Statistics 302 Homework 8

Desmond Fung

4/27/2019

Textbook Problems

7.13 Favorite Skittles Flavor Skittles are a popular fruity candy with five different flavors (colored green, orange, purple, red, and yellow). A sample of 66 people3 recorded their favorite flavor and the results are shown in Table 7.8. Perform a chi-square test, as indicated in the steps below, to see whether or not the flavors are equally popular.

(a) State the null and alternative hypotheses.

$$H_0 : p_g = p_o = p_p = p_r = p_y = 0.2 H_a : p_i \neq 0.2$$

(b) If every flavor of skittle were equally popular, how many people (in a sample of 66) would we expect to choose each?

Expected Count = n * p = 66*0.2 = 13.2

(c) How many degrees of freedom do we have?

df = k - 1 = 4

(d) Calculate the chi-square test statistic.

chi-square test statistic = 3.7 using statkey

(e) What is the conclusion about the popularity of the skittles flavors?

p-value os 0.4481 > 0.05, therefore, by the condition, if p-value > level of significance, then do not reject the null hypothesis, so there is no evidence to reject the null hypothesis.

7.17 Birth Date and Canadian Ice Hockey In his book Outliers: The Story of Success (2008), Malcolm Gladwell speculates that Canadian ice hockey players that are born early in the year have an advantage. This is because the birthdate cutoff for different levels of youth hockey leagues in Canada is January 1st, so youth hockey players who are born in January and February are slightly older than teammates born later in the year. Does this slight age advantage in the beginning lead to success later on?A2010 study6 examined the birthdate distribution of players in the Ontario Hockey League (OHL), a high-level and selective Canadian hockey league (ages 15-20), for the 2008-2009 season. The number of OHL players born during the 1st quarter (Jan-Mar), 2nd quarter (Apr-Jun), 3rd quarter (Jul-Sep), and 4th quarter (Oct-Dec) of the year is shown in Table 7.11. The overall percentage of live births in Canada (year 1989) are also provided for each quarter. Is this evidence that the birthdate distribution for OHL players differs significantly from the national proportions? State the null and alternative hypotheses, calculate the chi-square statistic, find the p-value, and state the conclusion in context.

$$H_0 : p_1 = 0.237; p_2 = 0.259, p_3 = 0.259, p_4 = 0.245$$

H_a: There is evidence that the birthdates distribution for OHL players differs significantly from the national proportions.

df = 3

Chi-square test statistic is 82.8

p-value = 0

Therefore by the condition, if p-value < level of significance, then reject the null hypothesis.So there is evedience that the birthdates distribution for OHL players differs significantly from the national proportions.

7.20 Can People Delay Death? A new study indicates that elderly people are able to postpone death for a short time to reach an important occasion. The researchers9 studied deaths from natural causes among 1200 elderly people of Chinese descent in California during six months before and after the Harbor Moon Festival. Thirty-three deaths occurred in the week before the Chinese festival, compared with an estimated 50.82 deaths expected in that period. In the week following the festival, 70 deaths occurred, compared with an estimated 52. "The numbers are so significant that it would be unlikely to occur by chance," said one of the researchers.

(a) Given the information in the problem, is the ??2 statistic likely to be relatively large or relatively small?

??2 is likely to be large

(b) Is the p-value likely to be relatively large or relatively small?

??2 is likely to be small

(c) In the week before the festival, which is higher: the observed count or the expected count? What does this tell us about the ability of elderly people to delay death?

From the given result, the expected count is greater than the observed count in the week before the festival because some aged person might be capable to delay death.

(d) What is the contribution to the ??2-statistic for the week before the festival?

Chi-square test statistic is 6.249

(e) In the week after the festival, which is higher: the observed count or the expected count? What does this tell us about the ability of elderly people to delay death?

From the given result, the expected count is greater than the observed count in the week before the festival because some aged person might be capable to delay death but not with very much delay.

