20121865_PearsonsChiSquare

Louise Branzuela

2022-06-01

Question 1

Null Hypothesis

A sensible null hypothesis would be that there is no systematic bias towards any of the colours, therefore the null hypothesis would be P(red) = P(orange) = P(yellow) = P(green) = 1/4.

Observed Values

```
o <- c(red = 10, orange = 16, yellow = 24, green = 12)
```

Expected Values

```
e <- c(red = 15.5, orange = 15.5, yellow = 15.5, green = 15.5)
```

However, since half a person is not possible, two of the colours would have 16 and and two would be 15. However, under the null hypothesis, these would be the expected values as its 62 divided by 4.

Pearson's chi square statistic sum

```
sum((o-e)^2/e)
```

```
## [1] 7.419355
```

The null distribution would be 3, as the degree of freedom is the number of cells minus one because the total is known so 4 - 1.

P-value and "more extreme" definition

P-value is the probability, if the null is true, of obtaining the observation or an observation more extreme. "More extreme" in this case would means that the p-value is larger than the observed sample mean.

P-value

```r
pchisq(7.419355, df=3, lower.tail=FALSE)
```

```
## [1] 0.05966718
```

Since the p-value is just short of the 5% critical value. We fail to reject the null hypothesis, which means that there is no systematic bias towards any colour.

Question 2

Justification for Hypothesis

The hypothesis would be a reasonable supposition as there is a large number of days in a year, each one of which have a small probability of a caller calling in.

Verifying Observations

```r
o <- c(c0=8, c1=12, c2=36, c3=54, c4=67,c5=66, c6=41, c7=37, c8=23, c9=10, c10=11)
sum(o)
```

```
## [1] 365
```

Number of calls in the year

```r
(8*0 + 12*1 + 36*2 + 54*3 + 67*4 + 66*5 + 41*6 + 37*7 + 23*8 + 10*9 + 11*10)
```

```
## [1] 1733
```

Average number of calls per day

```r
1733/365
```

```
## [1] 4.747945
```

Probability of having 0,1,2,...>=10 calls on any day

```
probs <- c(dpois(0:9, lambda=4.747945), ppois(9,lambda=4.747945,lower.tail=FALSE))
probs
```

```
##  [1] 0.008669493 0.041162275 0.097718108 0.154653401 0.183571460 0.174317439
##  [7] 0.137941602 0.093562734 0.055528840 0.029294209 0.023580440
```

```
sum(probs)
```

```
## [1] 1
```

Expected number of calls within the year.

```
e <- 365*probs
e
```

```
##  [1]  3.164365 15.024230 35.667109 56.448491 67.003583 63.625865 50.348685
##  [8] 34.150398 20.268026 10.692386  8.606861
```

Thus we expect to see about 3 days with 0 calls, 15 days with 1 call, etc.

Badness-of-fit

```
sum((o-e)^2/e)
```

```
## [1] 11.24837
```

P-value

```
B <- sum((o-e)^2/e)
pchisq(B,df=9,lower.tail=FALSE)
```

```
## [1] 0.2590679
```

Question 3

Expected Values

The expected number of balls would be a third of the total amount of draws as each ball has the same probability of being picked which is 1/3

```
12*(1/3)
```

```
## [1] 4
```

Calculating chi-square statistic of observation

```
o <- c(r=5, g=3, b=4)
e <- c(r=4, g=4, b=4)
B <- sum((o-e)^2/e)
B
```

```
## [1] 0.5
```

```
pchisq(B,df=2,lower.tail=FALSE)
```

