Homework 11

Your name and student ID

Today's date

```
BEGIN ASSIGNMENT
requirements: requirements.R
generate: true
```

Instructions

- Solutions will be released on Tuesday, April 27
- This semester, homework assignments are for practice only and will not be turned in for marks.

Helpful hints:

- Every function you need to use was taught during lecture! So you may need to revisit the lecture code to help you along by opening the relevant files on Datahub. Alternatively, you may wish to view the code in the condensed PDFs posted on the course website. Good luck!
- Knit your file early and often to minimize knitting errors! If you copy and paste code for the slides, you are bound to get an error that is hard to diagnose. Typing out the 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. This will save you and the GSIs from frustration! **You must knit correctly before submitting.**
- If your code runs off the page of the knitted PDF then you will LOSE POINTS! To avoid this, have a look at your knitted PDF and ensure all the code fits in the file (you can easily view it on Gradescope via the provided link after submitting). If it doesn't look right, go back to your .Rmd file and add spaces (new lines) using the return or enter key so that the code runs onto the next line.

Parental leave is often compensated to some degree, but the amount of compensation varies greatly. You read a research article that stated, "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 parental leave payment 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. [1 point] 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?

```
BEGIN QUESTION
name: p1
manual: true
```

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.

2. [1 point] What are the expected counts of leave-takers among households with incomes < $30,000? Assign each expected count to the appropriate variable. Make sure to remove the quotes. Round each number to 2 decimal places.

```
BEGIN QUESTION
name: p2
manual: false
points: 1
```

```r
# put your answer here
full_pay <- "REPLACE WITH NUMBER ROUNDED TO 2 DECIMALS"
partial_pay <- "REPLACE WITH NUMBER ROUNDED TO 2 DECIMALS"
no_pay <- "REPLACE WITH NUMBER ROUNDED TO 2 DECIMALS"

# BEGIN SOLUTION NO PROMPT
full_pay <- 105.75
partial_pay <- 36.00
no_pay <- 83.25
# END SOLUTION
```

```r
## Test ##
test_that("p2a", {
  expect_true(all.equal(full_pay, 105.75, tol = 0.01))
  print("Checking: full pay value")
})
```

```
## [1] "Checking: full pay value"
## Test passed
```

```r
## Test ##
test_that("p2b", {
  expect_true(all.equal(fpartial_pay, 105.75, tol = 0.01))
  print("Checking: partial pay value")
})
```

```
## -- Error (<text>:3:3): p2b ----------------------------------------------------
## Error: object 'fpartial_pay' not found
## Backtrace:
##  1. testthat::expect_true(all.equal(fpartial_pay, 105.75, tol = 0.01))
##  4. base::all.equal(fpartial_pay, 105.75, tol = 0.01)
```

```r
## Test ##
test_that("p2c", {
  expect_true(all.equal(full_pay, 83.25, tol = 0.01))
  print("Checking: no pay value")
})
```

```
## [1] "Checking: no pay value"
## -- Failure (<text>:3:3): p2c ---------------------------------------------------
## all.equal(full_pay, 83.25, tol = 0.01) is not TRUE
##
## `actual` is a character vector ('Mean relative difference: 0.212766')
## `expected` is a logical vector (TRUE)
```

3. [1 point] State the null hypothesis under which the above expected counts were computed.

```
BEGIN QUESTION
name: p3
manual: true
```

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%.)

4. [1 point] Compute the chi-square statistic. Round your answer to 2 decimal places.

```
BEGIN QUESTION
name: p4
manual: false
points: 1
```

```r
# put your answer here
chi_sq_answer <- "REPLACE WITH NUMBER ROUNDED TO 2 DECIMALS"

# BEGIN SOLUTION NO PROMPT
chi_sq_answer <- 68.66
# END SOLUTION
```

```r
## Test ##
test_that("p4", {
  expect_true(all.equal(chi_sq_answer, 68.66, tol = 0.01))
  print("Checking: value of chi sq to 2 decimals")
})
```

```
## [1] "Checking: value of chi sq to 2 decimals"
## Test passed
```

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

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

5. [1 point] Uncomment which cell (i.e. term in summation) contributes the most to the statistic.

```
BEGIN QUESTION
name: p5
manual: false
points: 1
```

```r
# UNCOMMENT THE CORRECT ANSWER

# largest_contribution <- "full pay"
# largest_contribution <- "partial pay"
# largest_contribution <- "no pay"

# BEGIN SOLUTION NO PROMPT
largest_contribution <- "no pay"
# END SOLUTION
```

```r
## Test ##
test_that("p5", {
  expect_true(largest_contribution == "no pay")
  print("Checking: selection")
})
```

```
## [1] "Checking: selection"
## Test passed
```

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 than that expected under the null hypothesis.

