Assignment 9 Solutions

Your name and student ID

Today's date

- Due date: November 19, 11:59pm (make sure to provide enough time for Gradescope submission to be uploaded if you include large image files).
- Late penalty: 50% late penalty if submitted within 24 hours of due date, no marks for assignments submitted thereafter.
- Submission process: Please submit a PDF of your assignment to Gradescope. You must tell Gradescope which questions are on which pages. If you can't see it properly on Gradescope, open the PDF in a PDF viewer at the same time so you can make the selections accurately, others points will be deducted since this makes grading much less efficient.

Helpful hint:

- Knit your file early and often to minimize knitting errors! If you copy and paste code, you are bound to get an error that is hard to diagnose. Hand-writing code is the way to smooth knitting! We recommend knitting your file each time after you write a few sentences/add a new code chunk, so you can detect the source of knitting errors more easily.

Question 1. [7 marks total] Parental leave is often compensated to some degree, but the amount of compensation varies greatly. You read a research article that stated that, "across people of all incomes, 47% of leave-takers received full pay during their leave, 16% received partial pay, and 37% received no pay."

After reading this, you wonder what the distribution of pay is for low income households. Suppose you conduct a survey of leave-takers within households earning less than $30,000 per year. You surveyed 225 people (selected in a random sample) and found that 51 received full pay, 33 received partial pay, and 141 received no pay.

1.a) [1 mark] You would like to investigate whether the distribution of pay for households earning < $30,000 is different from that of all income levels. Does this correspond to a chi-square test of independence or a chi-square test for goodness of fit?

Solution: This corresponds to a chi-square test for goodness of fit. The reason for this is because we only have one sample (from low income households) and are comparing their observed counts for each category to a provided distribution.

1.b) [2 marks] What are the expected counts of leave-takers among households with incomes < $30,000. State the null hypothesis under which these expected counts were computed.

H_0: The null hypothesis is that the leave distribution would equal that which you read in the research article (i.e., that the proportion receiving full pay equals 47%, the proportion recieving partial pay is 16%, and the proportion with no pay is 37%.)

Expected counts under the null:

- 0.47*225 = 105.75
- 0.16*225 = 36
- 0.37*225 = 83.25

1.c) [2 marks] Compute the chi-square statistic and comment on the cell that contributes the most to the statistic.

$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$

= (105.75 - 51)^2/105.75 + (36 - 33)^2/36 + (83.25 - 141)^2/83.25 = 28.34574 + 0.25 + 40.06081 = 68.65656

The largest contribution comes from the deviation in the people that recieve no pay to go on parental leave. We see a much higher number of no pay among low income households that that expected under the null hypothesis.

1.d) [2 marks] Compute the p-value for your test statistic and conclude whether you believe there is evidence against the null hypothesis in favor of the alternative hypothesis.

```
pchisq(q = 68.65656, df = 2, lower.tail = F)
```

```
## [1] 1.234291e-15
```

The probability of seeing this chi-square statistic is very tiny (<0.001) under the null hypothesis. Thus we conclude there is evidence in favor of the alternative hypothesis that the distribution of leave is different for low income households vs. that specified in the research article.

Question 2. [9 marks total] Human papillomavirus (HPV) is a very common STI that most sexually active persons will encounter during their lifetimes. While many people clear the virus, certain strands can lead to adverse health outcomes such as genital warts and cervical cancer.

Suppose that you selected a random sample from a population and collected these data on age and HPV status for the sample:

Age Group	HPV +	HPV -	Row total
14-19	160	492	652 (33.9%)
20-24	85	104	189 (9.8%)
25-29	48	126	174 (9.1%)
30-39	90	238	328 (17.1%)
40-49	82	242	324 (16.9%)
50-59	50	204	254 (13.2%)
Col total	515 (26.8%)	1406 (73.2%)	1921

2.a) [1 mark] Which variable is explanatory and which is response?