(f) What is the contribution to the ??2-statistic for the week after the festival?

The Chi-square test statistic for the week before the festival is 6.231

(g) The researchers tell us that in a control group of elderly people in California who are not of Chinese descent, the same effect was not seen. Why did the researchers also include a control group?

The control group consists of the aged persons who attend the Harbor Moon Festival. The effect was found in this control group where the persons had a meaningful event, thus the control group shows a significant difference in the event.

7.40 Metal Tags on Penguins In Exercise 6.178 on page 403 we perform a test for the difference in the proportion of penguins who survive over a 10-year period, between penguins tagged with metal tags and those tagged with electronic tags. We are interested in testing whether the type of tag has an effect on penguin survival rate, this time using a chi-square test. In the study, 33 of the 167 metal-tagged penguins survived while 68 of the 189 electronic-tagged penguins survived.

(a) Create a two-way table from the information given.

```
trial <- matrix(c(33, 68, 101, 134, 121, 255,
    167, 189, 356), ncol = 3)
colnames(trial) <- c("metal", "electronic", "row total")
rownames(trial) <- c("survived", "died", "total")
trial.table <- as.table(trial)
trial.table
```

```
         metal electronic row total
survived    33        134       167
died        68        121       189
total      101        255       356
```

(b) State the null and alternative hypotheses.

H_o There is no difference between type of tag and survial rate of penguins H_a There is some difference between type of tag and survial rate of penguins

(c) Give a table with the expected counts for each of the four categories.

```
trial <- matrix(c(47.38, 53.62, 119.62, 135.38),
    ncol = 2)
colnames(trial) <- c("metal", "electronic")
rownames(trial) <- c("survived", "died")
trial.table <- as.table(trial)
trial.table
```

```
          metal electronic
survived  47.38     119.62
died      53.62     135.38
```

(d) Calculate the chi-square test statistic.

The Chi-square test statistic with one degree of freedom is 11.476

(e) Determine the p-value and state the conclusion.

p-value = 0.001

Therefore by the condition, if p-value $<$ level of significance, then reject the null hypothesis.

7.43 Favorite Skittles Flavor? Exercise 7.13 on page 472 discusses a sample of people choosing their favorite Skittles flavor by color (green, orange, purple, red, or yellow). A separate poll sampled 91 people, again asking them their favorite skittle flavor, but rather than by color they asked by the actual flavor (lime, orange, grape, strawberry, and lemon, respectively).16 Table 7.29 shows the results from both polls. Does the way people choose their favorite Skittles type, by color or flavor, appear to be related to which type is chosen?

(a) State the null and alternative hypotheses.

H_o Skittles popularity is independent on method of choosing color against flavor H_a Skittles popularity is dependent on method of choosing color against flavor

(b) Give a table with the expected counts for each of the 10 cells.

```
trial <- matrix(c(13.03, 10.51, 14.29, 19.76,
    8.41, 17.97, 14.49, 19.71, 27.24, 11.59),
    ncol = 5)
```

```
colnames(trial) <- c("Green", "Orange", "Purple",
    "Red", "Yellow")
rownames(trial) <- c("color", "flavor")
trial.table <- as.table(trial)
trial.table
```

```
       Green Orange Purple   Red Yellow
color  13.03  14.29   8.41 14.49  27.24
flavor 10.51  19.76  17.97 19.71  11.59
```

(c) Are the expected counts large enough for a chi-square test?

Yes since expected count > 5

(d) How many degrees of freedom do we have for this test?

degree of freedom is 4

(e) Calculate the chi-square test statistic.

The Chi-square test statistic with 4 degree of freedom is 9.069

(f) Determine the p-value. Do we find evidence that method of choice affects which is chosen?

p-value = 0.059

Therefore by the condition, if p-value $>$ level of significance, then we do not reject the null hypothesis.

