Homework 7: Sections 7.1 - 7.2
KEY

STA209-04: Applied Statistics

Assigned: 04/12/2019
Due: 04/19/2019

Total Possible Points: 37

From the Book:

Questions: 7.15, 7.21, 7.23, 7.41, 7.45, 7.53

7.15 [7 pts] Between 2008 and 2011, the age distribution of users of social networking sites such as Facebook changed dramatically. In 2008, 70% of users were 35 years of or younger. In 2011, the age distribution was much more spread out. Table 7.10 shows the age distribution of 975 users of social networking sites from a survey reported in June 2011.

Table 7.10

Age	18-22	22-35	36-49	50-65	65+
Frequency	156	312	253	195	59

a) [5 pts] Test an assumption that users are equally likely to be in each of the five age groups listed. Show all details of the test.

Let p_1 denote the population proportion of 18-22 year olds, p_2 23-35, and so on.

$$H_0 : p_1 = p_2 = p_3 = p_4 = p_5 = 0.20$$

$$H_A : p_i \neq 0.20 \text{for at least one } i$$

Age	18-22	22-35	36-49	50-65	65+
Observed	156	312	253	195	59
Expected	$975(0.20) = 195$	195	195	195	195

$$\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i} = 190.1026$$

Using the test statistic, we turn to a $\chi^2_{df=4}$ distribution to obtain a p-value of virtually 0. We reject the null hypothesis and conclude that users are not all equally likely to be in each of the five age groups listed.

b) [2 pts] Which age group contributes the largest amount to the sum for the χ^2 test statistic? For this age group, is the observed count smaller or larger than the expected count?

The 65+ group contributes the largest amount to the sum of the test statistic. The observed count is much smaller than the expected count for this group.

7.21 [6 pts] Most medical school graduates in the US enter their residency programs at teaching hospitals in July. A recent study suggests that a spike in deaths due to medication errors coincides with this influx of new practitioners. The study indicates that the number of deaths is significantly higher than expected in July.

a) What type of statistical analysis was probably done to arrive at this conclusion?

A chi^2 goodness of fit test was likely performed. The counts of deaths due to medication errors were likely collected each month, and the test was whether deaths were equally likely under each month.

b) Is the χ^2 statistic likely to be relatively large or relatively small?

The statistic should be relatively large given that the study indicates a spike in July.

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

The p-value should be relatively small given that the statistic is relatively large.

d) What does the relevant categorical variable record?

The month corresponding to a death due to medication error.

e) What cell contributes the most to the χ^2 statistics?

The cell corresponding to July.

f) In the cell referred to in part e), which is higher: the observed count or the expected count?

The observed count is higher.

7.23 [7 pts] Movies based on Ian Fleming's novels starring British secret agent James Bond have become one of the longest running film series to date. As of 2016, six different actors have portrayed the secret agent. Which actor is the best James Bond? A sample of responses to this question is shown in Table 7.16.

Table 7.16

Actor	Frequency
Sean Connery	98
George Lazenby	5
Roger Moore	23
Timothy Dalton	9
Pierce Brosnan	25
Daniel Craig	51

a) [3 pts] Does the sample provide evidence of a significant difference in popularity among the six actors, at a 5% significance level?

The appropriate test to perform here is a chi^2 goodness of fit test. Let p_1 refer to the proportion who prefer Sean Connery, p_2 for George Lazenby, and so on.

$$H_0 : p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = 0.167$$

$$H_A : p_i \neq 0.167 \text{for at least one } i$$

Actor	Observed	Expected
Sean Connery	98	$211(0.167) = 35.167$
George Lazenby	5	35.167
Roger Moore	23	35.167
Timothy Dalton	9	35.167
Pierce Brosnan	25	35.167
Daniel Craig	51	35.167

$\chi^2 = 171.89$ which yields a p-value of virtually 0 under a $\chi^2_{df=5}$. We reject the null hypothesis and conclude that there is a difference in popularity among the Bond actors.

b) [3 pts] Repeat the test from part a) if we ignore the results for George Lazenby, who only appeared in one Bond film. Do we find evidence of a significant difference in popularity among the remaining five actors?

Repeating the above test without George Lazenby, we obtain the following table:

Actor	Observed	Expected
Sean Connery	98	$206(0.20) = 41.2$
Roger Moore	23	41.2
Timothy Dalton	9	41.2
Pierce Brosnan	25	41.2
Daniel Craig	51	41.2

We compute $\chi^2 = 120.21$ which yields a p-value of virtually 0 under a $\chi^2_{df=4}$ distribution. Again, we reject the null hypothesis and conclude that there is a difference in popularity among the Bond actors (even after removing George Lazenby from consideration).

c) [1 pts] The message from Chapter 1 still holds true: Pay attention to where the data come from! These data come from a poll held on a James Bond fan site. Can we generalize the results of this poll to the movie-watching population?