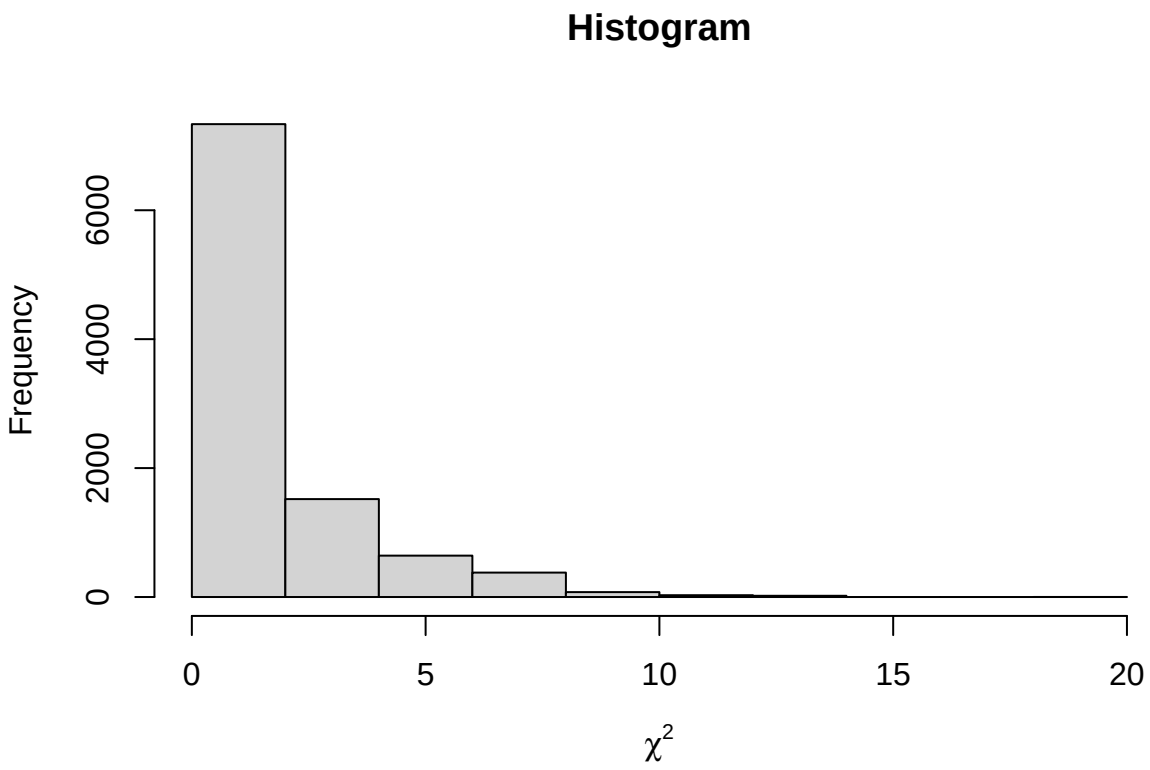
```
## [1] 0.7788008
```

There are 2 degrees of freedom as there are 3 cells minus 1 which equals to 2.

