140.615.HW.Number.Jin.Vincent

Vincent Jin

2023-03-30

Homework 06

Vincent Jin

1.

Consider the counts of 5 outcome classes in the table below.

$$\begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 17 & 19 & 22 & 25 & 17 \end{bmatrix}$$

Do the outcomes 1 – 5 look equally likely? Use a χ^2 and a likelihood ratio test to answer that question.

Answer

For our hypothesis that outcomes 1-5 look equal likely, the p for each outcome will be 1 / 5 = 0.2.

```
x <- c(17,19,22,25,17)
p <- c(0.2, 0.2, 0.2, 0.2, 0.2)
cat("The results from chi-square test:\n")
```

```
## The results from chi-square test:
```

```
chisq.test(x, p = p)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  x
## X-squared = 2.4, df = 4, p-value = 0.6626
```

```
e <- sum(x) * p
cat("\nThe results from likelihood ratio test:\n")
```

```
##
## The results from likelihood ratio test:
```

```
lrt <- 2*sum(x*log(x/e))
pchisq(lrt,length(x)-1,lower.tail=FALSE)
```

```
## [1] 0.67161
```

The Chi-square suggested a p-value of 0.6626 under 4 degree of freedom and the likelihood ratio test also suggested a p-valuie of 0.6716 under 4 degree of freedom. In both test we failed to reject the null hypothesis of outcome 1-5 are equal likely, therefore, the outcomes 1-5 did look equal likely.

2.

Researchers studied a mutant type of flax seed. The amount of palmitic acid in the flax seed was an important factor in the search; a related factor was whether the seed was brown or variegated. The seeds were classified into six categories, as shown below. According to a hypothesized genetic model, the six combinations should occur in a 3:6:3:1:2:1 ratio. Use the χ^2 test to find out whether or not the data are consistent with this Mendelian model. Calculate the expected numbers for the six categories under the null hypothesis, derive the test statistic, and calculate the p-value. Confirm your findings using the built-in function in R.

$$\begin{bmatrix} Color & Acid\ Level & Observed \\ Brown & Low & 15 \\ Brown & Medium & 26 \\ Brown & High & 15 \\ Varieagated & Low & 0 \\ Varieagated & Medium & 8 \\ Varieagated & High & 8 \end{bmatrix}$$

Answer

```
o <- c(15, 26, 15, 0, 8, 8)
p <- c(3, 6, 3, 1, 2, 1)
p <- p / sum(p)
e <- sum(o) * p
xsq <- sum((o-e)^2/e)
pchisq(xsq,length(o) - 1,lower.tail=FALSE)
```

```
## [1] 0.1733389
```

```
chisq.test(o, p = p)
```

```
## Warning in chisq.test(o, p = p): Chi-squared approximation may be incorrect

##
##  Chi-squared test for given probabilities
##
## data:  o
## X-squared = 7.7037, df = 5, p-value = 0.1733
```

The chi-square test suggested a p-value of 0.1733 under 5 degree of freedom, so that we failed to reject the null hypothesis of data consistent with Mendelian model. This is also being confirmed by the built-in function in R.

3.

A researcher is interested in the co-infection of subjects with Hepatitis C and HIV. In 150 study subjects randomly chosen from a high risk population, he observed 10 subjects Hepatitis C and HIV, 14 with HIV only, 25 with Hepatitis C only, and 101 subjects infected with neither.

(a)

Is there evidence that the infections are not independent of each other? Derive the test statistic analytically, and calculate the p-value. Confirm your answer using a built in function.

Answer

Based on the information given, we can re-write the observed data as a table:

$$\begin{bmatrix} & & HIV & \\ & & Yes & No \\ Hep\ C & Yes & 10 & 25 \\ & No & 14 & 101 \end{bmatrix}$$

For expected value, we can calculate as column rows times the proportion of row totals. For example, the expected value for HIV and Hepatitis C co-infection is: (10 + 14) * (25 + 10) / 150 = 5.6

Calcualte for each cell we can write the table for expected values as:

$$\begin{bmatrix} & & HIV & \\ & & Yes & No \\ Hep\ C & Yes & 5.6 & 29.4 \\ & No & 18.4 & 96.6 \end{bmatrix}$$

```r
o <- c(10, 25, 14, 101)
e <- c(5.6, 29.4, 18.4, 96.6)
xsq <- sum((o-e)^2/e)
pchisq(xsq, (2 - 1) * (2 - 1), lower.tail=FALSE)
```

```
## [1] 0.02050673
```

Using built-in function to double check:

```r
o <- rbind(c(10, 25),c(14, 101))
chisq.test(o, correct = FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  o
## X-squared = 5.3682, df = 1, p-value = 0.02051
```

The chi-square test suggested a p-value of 0.0202 under 1 degree of freedom, so that we reject the null hypothesis of the two infections are independent and conclude that the two infections are not independent.