6. [1 point] Compute the p-value for your test statistic. Round your answer to 2 decimal places.

```
BEGIN QUESTION
name: p6
manual: false
points: 1
```

```r
# put your answer here
p_value <- "REPLACE WITH NUMBER ROUNDED TO 2 DECIMALS"

# BEGIN SOLUTION NO PROMPT
p_value <- round(pchisq(q = 68.65656, df = 2, lower.tail = F), 2)
p_value <- 0.00
# END SOLUTION
```

```r
## Test ##
test_that("p6", {
  expect_true(all.equal(p_value, 0, tol = 0.01))
  print("Checking: p value to 2 decimal places")
})
```

```
## [1] "Checking: p value to 2 decimal places"
## Test passed
```

7. [1 point] Conclude whether you believe there is evidence against the null hypothesis in favor of the alternative hypothesis. Answer this by uncommenting the appropriate conclusion.

```
BEGIN QUESTION
name: p7
manual: false
points: 1
```

```r
# UNCOMMENT THE CORRECT ANSWER

# conclusion <- "in favor of null"
# conclusion <- "against null"

# BEGIN SOLUTION NO PROMPT
conclusion <- "against null"
# END SOLUTION
```

```r
## Test ##
test_that("p7", {
  expect_true(conclusion == "against null")
  print("Checking: selection")
})
```

```
## [1] "Checking: selection"
## Test passed
```

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.

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

8. [1 point] Which variable is explanatory and which is response? Uncomment the appropriate answer.

```
BEGIN QUESTION
name: p8
manual: false
points: 1
```

```
# UNCOMMENT THE CORRECT ANSWER

# variable_type <- c("explanatory: age group", "response: HPV status")
# variable_type <- c("explanatory: HPV status", "response: age group")

# BEGIN SOLUTION NO PROMPT
variable_type <- c("explanatory: age group", "response: HPV status")
# END SOLUTION
```

```
## Test ##
test_that("p8a", {
  expect_true(variable_type[1] == "explanatory: age group")
  print("Checking: explanatory")
})
```

```
## [1] "Checking: explanatory"
## Test passed
```

```
## Test ##
test_that("p8b", {
  expect_true(variable_type[2] == "response: HPV status")
  print("Checking: response")
})
```

```
## [1] "Checking: response"
## Test passed
```

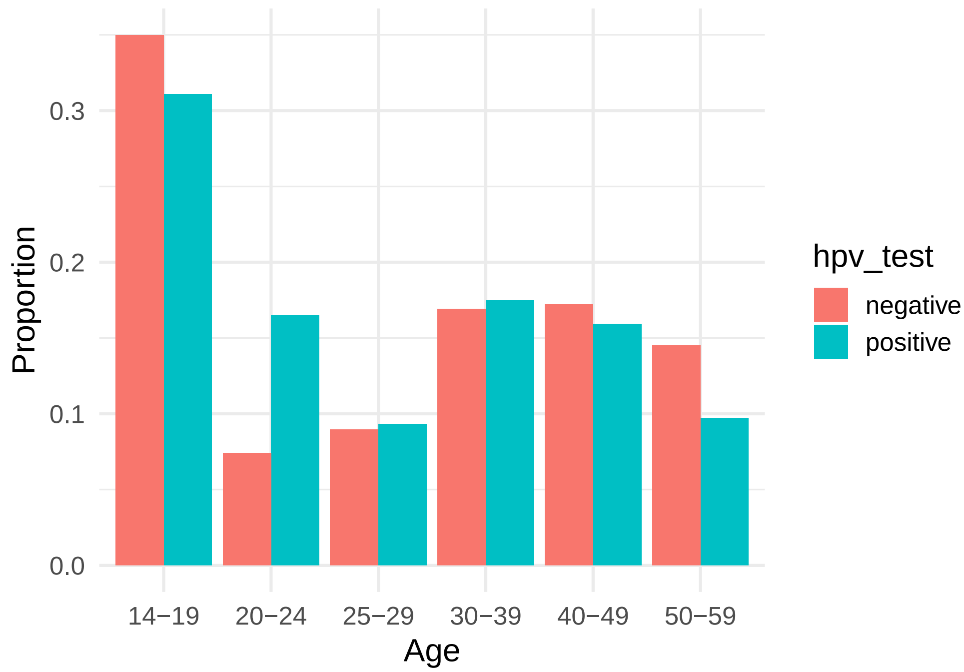
9. [2 points] Formulate null and alternative hypotheses using these data to test whether there is a relationship between age group and HPV status. State these hypotheses using the language or notation of conditional distributions.

```
BEGIN QUESTION
name: p9
manual: true
```

