02 Quizz 1; Chi Dist

Contents

Quizz 1

Q1; Comparison of Observation to Base Ratio

A set of 104 university students were surveyed, asking for their favourite pizza type, with the following results.

Hawaiian	Barbecue	Supreme
32	14	58

The Australian population has the following pizza preferences.

Hawaian	Barbecue	Supreme
0.5	0.3	0.2

Calculate the chi squared distance between the student sample and the Australian population.

Solution

First enter the values, really carefully:

```
pizza <- c(32, 14, 58)
aus <- c(0.5, .3, 0.2)
```

This can either be calculated manually:

```
e <- sum(pizza)*aus
o <- pizza

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

```
## [1] 83.70513
```

Or the built in function can be used:

```
chisq.test(pizza, p = aus, rescale.p = TRUE)
```

```
##
##  Chi-squared test for given probabilities
##
## data: pizza
## X-squared = 83.705, df = 2, p-value < 2.2e-16
```

Q2; Comparing Populations (Assuming they're identically distributed)

A survery was performed on 100 people, asking for their city of birth and their eye colour. The results are tabulated below.

Brown	Hazel	Blue	
Sydney	10.00	30.00	20.00
Melbourne	20.00	10.00	10.00

If we assume that city of birth has no effect on eye colour, calculate the expected number of people that are born in Melbourne and have brown eyes.

Solution

So basically here:

- H0
  - Populations are distributed with equivalent Proportions
- Ha
  - Populations have different Proportions.

First Enter the Data

```
# Create Vectors
sydney    <- c("Brown" = 10,"Hazel" = 30,"Blue" = 20)
melbourne <- c("Brown" = 20,"Hazel" = 10,"Blue" = 10)
```

Create a Matrix

```
eye_Mat <- rbind(sydney, melbourne)
```

Determine the expected distributions

Outer Product Method

```
obs_count <- rowSums(eye_Mat)
feature_proportions <- colSums(eye_Mat)/sum(eye_Mat)

outer(rowSums(eye_Mat), colSums(eye_Mat)/sum(eye_Mat))
```

```
##           Brown Hazel Blue
## sydney       18    24   18
## melbourne    12    16   12
```

Matrix Method

```
obs_count
```

```
##    sydney melbourne
##        60        40
```

```
feature_proportions
```

```
## Brown Hazel Blue
##   0.3   0.4  0.3
```

```
as.matrix(obs_count) %*% t(as.matrix(feature_proportions))
```

```
##           Brown Hazel Blue
## sydney       18    24   18
## melbourne    12    16   12
```

Row and Col Sum Method

Another way to remember it is that the expected value is:

$$e_{ij} = \frac{\text{rowTotal} * \text{colTotal}}{\text{grandTotal}}$$

Perform a Chi Test

```
(cht <- chisq.test(x = eye_Mat, simulate.p.value = TRUE, B = 10^4))
```

```
##
## 	Pearson's Chi-squared test with simulated p-value
## 	(based on 10000 replicates)
##
## data: eye_Mat
## X-squared = 13.194, df = NA, p-value = 0.0017
```

Hence the probability of there detecting a difference between the two populations, under the assumption that the populations are identical is 0.002

Q3; Test For Uniform Distribution

A Bank is open everyday from 10am to 3pm and closed for lunch between 12pm and 1am. For one particular day the tellers serve the following number of customers in each hour.

10-11	11-12	1-2	2-3
80.00	70.00	69.00	61.00

Management would like to know if this is consistent with a uniform distribution of customers across the day. What is the expected number of customers per hour, if a uniform distribution is correct?

Solution

A uniform distribution would be such that each value is equal, hence:

```
bank <- c(80, 70, 69, 61)
e <- mean(bank)
```

The probability of incorrectly asserting that the values do differ over time (under the assumption that all values are identical) is given by:

```
chisq.test(bank, rescale.p = TRUE)
```

```
##
##  Chi-squared test for given probabilities
##
## data: bank
## X-squared = 2.6, df = 3, p-value = 0.4575
```

This is equivalent to specifying equal proportions, that's the default assumption made by ***R***:

```
chisq.test(bank, p = c(0.25, 0.25, 0.25, 0.25), rescale.p = TRUE)
```

```
##
##  Chi-squared test for given probabilities
##
## data: bank
## X-squared = 2.6, df = 3, p-value = 0.4575
```

In this case there is not enough evidence to reject the assumption that the distriution varies with time.

Q4; Expected Frequencies

A large sample survey of Australian businesses showed the following percentage of computers running the following operating systems.

Windows	OS X	Linux
30.00	50.00	20.00

Compute the expected number of computers using Linux from a sample of size 32.

Solution

Just multiply the proportion through:

```
comp <- c(30, 50, 20)
comp_rate <- comp/sum(comp)

comp_rate[3]*32
```

```
## [1] 6.4
```

hence the expected number of computers running linux is 6.4.