homework2

Tamires Amorim, Meirou Guan, Yamei Li

9/20/2020

Lab # 1

Assessing the fairness of the dice, the goal of this project is to judge the Experiment Protocols built for assessing if the die is fair or not fair and draw conclusions from the findings based on statistics theory. We first judged the Experiment Protocols as follow:

PP1. For the first protocol, the group came to a conclusion that there is not there is not enough information about the die, for instance, the material it is made of, if it had two sides with the number six with a probability of 2/6, or had any environmental influence when rolling that could result in biased outcomes. When observing only one roll, it would return two possible outcomes a 6 with probability 1/6 or 16.66% or any other number with probability of 5/6 or 83.33%. It is also observed that the statement “if get 6 then conclude the dice is not fair”, is wrong since we cannot prove with a 100% certainty that the dice is unfair, given that we work with probabilities and there is always a chance of making wrong conclusions in accepting or rejecting the null hypothesis (H0).

PP2. The group decided to do Empirical Observations and use R to judge the fairness of the die. First, it was assumed the die was fair and after rolling it 30 times on R, we would be able to establish the comparison with the outcomes expected for a fair die. In other words, we would expect each number on the die to be rolled about 1/6, with expected value of the random variable X defined as: E(X)=np =30\* 1/6 =5. Thus, the fairness was evaluated following the Binomial Distribution where outcome 6 is success and outcomes 1,2,3,4 or is failure. Furthermore, out of 30 rolls the expected success of returning number 6 would be 5 times, and the die is decided to be fair. According to binomial distribution, the standard deviation=sqrt(30\* 1/6\*5/6)=2.04.

Follow the R simulation results:

a<-sample(1:6,size=30,replace = T)  
print(a)

## [1] 1 5 5 1 3 6 1 1 4 6 3 4 1 4 3 5 5 2 2 6 4 5 2 3 6 5 6 4 4 3

table(a)

## a  
## 1 2 3 4 5 6   
## 5 3 5 6 6 5

As observed, we did not obtain the outcome 6, for five times as predicted in theory. Next, we identified the hypothesis for testing the results. (1)we can write the null and alternative hypotheses as follows: Ho: p1=p2=p3=p4=p5=p6=1/6 Ha: there is some difference among the probabilities (2) The sample data are randomly selected, and the np>5 and n(1-p)>5, the condition for a binomial distribution are met, then we can use chi-square statistic to test. We set the level of significance given in the problem is alpha=0.05. (3) We use chi-square to test:

chisq.test(table(a),p=rep(1/6,6))

##   
## Chi-squared test for given probabilities  
##   
## data: table(a)  
## X-squared = 1.2, df = 5, p-value = 0.9449

1. Conclusion P-value=0.6692>0.05,so the conclusion is to fail to reject the null hypothesis.In other words, there is not sufficient evidence at the 0.05 level of significance to support there is some difference among the probabilities, concluding that the die is fair.

Following we observed that the probability of getting 5 or less 6 from 30 roll die is 0.616447. We also calculated the probability of the time number 6 comes up equal to 6,7,8. The conclusion is that when the times of outcome 6 increases by one, the probability will decrease a lot.

x<-pbinom(5,30,1/6)  
print(x)

## [1] 0.616447

x<-dbinom(6,30,1/6)  
print(x)

## [1] 0.1600901

x<-dbinom(7,30,1/6)  
print(x)

## [1] 0.1097761

x<-dbinom(8,30,1/6)  
print(x)

## [1] 0.06312124

PP3. We rolled the die 100 times on R, and the frequency distribution was uneven, with a large gap. It seems that the probability of each number of the die is not equal, but if the number of repetition continues to increase to 1000 times, we can find from the graph that the frequency of each number of the die gradually tightens to a straight line. We concluded that the frequency distribution obeys the uniform distribution.

Roll 100 times, the average of the result is 3.14.

a<-sample(1:6,size=100,replace = T)  
mean(a)

## [1] 3.61

According to PP2, we know that the probability of number 6 outcome follows the standard normal distribution. (1) we can write the null and alternative hypotheses as follows: Ho: p1=p2=p3=p4=p5=p6 Ha: there is some difference among the probabilities (2) The sample data are randomly selected, and the np>5 and n(1-p)>5,the conditions for the binomial distribution are met, then, we can use chi-square statistic to test. We set the level of significance given in the problem as alpha=0.05. (3) We use chi-square to test.

a<-sample(1:6,size=30,replace = T)  
print(a)

## [1] 3 2 4 2 1 1 2 2 1 4 2 2 5 3 5 6 5 5 6 1 2 6 2 1 3 4 3 2 2 1

table(a)

## a  
## 1 2 3 4 5 6   
## 6 10 4 3 4 3

library(ggplot2)  
qplot(a,binwidth=1)



chisq.test(table(a),p=rep(1/6,6))