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.

10. [1 point] 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? Explain why. (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?)




```
BEGIN QUESTION
name: p10
manual: true
```

Cells corresponding to the 20-24 year-olds will likely contribute the most to the chi-square statistic because they exhibit the largest observed difference between HPV- and HPV+ individuals. (Additionally, though not requred for full marks, one might mention that the low overall proportion for 20-24 year olds means the denominator for the 20-24 y.o. Chi Square term will be relatively small)

11. [2 points] 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. Assign each appropriate cell/letter to the variable in the code. Round each number to 2 decimal places.

Expected counts:

Age Group	HPV +	HPV -
14-19	A	H
20-24	B	I
25-29	C	J
30-39	D	K
40-49	E	L
50-59	G	M

```
BEGIN QUESTION
name: p11
manual: false
points: 2
```

```r
# put your answer here
A <- "REPLACE WITH NUMBER ROUNDED TO 2 DECIMALS"
B <- "REPLACE WITH NUMBER ROUNDED TO 2 DECIMALS"
C <- "REPLACE WITH NUMBER ROUNDED TO 2 DECIMALS"
D <- "REPLACE WITH NUMBER ROUNDED TO 2 DECIMALS"
E <- "REPLACE WITH NUMBER ROUNDED TO 2 DECIMALS"
G <- "REPLACE WITH NUMBER ROUNDED TO 2 DECIMALS"
H <- "REPLACE WITH NUMBER ROUNDED TO 2 DECIMALS"
I <- "REPLACE WITH NUMBER ROUNDED TO 2 DECIMALS"
J <- "REPLACE WITH NUMBER ROUNDED TO 2 DECIMALS"
K <- "REPLACE WITH NUMBER ROUNDED TO 2 DECIMALS"
L <- "REPLACE WITH NUMBER ROUNDED TO 2 DECIMALS"
M <- "REPLACE WITH NUMBER ROUNDED TO 2 DECIMALS"

# BEGIN SOLUTION NO PROMPT
A <- 174.79
B <- 50.67
C <- 46.65
D <- 87.93
E <- 86.86
G <- 68.09
H <- 477.21
I <- 138.33
J <- 127.35
K <- 240.07
L <- 237.14
M <- 185.91
# END SOLUTION
```

```r
## Test ##
test_that("p11a", {
  expect_true(all.equal(A, 174.79, tol = 0.01))
```

```r
  print("Checking: A to 2 decimal places")
})
```

```
## [1] "Checking: A to 2 decimal places"
## Test passed
```

```r
## Test ##
test_that("p11b", {
  expect_true(all.equal(B, 50.67, tol = 0.01))
  print("Checking: B to 2 decimal places")
})
```

```
## [1] "Checking: B to 2 decimal places"
## Test passed
```

```r
## Test ##
test_that("p11c", {
  expect_true(all.equal(C, 46.65, tol = 0.01))
  print("Checking: C to 2 decimal places")
})
```

```
## [1] "Checking: C to 2 decimal places"
## Test passed
```

```r
## Test ##
test_that("p11d", {
  expect_true(all.equal(D, 87.93, tol = 0.01))
  print("Checking: D to 2 decimal places")
})
```

```
## [1] "Checking: D to 2 decimal places"
## Test passed
```

```r
## Test ##
test_that("p11e", {
  expect_true(all.equal(E, 86.86, tol = 0.01))
  print("Checking: E to 2 decimal places")
})
```

```
## [1] "Checking: E to 2 decimal places"
## Test passed
```

```r
## Test ##
test_that("p11f", {
  expect_true(all.equal(G, 68.09, tol = 0.01))
  print("Checking: G to 2 decimal places")
})
```

```
## [1] "Checking: G to 2 decimal places"
## Test passed
```

```r
## Test ##
test_that("p11g", {
  expect_true(all.equal(H, 477.21, tol = 0.01))
  print("Checking: H to 2 decimal places")
})
```

```
## [1] "Checking: H to 2 decimal places"
## Test passed
```

```r
## Test ##
test_that("p11h", {
  expect_true(all.equal(I, 138.33, tol = 0.01))
  print("Checking: I to 2 decimal places")
})
```

```
## [1] "Checking: I to 2 decimal places"
## Test passed
```

```r
## Test ##
test_that("p11i", {
  expect_true(all.equal(J, 127.35, tol = 0.01))
  print("Checking: J to 2 decimal places")
})
```

```
## [1] "Checking: J to 2 decimal places"
## Test passed
```

```r
## Test ##
test_that("p11j", {
  expect_true(all.equal(K, 240.07, tol = 0.01))
  print("Checking: K to 2 decimal places")
})
```

```
## [1] "Checking: K to 2 decimal places"
## Test passed
```

```r
## Test ##
test_that("p11k", {
  expect_true(all.equal(L, 237.14, tol = 0.01))
  print("Checking: L to 2 decimal places")
})
```

```
## [1] "Checking: L to 2 decimal places"
## Test passed
```

```r
## Test ##
test_that("p11l", {
  expect_true(all.equal(M, 185.91, tol = 0.01))
  print("Checking: M to 2 decimal places")
})
```

```
## [1] "Checking: M to 2 decimal places"
## Test passed
```