7.52 Another Test for Cocaine Addiction Exercise 7.41 on page 485 describes an experiment on helping cocaine addicts break the cocaine addiction, in which cocaine addicts were randomized to take desipramine, lithium, or a placebo. The results (relapse or no relapse after six weeks) are summarized in Table 7.35.

(a) In Exercise 7.41, we calculate a ??2 statistic of 10.5 and use a ??2 distribution to calculate a p-value of 0.005 using these data, but we also could have used a randomization distribution. How would you use cards to generate a randomization sample? What would you write on the cards, how many cards would there be of each type, and what would you do with the cards?

Write relapse or no relapse on the cards. From the given table, it is observed that the number of relapse cards is 48 and the number of no relapse cards is 24. All the cards are shuffled, and then segregate them into three equal packs namely, despiramine, lithium, and placebo.

(b) If you generated 1000 randomization samples according to your procedure from part (a) and calculated the ??2 statistic for each, approximately how many of these statistics do you expect would be greater than or equal to the ??2 statistic of 10.5 found using the original sample?

Since p-value is 0.005, 5 out of 1000 randomization test statistic will be greater or equal to the observed statistic.

R Problem 1 The function chisq.test() can be used for chi-square tests. For goodness-of-fit tests,provide a single array of counts and an optional array of the same length of probabilities for each case. (Ruses uniform probabilities by default.) For example, in class for the Rock, Paper, Scissors data, we hadtotal counts of 88, 74, and 66. Here is an example.

```
## 1.1

x <- c(66, 39, 14)
chisq.test(x)
```

```
    Chi-squared test for given probabilities

data:  x
X-squared = 34.101, df = 2, p-value = 3.936e-08
```

```
## 1.2
p0 <- c(0.5, 0.3, 0.2)
chisq.test(x, p = p0)
```

```
    Chi-squared test for given probabilities

data:  x
X-squared = 5.0504, df = 2, p-value = 0.08004
```

```
expt <- sum(x) * p0
test_chi <- sum((x - expt)^2/expt)
p_value <- pchisq(test_chi, df = length(x) - 1,
    lower.tail = F)
p_value
```

```
[1] 0.0800415
```

```
## 1.3

chisq.test(x, simulate.p.value = TRUE, B = 10000)
```

```
    Chi-squared test for given probabilities with simulated p-value
    (based on 10000 replicates)

data:  x
X-squared = 34.101, df = NA, p-value = 9.999e-05
```

R problem 2: Here you will learn to usechisq.test()for tests of association where the counts are inmatrices. Compare the answer you get with this function with the answer from your hand calculation onExercise 7.43.

```
## 2.1

x <- matrix(c(18, 13, 9, 16, 15, 19, 13, 34, 11,
    9), nrow = 2, ncol = 5)
chisq.test(x)
```

```
    Pearson's Chi-squared test

data:  x
X-squared = 9.0691, df = 4, p-value = 0.0594
```

```
row_sum <- apply(x, 1, sum)
colum_sum <- apply(x, 2, sum)
expt <- row_sum %*% t(colum_sum)/sum(x)
test_chi <- sum((x - expt)^2/expt)
p_value <- pchisq(test_chi, df = (2 - 1) * (5 -
    1), lower.tail = F)
p_value
```

```
[1] 0.05939575
```

R problem 3: This problem tests if data comes from a binomial distribution without specifying the valueofp.

```
## 3.1

prop_boy <- (0 * 5844 + 1 * 13079 + 2 * 6545)/(2 *
    (5844 + 13079 + 6545))
prop_girl <- 1 - prop_boy


## 3.2

p_binom <- dbinom(seq(0, 2, by = 1), size = 2,
    prob = prop_boy)
obs <- c(5844, 13079, 6545)
expt <- sum(obs) * p_binom
test_chi <- sum((obs - expt)^2/expt)
test_chi
```

```
[1] 19.78413
```

```
p_value <- pchisq(test_chi, df = 1, lower.tail = F)
p_value
```

```
[1] 8.669935e-06
```