Assignment 09

Your name and student ID

Today's date

- Due date: Tuesday, November 10th.
- Remember: autograder is meant as sanity check ONLY. It will not tell you if you have the correct answer. It will tell you if you are in the ball park of the answer so *CHECK YOUR WORK*.
- Submission process: Follow the submission instructions on the final page. Make sure you do not remove any `\newpage` tags or rename this file, as this will break the submission.

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.*

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?

[TODO: YOUR ANSWER HERE]

BEGIN 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. # END SOLUTION

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.

```
# 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
full_pay <- 105.75
partial_pay <- 36.00
no_pay <- 83.25
# END SOLUTION

check_problem2()
```

```
## [1] "Checkpoint 1 Passed: Full pay is numeric"
## [1] "Checkpoint 2 Passed: Partial pay is numeric"
## [1] "Checkpoint 3 Passed: no pay is numeric"
##
## Problem 2
## Checkpoints Passed: 3
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

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

[TODO: YOUR ANSWER HERE]

BEGIN SOLUTION

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%.) # END SOLUTION

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

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

# BEGIN SOLUTION
chi_sq_answer <- 68.66
# END SOLUTION

check_problem4()
```

```
## [1] "Checkpoint 1 Passed: You answer is numeric!"
## [1] "Checkpoint 2 Passed: You answer is correct!"
##
## Problem 4
## Checkpoints Passed: 2
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

BEGIN SOLUTION

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

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

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

```
# UNCOMMENT THE CORRECT ANSWER

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

# BEGIN SOLUTION
largest_contribution <- "no pay"
# END SOLUTION

check_problem5()
```

```
## [1] "Checkpoint 1 Passed: You answer is correct"
##
## Problem 5
## Checkpoints Passed: 1
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

BEGIN SOLUTION

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. # END SOLUTION

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

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

# BEGIN SOLUTION
p_value <- round(pchisq(q = 68.65656, df = 2, lower.tail = F), 2)
p_value <- 0.00
# END SOLUTION

check_problem6()
```

```
## [1] "Checkpoint 1 Passed: Your answer is corrert range!"
## [1] "Checkpoint 2 Passed: Your answer is correct!"
##
## Problem 6
## Checkpoints Passed: 2
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

7. [1 point] Conclude whether you believe there is evidence against the null hypothesis in favor of the alternative hypothesis under the significance level of 0.001. Answer this by uncommenting the appropriate conclusion.

```
# UNCOMMENT THE CORRECT ANSWER

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

# BEGIN SOLUTION
conclusion <- "against null"
# END SOLUTION

check_problem7()
```

```
## [1] "Checkpoint 1 Passed: You answer is correct!"
##
## Problem 7
## Checkpoints Passed: 1
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

BEGIN SOLUTION

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. # END SOLUTION

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.

```
# UNCOMMENT THE CORRECT ANSWER

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

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

check_problem8()
```

```
## [1] "Checkpoint 1 Passed: Your answer is correct!"
##
## Problem 8
## Checkpoints Passed: 1
## Checkpoints Errored: 0
## 100% passed
## --------
## 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.

[TODO: YOUR ANSWER HERE]