(b)

Is there evidence against the assumption that the prevalence of Hepatitis C and HIV are the same in this population? Derive the test statistic analytically, and calculate the p-value. Confirm your answer using a built in function.

Answer

To test the prevalence of Hepatitis C and HIV are the same, we are trying to assess if $P_{Hep\ C+} = P_{+HIV}$. In other words, we are testing if $P_{Hep\ C+HIV-} = P_{Hep\ C-HIV+}$. Therefore, we can conduct the chi-square statistic as:

```
xsq <- (25 - 14) ^ 2 / (25 + 14)
pchisq(xsq, 1, lower.tail=FALSE)
```

```
## [1] 0.07816909
```

Use built-in function to double check:

```
mcnemar.test(o, correct = FALSE)
```

```
##
##  McNemar's Chi-squared test
##
## data:  o
## McNemar's chi-squared = 3.1026, df = 1, p-value = 0.07817
```

The McNemar's test suggested a p-value of 0.0782 under 1 degree of freedom, so that we failed to reject the null hypothesis and conclude that the prevalence of two infections are the same.

4.

Among 1,000 subjects in a study, we see 649 AA, 300 AB, and 51 BB at one locus, and 640 AA, 360 AB and 0 BB at another locus. Use both the χ^2 and the likelihood ratio test to find out whether or not the loci are in Hardy-Weinberg equilibrium.

Answer

```
loci1 <- c(649, 300, 51)
loci2 <- c(640, 360, 0)
```

To determine if the loci are in Hardy-Weinberg equilibrium, we need to do tests separately for each loci:

```
cat("for loci1:\n")
```

```
## for loci1:
```

```
fhat <- (loci1[1] + loci1[2] / 2) / sum(loci1)
p1 <- c(fhat^2,2*fhat*(1-fhat),(1-fhat)^2)
e1 <- sum(loci1) * p1
lrt1 <- 2 * sum(loci1 * log(loci1 / e1))
cat("The result from LRT for loci 1 is:\n")
```

```
## The result from LRT for loci 1 is:
```

```r
pchisq(lrt1, 1, lower.tail=FALSE)
```

```
## [1] 0.04112428
```

```r
cat("\nThe result from chi-square test for loci 1 is:\n")
```

```
##
## The result from chi-square test for loci 1 is:
```

```r
xsq <- sum((loci1 - e1) ^ 2 / e1)
pchisq(xsq, 1, lower.tail=FALSE)
```

```
## [1] 0.03688831
```

For loci1, the LRT suggested a p-value of 0.0411 under 1 degree of freedom so that we reject the null hypothesis of loci1 are in Hardy-Weinberg equilibrium. The chi-square test suggested a p-value of 0.0369 under 1 degree of freedom so that we also reject the null hypothesis of loci1 are in Hardy-Weinberg equilibrium.

```r
cat("for loci2:\n")
```

```
## for loci2:
```

```r
fhat <- (loci2[1] + loci2[2] / 2) / sum(loci2)
p2 <- c(fhat^2,2*fhat*(1-fhat),(1-fhat)^2)
e2 <- sum(loci2) * p2
lrt2 <- 2 * sum(loci2 * log(loci2 / e2), na.rm = TRUE)
cat("The result from LRT for loci 2 is:\n")
```

```
## The result from LRT for loci 2 is:
```

```r
pchisq(lrt2, 1, lower.tail=FALSE)
```

```
## [1] 4.421128e-19
```

```r
cat("\nThe result from chi-square test for loci 2 is:\n")
```

```
##
## The result from chi-square test for loci 2 is:
```

```r
xsq <- sum((loci2 - e2) ^ 2/e2)
pchisq(xsq, 1, lower.tail=FALSE)
```

```
## [1] 3.877245e-12
```

For loci2, the LRT suggested a p-value less than 0.01 under 1 degree of freedom so that we reject the null hypothesis of loci2 are in Hardy-Weinberg equilibrium. The chi-square test suggested a p-value less than 0.01 under 1 degree of freedom so that we also reject the null hypothesis of loci2 are in Hardy-Weinberg equilibrium.

For both of the locis, we rejected the null hypothesis and conclude that the locis are not in Hardy-Weinberg equilibrium.