No, the sample is likely biased and not representative of the movie-watching population. Not all movie-goers have seen all James Bond films, and for those that have, not all would be so passionate as to fill out a survey on a fan site.

7.41 [5 pts] In Exercise 6.148 on page 445 we perform a test for the difference in the proportion of penguins who survive over a ten 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, 10 out of the 50 metal-tagged penguins survived while 18 out of the 50 electronic-tagged penguins survived.

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

Using the provided information, the following table is constructed:

	Survived	Died
Metal Tag	10	40
Electronic Tag	18	32

b) State the null and alternative hypotheses.

The null hypothesis is that tag type and survival are not associated with one another (i.e. independent). The alternative hypothesis is that tag type and survival are associated.

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

The expected counts for each category are shown in the table below:

	Survived	**Died**
Metal Tag	$50(0.28) = 14$	$50(0.72) = 36$
Electronic Tag	14	36

d) Calculate the chi-square test statistic

We compute $\chi^2 = 3.17$.

e) Determine the p-value and state the conclusion using a 5% significance level.

Using a $\chi^2_{df=1}$ distribution, we obtain a p-value of 0.075. We fail to reject the null hypothesis of independence. While we failed to reject the null hypothesis, there is marginal evidence of an association, given the relatively small p-value and the (non-minute) disparity between the observed and expected counts.

7.45 [6 pts] 478 middle school (grades 4 to 6) students from three school districts in Michigan were asked whether good grades, athletic ability, or popularity was most important to them. The results are shown below, broken down by gender.

	Grades	Sports	Popular
Boy	117	60	50
Girl	130	30	91

a) Do these data provide evidence that grades, sports, and popularity are not all equally valued among middle school students in these school districts? State the null and alternative hypotheses, calculate a test statistic, find a p-value, and answer the question.

A goodness of fittest is appropriate here. $H_0 : p_g = p_s = p_p = 0.33$; $H_A : p_i \neq 0.33$for at least one p_i. After collapsing both categories of sex into a single row of counts, we can compute a table of expected counts in order to find the value of our test statistic, $\chi^2 = 80.52$. Using a $\chi^2_{df=2}$ distribution, we obtain a p-value of virtually 0. We reject the null hypothesis and conclude that there is a difference in value between grades, sports, and popularity among middle school students.

b) Do middle school boys and girls have different priorities regarding grades, sports, and popularity? State the null and alternative hypotheses, calculate a test statistic, find a p-value, and answer the question.

Here, a test for association is appropriate. The null hypothesis is that sex and priority among grades, sports, and popularity are not associated. The alternative hypothesis is that they are. Computing our test statistic, we obtain $\chi^2 = 21.46$. Using a $\chi^2_{df=2}$ distribution, we obtain a p-value of virtually 0. We reject the null hypothesis and conclude that sex and preference among grades, sports, and popularity are associated (i.e. they differ among middle school boys and girls).

7.53 [6 pts] The study on genetics and fast-twitch muscles includes a sample of elite sprinters,a sample of elite endurance athletes, and a control group of non-athletes. Is there an association between genetic allele classification (R or X) and group (sprinter, endurance, control)? Computer output is shown for this chi-square test. In each cell, the top number is the observed count, the middle number is the expected count, and the bottom number is the contribution to the chi-square statistic.

	R	X	Total
Control	244	192	436
	251.42	184.58	
	0.219	0.299	
Sprint	77	30	107
	61.70	45.30	
	3.792	5.166	
Endurance	104	90	196
	111.87	82.13	
	0.554	0.755	
Total	425	312	737

Chi-Sq = 10.785, DF = 2, P-Value = 0.005

a) How many endurance athletes were included in the study?

196.

b) What is the expected count for sprinters with the R allele? For this cell, what is the contribution to the chi-square statistic? Verify both values by computing them yourself.

61.70, and 3.792. Verification:$107\left(\frac{425}{737}\right) = 61.703$; $\frac{(77-61.70)^2}{61.70} = 3.794$

c) What are the degrees of freedom for the test? Verify this value by computing it yourself.

$DF = 2(1) = 2$.

d) What is the chi-square test statistic? What is the p-value? What is the conclusion of the test?

$\chi^2 = 10.785$, p-value is 0.005, and we reject the null hypothesis. There is an association between genetic allele classification and group (sprinter, endurance, control).

e) Which cell contributes the most to the chi-square statistic? For this cell, is the observed count greater than or less than the expected count?

The Sprint,X cell contributes most to the chi-square statistic. In this cell, the observed count is less than the expected count (i.e. 30 vs 45.30).

f) Which allele is most over-represented in sprinters? Which allele is most over-represented in endurance athletes?

The R allele is most over-represented in sprinters since the observed count is greater than the expected count. The X allele is over-represented in endurance athletes since the observed count is greater than the expected count.