BEGIN SOLUTION

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

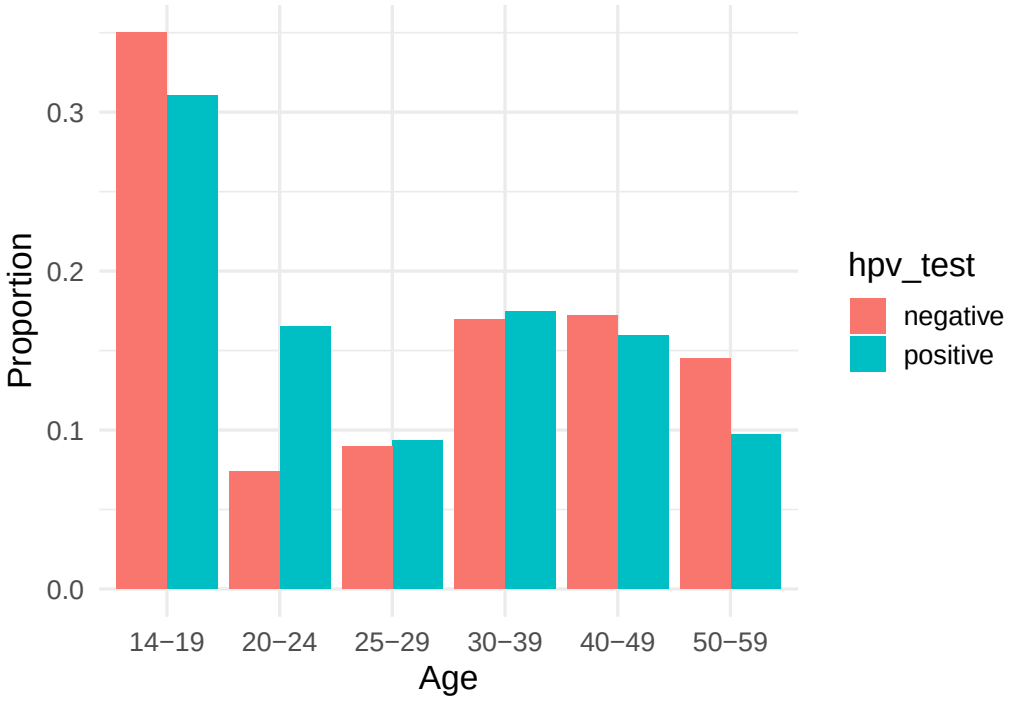
H_a : The conditional distribution of age is different for HPV + and HPV - individuals. # END SOLUTION

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?)

[TODO: YOUR ANSWER HERE]

BEGIN SOLUTION

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) # END SOLUTION



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

```
# 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
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

check_problem11()
```

```
## [1] "Checkpoint 1 Passed: A is correct"
## [1] "Checkpoint 2 Passed: B is correct"
## [1] "Checkpoint 3 Passed: C is correct"
## [1] "Checkpoint 4 Passed: D is correct"
## [1] "Checkpoint 5 Passed: E is correct"
## [1] "Checkpoint 6 Passed: G is correct"
## [1] "Checkpoint 7 Passed: H is correct"
```

```
## [1] "Checkpoint 8 Passed: I is correct"
## [1] "Checkpoint 9 Passed: J is correct"
## [1] "Checkpoint 10 Passed: K is correct"
## [1] "Checkpoint 11 Passed: L is correct"
## [1] "Checkpoint 12 Passed: M is correct"
##
## Problem 11
## Checkpoints Passed: 12
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

BEGIN SOLUTION

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

END SOLUTION

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

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

# BEGIN SOLUTION
chi_sq_p12 <- 40.55
# END SOLUTION

check_problem12()
```

```
## [1] "Checkpoint 1 Passed: You answer is numeric!"
## [1] "Checkpoint 2 Passed: Correct Answer!"
##
## Problem 12
## Checkpoints Passed: 2
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

BEGIN SOLUTION

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

= [(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

END SOLUTION

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

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

# BEGIN SOLUTION
p_value_p13 <- round(pchisq(q = 40.55353, df = 5, lower.tail = F), 2)
p_value_p13 <- 0.00
# END SOLUTION

check_problem13()
```

```
## [1] "Checkpoint 1 Passed: You answer is numeric!"
## [1] "Checkpoint 2 Passed: Correct Answer!"
##
## Problem 13
## Checkpoints Passed: 2
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

BEGIN SOLUTION

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