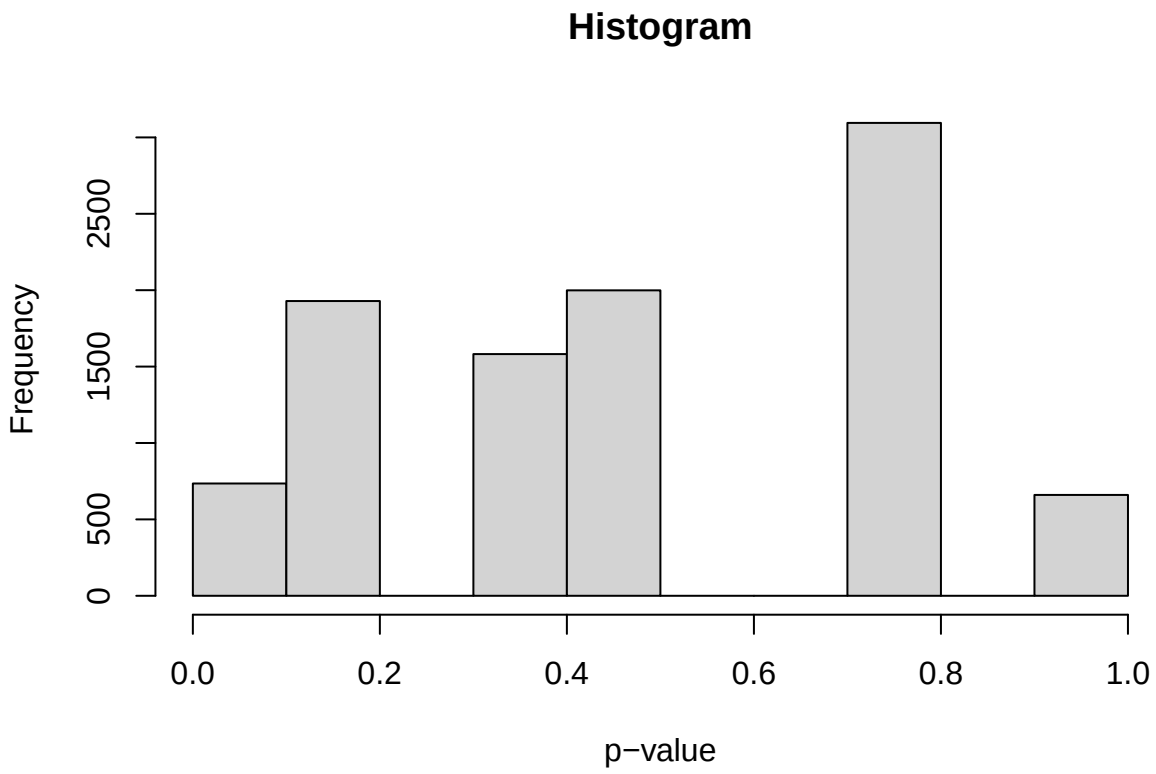
Generating sequences of random draws and histograms for their chi-square statistic and p-value

```
B_seq=1:1e4
p_value=1:1e4
for(i in 1:1e4) {
  S=sample(c("Red","Green","Blue"),size=12,replace =TRUE)
  Red=sum(S=="Red")
  Green=sum(S=="Green")
  Blue=sum(S=="Blue")
  B_seq[i] <- (Red-4)^2/4+(Green-4)^2/4+(Blue-4)^2/4
  p_value[i] <- pchisq(B_seq[i],df=2,lower.tail=FALSE)
}
```

```
hist(B_seq, breaks=10, xlab=expression(chi^2),main="Histogram")
```



```
hist(p_value, breaks=10, xlab="p-value",main="Histogram")
```



Question 4

Example

For the past year, I've logged the amount of League of Legends games that I've played in each day. I wish to test the hypothesis that the number of games that I play on any given day is a Poisson distribution.

Total number of games I've played in a year

```
(36*0 + 42*1 + 52*2 + 89*3 + 64*4 + 59*5 + 17*6 + 7*6)
```

```
## [1] 1108
```

Average number of games I play in any day

```
1108/365
```

```
## [1] 3.035616
```

Probability of playing 0,1,2,. . . >=6 games on any day

```
probs <- c(dpois(0:6, lambda=3.035616), ppois(6,lambda=3.035616,lower.tail=FALSE))
probs
```

```
## [1] 0.04804506 0.14584635 0.22136675 0.22399482 0.16999056 0.10320521 0.05221523
## [8] 0.03533601
```

```
sum(probs)
```

```
## [1] 1
```

Observed Values

```
o <- c(g0=36, g1=42, g2=52, g3=89, g4=64,g5=59, g6=17, g7=6)
o
```

```
## g0 g1 g2 g3 g4 g5 g6 g7
## 36 42 52 89 64 59 17  6
```

Expected Values

```
e <- 365*probs
e
```

```
## [1] 17.53645 53.23392 80.79886 81.75811 62.04656 37.66990 19.05856 12.89764
```

Badness-of-fit

```
sum((o-e)^2/e)
```

```
## [1] 48.7671
```

The degrees of freedom would be 8 - 1 - 1 = 6. This is because there are 8 categories, minus one because the total is known, and minus another one because the expectation values uses an estimation from lambda.

P-value

```
B <- sum((o-e)^2/e)
pchisq(B,df=6,lower.tail=FALSE)
```

```
## [1] 8.300541e-09
```

Since the p-value is smaller than the critical probability of 0.05, this means that we reject the null hypothesis and that the observation is not a Poisson distribution.