIgnificance level  
qchisq(1-a, df)

## [1] 11.0705

We see that calculated X^2 < table X^2. Hence, we may accept H\_0. In other words, we may say that the probability of male births is 0.5. Hence, we may say that the probability of female births equals that of male births.

#  
#

QUESTION 4

#

Here, given the following data, we check if me winning PvP Minecraft battles follows binonmial distribution. Note that the order of number of wins in a row is 5, 4, 3, 2, 1, 0

q4\_obf = c(2, 5, 8, 10, 20, 14)  
q4\_x = c(5, 4, 3, 2, 1, 0)  
q4\_sum\_f = sum(q4\_obf)  
q4\_sum\_f # Winning streaks

## [1] 59

q4\_sum\_xf = sum(q4\_obf \* q4\_x)  
q4\_sum\_xf # Total wins

## [1] 94

q4\_xbar = q4\_sum\_xf / q4\_sum\_f # Average size of winning streak  
# But, in binomial distribution, x\_bar = np  
# Here, n = max no. of trials = 5, p = probability of win = ?  
# p = x\_bar / n  
q4\_p = q4\_xbar / 5  
q4\_p

## [1] 0.3186441

Hence, the expected frequencies are…

q4\_exf = dbinom(c(5, 4, 3, 2, 1, 0), size = 5, prob = q4\_p) \* q4\_sum\_f  
q4\_exf

## [1] 0.1938122 2.0721415 8.8617115 18.9489789 20.2592806 8.6640753

Checking if sum of observed frequencies equals the sum of expected frequencies…

sum(q4\_obf)

## [1] 59

sum(q4\_exf)

## [1] 59

Calculated Chi square value is given by

# X^2 = sum((obf - exf)^2 / exf)  
q4\_X2 = sum((q4\_obf - q4\_exf)^2 / q4\_exf)  
q4\_X2

## [1] 28.56894

To check the table X^2 value, we need the degrees of freedom, and significance level…

df = 6 - 1 # Degrees of freedom  
a = 0.05 # SIgnificance level  
qchisq(1-a, df)

## [1] 11.0705

We see that calculated X^2 > table X^2. Hence, we may reject H\_0. In other words, we may not say that me winning in PvP Minecraft battles follows binomial distribution. Very sad times.