```
pchisq(q = 40.55353, df = 5, lower.tail = F)
```

```
## [1] 1.154754e-07
```

END SOLUTION

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

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

# BEGIN SOLUTION
conclusion_p14 <- "against null"
# END SOLUTION

check_problem14()
```

```
## [1] "Checkpoint 1 Passed: You answer is correct"
##
## Problem 14
## Checkpoints Passed: 1
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

BEGIN 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. # END SOLUTION

15. [3.5 marks] Fill in the blanks. Assign the corresponding object (`a`, `b`, `c`, `d`, `e`, `f`, `g`) with your answer.

The bootstrap method is used to compute _____**a**_____, while the permutation test is used to conduct _____**b**_____.

Bootstrapping involves taking repeated simple random samples _____**c**_____ replacement from the original sample of the _____**d**_____ 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 _____**e**_____ and the _____**f**_____ and _____**g**_____ quantiles are computed to calculate a 95% confidence interval.

```
a <- "YOUR ANSWER HERE"
b <- "YOUR ANSWER HERE"
c <- "YOUR ANSWER HERE"
d <- "YOUR ANSWER HERE"
e <- "YOUR ANSWER HERE"
f <- "YOUR ANSWER HERE"
g <- "YOUR ANSWER HERE"

# BEGIN SOLUTION
a <- "confidence intervals"
b <- "hypothesis tests"
c <- "with"
d <- "same"
e <- "histogram"
f <- "2.5th"
g <- "97.5th"
# END SOLUTION

check_problem15()
```

```
## [1] "Checkpoint 1 Passed: Blank a is correct!"
## [1] "Checkpoint 2 Passed: Blank b is correct!"
## [1] "Checkpoint 3 Passed: Blank c is correct!"
## [1] "Checkpoint 4 Passed: Blank d is correct!"
## [1] "Checkpoint 5 Passed: Blank e is correct!"
## [1] "Checkpoint 6 Passed: Blank f is correct!"
## [1] "Checkpoint 7 Passed: Blank g is correct!"
##
## Problem 15
## Checkpoints Passed: 7
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

Check your score

Click on the middle icon on the top right of this code chunk (with the downwards gray arrow and green bar) to run all your code in order. Then, run this chunk to check your score.

```
# Just run this chunk.
total_score()
```

```
##                      Test Points_Possible          Type
## Problem 1  NOT YET GRADED               1 free-response
## Problem 2          PASSED               1    autograded
```

```
## Problem 3  NOT YET GRADED              1 free-response
## Problem 4           PASSED              1    autograded
## Problem 5           PASSED              1    autograded
## Problem 6           PASSED              1    autograded
## Problem 7           PASSED              1    autograded
## Problem 8           PASSED              1    autograded
## Problem 9  NOT YET GRADED              2 free-response
## Problem 10 NOT YET GRADED              1 free-response
## Problem 11          PASSED              2    autograded
## Problem 12          PASSED              1    autograded
## Problem 13          PASSED              1    autograded
## Problem 14          PASSED              1    autograded
## Problem 15          PASSED            3.5    autograded
```

Submission

For assignments in this class, you'll be submitting using the **Terminal** tab in the pane below. In order for the submission to work properly, make sure that:

1. Any image files you add that are needed to knit the file are in the `src` folder and file paths are specified accordingly.
2. You **have not changed the file name** of the assignment.
3. The file is saved (the file name in the tab should be **black**, not red with an asterisk).
4. The file knits properly.

Once you have checked these items, you can proceed to submit your assignment.

1. Click on the **Terminal** tab in the pane below.
2. Copy-paste the following line of code into the terminal and press enter.

cd; cd ph142-sp20/hw/hw11; python3 turn_in.py

3. Follow the prompts to enter your Gradescope username and password. When entering your password, you won't see anything come up on the screen–don't worry! This is just for security purposes–just keep typing and hit enter.
4. If the submission is successful, you should see "Submission successful!" appear as output.
5. If the submission fails, try to diagnose the issue using the error messages–if you have problems, post on Piazza.

The late policy will be strictly enforced, **no matter the reason**, including submission issues, so be sure to submit early enough to have time to diagnose issues if problems arise.