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

12. [1 point] Calculate the test statistic. Round your answer to 2 decimal places.

```
BEGIN QUESTION
name: p12
manual: false
points: 1
```

```r
# put your answer here
chi_sq_p12 <- "REPLACE WITH NUMBER ROUNDED TO 2 DECIMALS"

# BEGIN SOLUTION NO PROMPT
chi_sq_p12 <- 40.55
# END SOLUTION
```

```r
## Test ##
test_that("p12", {
  expect_true(all.equal(chi_sq_p12, 40.55, tol = 0.01))
  print("Checking: statistic to 2 decimal places")
})
```

```
## [1] "Checking: statistic to 2 decimal places"
## Test passed
```

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

= [(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 = 40.55353

13. [1 point] Calculate the p-value for your test statistic. Round your answer to 2 decimal places.

```
BEGIN QUESTION
name: p13
manual: false
points: 1
```

```r
p_value_p13 <- "REPLACE WITH NUMBER ROUNDED TO 2 DECIMALS"

# BEGIN SOLUTION NO PROMPT
df <- (6-1)*(2-1)  # 5
p_value_p13 <- round(pchisq(q = 40.55353, df = 5, lower.tail = F), 2)
p_value_p13 <- 0.00
# END SOLUTION
```

```r
## Test ##
test_that("p13a", {
  expect_true(p_value_p13 < 1 & p_value_p13 > -0.1)
  print("Checking: range of p-value")
})
```

```
## [1] "Checking: range of p-value"
## Test passed
```

```r
## Test ##
test_that("p13b", {
  expect_true(all.equal(p_value_p13, 0.00, tol = 0.0001))
  print("Checking: p-value to 2 decimal places")
})
```

```
## [1] "Checking: p-value to 2 decimal places"
## Test passed
```

14. [1 point] Assess whether there is evidence against the null in favor of the alternative. Answer this by uncommenting the appropriate conclusion.

```
BEGIN QUESTION
name: p14
manual: false
points: 1
```

```r
# UNCOMMENT THE CORRECT ANSWER
# conclusion_p14 <- "in favor of null"
# conclusion_p14 <- "against null"

# BEGIN SOLUTION NO PROMPT
conclusion_p14 <- "against null"
# END SOLUTION
```

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.