Explanatory: Age group

Response: HPV status

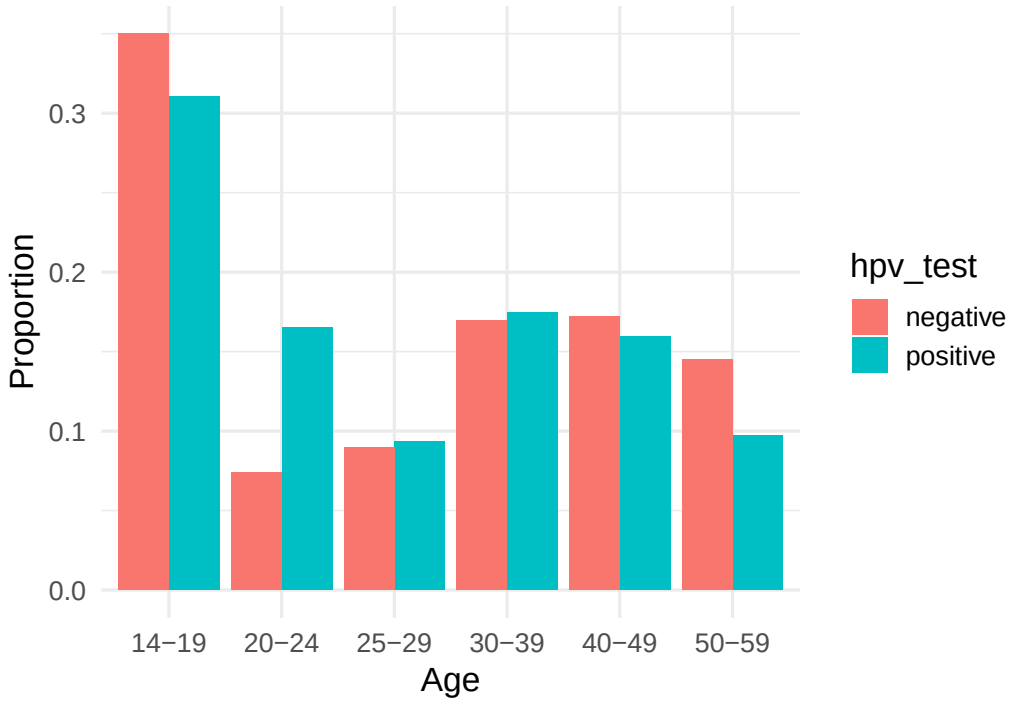
2.b) [2 marks] Formulate null and alternative hypotheses using these data to test whether there is a relationship between age group and HPV status. State these hypotheses using the language or notation of conditional distributions.

H_0 : The conditional distribution of age is the same for HPV + and HPV - individuals.

H_a : The conditional distribution of age is different for HPV + and HPV - individuals.

2.c) [1 mark] Run the code below to examine the conditional distribution of age by HPV status. Based on this plot, which age group will contribute the most to the chi-square statistic? (That is, can you tell based on this plot when the observed count will differ most from the expected count under the null hypothesis of no relationship between age group and HPV status?)

Cells corresponding to the 20-24 year-olds will contribute the most to the chi-square statistic because they exhibit the largest observed difference between HPV- and HPV+ individuals.



2.d) [2 marks] Fill out the table of expected counts under the null hypothesis of no association between age group and HPV status. You don't need to show your work, but make sure you can calculate the expected counts by hand, using a calculator.

Expected counts:

Age Group	HPV +	HPV -
14-19	652*515/1921 = 174.7944	652*1406/1921 = 477.2056
20-24	189*515/1921 = 50.66892	189*1406/1921 = 138.3311
25-29	174*515/1921 = 46.64758	174*1406/1921 = 127.3524
30-39	328*515/1921 = 87.93337	328*1406/1921 = 240.0666
40-49	324*515/1921 = 86.86101	324*1406/1921 = 237.139
50-59	254*515/1921 = 68.09474	254*1406/1921 = 185.9053

2.e) [3 marks] Calculate the test statistic and p-value and assess whether there is evidence against the null in favor of the alternative.

$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$

= [(174.7944 - 160)^2/174.7944] + [(477.2056 - 492)^2/477.2056] + [(50.66892 - 85)^2/50.66892] + [(138.3311 - 104)^2/138.3311] + [(46.64758 - 48)^2/46.64758] + [(127.3524 - 126)^2/127.3524] + [(87.93337 - 90)^2/87.93337] + [(240.0666 - 238)^2/240.0666] + [(86.86101 - 82)^2/86.86101] + [(237.139 - 242)^2/237.139] + [(68.09474 - 50)^2/68.09474] + [(185.9053 - 204)^2/185.9053] = 1.252181 + 0.4586582 + 23.26126 + 8.520314 + 0.03920975 + 0.01436161 + 0.04857041 + 0.01779021 + 0.2720371 + 0.09964334 + 4.808295 + 1.761209 = 40.55453

Degrees of Freedom = (6-1)*(2-1) = 5

```
pchisq(q = 40.55453, df = 5, lower.tail = F)
```

```
## [1] 1.154217e-07
```

The probability of seeing this chi-square statistic under the null hypothesis that the conditional distribution

of age is the same for HPV- and HPV+ is very small. Thus we conclude that there is evidence in favour of the alternative hypothesis that there is an association between age and HPV status.

Question 3. [7 marks] Fill in the blanks. Please let the asterisks so your answers are bolded in the rendered file.

The bootstrap method is used to compute **confidence intervals**, while the permutation test is used to conduct **hypothesis tests**.

Bootstrapping involves taking repeated simple random samples **with** replacement from the original sample of the **same** size as the original sample. For each bootstrap, the statistic of interest is calculated (say the median). These bootstrapped statistics are then plotted on a **histogram** and the **2.5th** and **97.5th** quantiles are computed to calculate a 95% confidence interval.