##   
## Chi-squared test for given probabilities  
##   
## data: table(a)  
## X-squared = 7.2, df = 5, p-value = 0.2062

a<-sample(1:6,size=100,replace = T)  
print(a)

## [1] 6 4 4 5 1 2 1 4 6 1 1 2 6 2 5 5 5 4 5 5 2 5 3 3 1 4 4 5 2 2 2 4 6 2 6 6 5  
## [38] 4 1 2 1 4 5 6 5 2 5 3 1 3 5 5 1 5 4 3 3 3 3 4 4 3 5 3 4 4 3 4 1 3 4 3 6 6  
## [75] 2 5 6 1 3 3 3 6 6 2 5 4 4 4 1 3 4 5 6 5 5 6 2 3 4 3

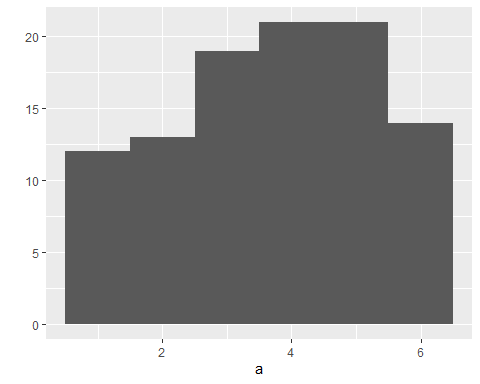
table(a)

## a  
## 1 2 3 4 5 6   
## 12 13 19 21 21 14

chisq.test(table(a),p=rep(1/6,6))

##   
## Chi-squared test for given probabilities  
##   
## data: table(a)  
## X-squared = 5.12, df = 5, p-value = 0.4014

qplot(a,binwidth=1)



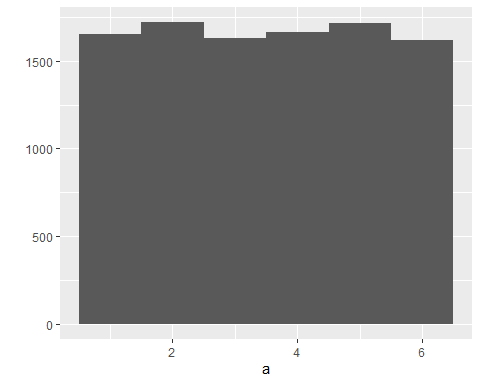
library(ggplot2)  
a<-sample(1:6,size=10000,replace = T)  
table(a)

## a  
## 1 2 3 4 5 6   
## 1653 1721 1629 1665 1712 1620

chisq.test(table(a),p=rep(1/6,6))

##   
## Chi-squared test for given probabilities  
##   
## data: table(a)  
## X-squared = 5.276, df = 5, p-value = 0.3831

qplot(a,binwidth=1)



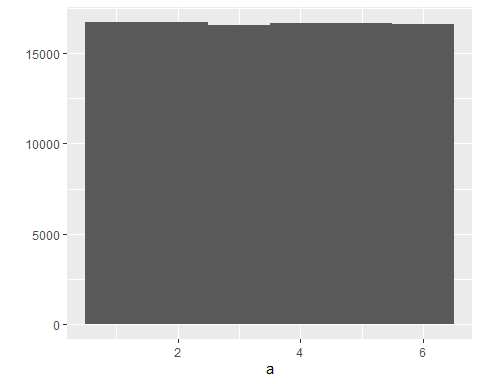
library(ggplot2)  
a<-sample(1:6,size=100000,replace = T)  
table(a)

## a  
## 1 2 3 4 5 6   
## 16733 16710 16577 16691 16678 16611

chisq.test(table(a),p=rep(1/6,6))

##   
## Chi-squared test for given probabilities  
##   
## data: table(a)  
## X-squared = 1.0882, df = 5, p-value = 0.9551

qplot(a,binwidth=1)



1. Conclusion all the P-value are greater than 0.05,so the conclusion is to fail to reject the null hypothesis.In other words, there is not sufficient evidence at the 0.05 level of significance to support there is some difference among the probabilities, so the dice is fair.

Final remarks from the experiment:

The results obtained on this project are not enough to draw a final conclusion about the fairness of a die. Only after using other resources for statistical testing (given the lack of R coding knowledge), we were able to try to minimize the chances of having type I and II errors, and establish a degree of confidence in the results obtained in here.

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

summary(cars)

## speed dist   
## Min. : 4.0 Min. : 2.00   
## 1st Qu.:12.0 1st Qu.: 26.00   
## Median :15.0 Median : 36.00   
## Mean :15.4 Mean : 42.98   
## 3rd Qu.:19.0 3rd Qu.: 56.00   
## Max. :25.0 Max. :120.00

## Including Plots

You can also embed plots, for example:



Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.