# Problem Set 9

Your name and student ID

Today's date

**Instructions**

- Solutions will be released by Monday, November 14th.
- This semester, problem sets 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!
- To avoid code running off the page, have a look at your knitted PDF and ensure all the code fits in the file. 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.

## Section 1: Parental Leave

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. 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?**

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. What are the expected counts of leave-takers among households with incomes < $30,000? Assign p2 to a vector of the expected counts for full pay, partial pay, and no pay. Round each number to 2 decimal places.**

```
. = " # BEGIN PROMPT
p2 <- NULL # YOUR CODE HERE
p2
" # END PROMPT


# BEGIN SOLUTION
p2 <- c(105.75, 36.00, 83.25)
# END SOLUTION
```

```
test_that("p2a", {
  expect_true(all.equal(p2[1], 105.75, tol = 0.01))
  print("Checking: Expected value for full pay is correct")
})
```

```
## [1] "Checking: Expected value for full pay is correct"
## Test passed
```

```
test_that("p2b", {
  expect_true(all.equal(p2[2], 36.00, tol = 0.01))
  print("Checking: Expected value for partial pay is correct")
})
```

```
## [1] "Checking: Expected value for partial pay is correct"
## Test passed
```

```
test_that("p2c", {
  expect_true(all.equal(p2[3], 83.25, tol = 0.01))
  print("Checking: Expected value for no pay is correct")
})
```

```
## [1] "Checking: Expected value for no pay is correct"
## Test passed
```

**3. State the null hypothesis under which the above expected counts were computed.**

The null hypothesis is that the distribution of parental leave would equal that stated in the research article (i.e., that the proportion receiving full pay equals 47%, the proportion receiving partial pay is 16%, and the proportion with no pay is 37%).

**4. Compute the chi-squared statistic by hand and assign the test statistic to p4. Round your answer to 2 decimal places.**

```
. = " # BEGIN PROMPT
p4 <- NULL # YOUR CODE HERE
p4
" # END PROMPT


# BEGIN SOLUTION
p4 <- 68.66
# END SOLUTION
```

```
test_that("p4a", {
  expect_true(all.equal(p4, 68.66, tol = 0.01))
  print("Checking: Correct chi-square test statistic")
})
```

```
## [1] "Checking: Correct chi-square test statistic"
## Test passed
```

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

$= (105.75 - 51)\hat{}2/105.75 + (36 - 33)\hat{}2/36 + (83.25 - 141)\hat{}2/83.25 = 28.34574 + 0.25 + 40.06081 = 68.65656$

**5. Which cell (i.e. term in the summation) contributes the most to the test statistic? Assign the letter of your answer choice (e.g., "a") to `p5`.**

a) full pay
b) partial pay
c) no pay

```
. = " # BEGIN PROMPT
p5 <- NULL # YOUR ANSWER CHOICE HERE
p5
" # END PROMPT


# BEGIN SOLUTION
p5 <- "c"
# END SOLUTION
```

```
test_that("p5a", {
  expect_true(p5 == "c")
  print("Checking: Correct answer choice")
})
```

```
## [1] "Checking: Correct answer choice"
## 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. Compute the p-value for the test statistic you calculated above. Round your answer to 2 decimal places.**

```
. = " # BEGIN PROMPT
p6 <- NULL # YOUR CODE HERE
p6
" # END PROMPT


# BEGIN SOLUTION
p_value <- round(pchisq(q = 68.65656, df = 2, lower.tail = F), 2)
p6 <- 0.00
# END SOLUTION
```

```
test_that("p6a", {
  expect_true(all.equal(p6, 0.00, tol = 0.01))
  print("Checking: Correct p-value")
})
```

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

**7. Do you believe there is evidence against the null hypothesis in favor of the alternative hypothesis assuming a significance level of 0.001? Assign the letter of your answer choice (e.g., "a") to p7.**

a) in favor of null
b) against null

```
. = " # BEGIN PROMPT
p7 <- NULL # YOUR ANSWER CHOICE HERE
p7
" # END PROMPT


# BEGIN SOLUTION
p7 <- "b"
# END SOLUTION
```

```
test_that("p7a", {
  expect_true(p7 == "b")
  print("Checking: Correct answer choice")
})
```

```
## [1] "Checking: Correct answer choice"
## 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.

## Section 2: HPV

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. Which variable is explanatory and which is response? Assign the letter of your answer choice (e.g., "a") to p8.**

a) explanatory: age group, response: HPV status
b) explanatory: HPV status, response: age group

```
. = " # BEGIN PROMPT
p8 <- NULL # YOUR ANSWER CHOICE HERE
p8
" # END PROMPT


# BEGIN SOLUTION
p8 <- "a"
# END SOLUTION
```

```
test_that("p8a", {
  expect_true(p8 == "a")
  print("Checking: Correct answer choice")
})
```

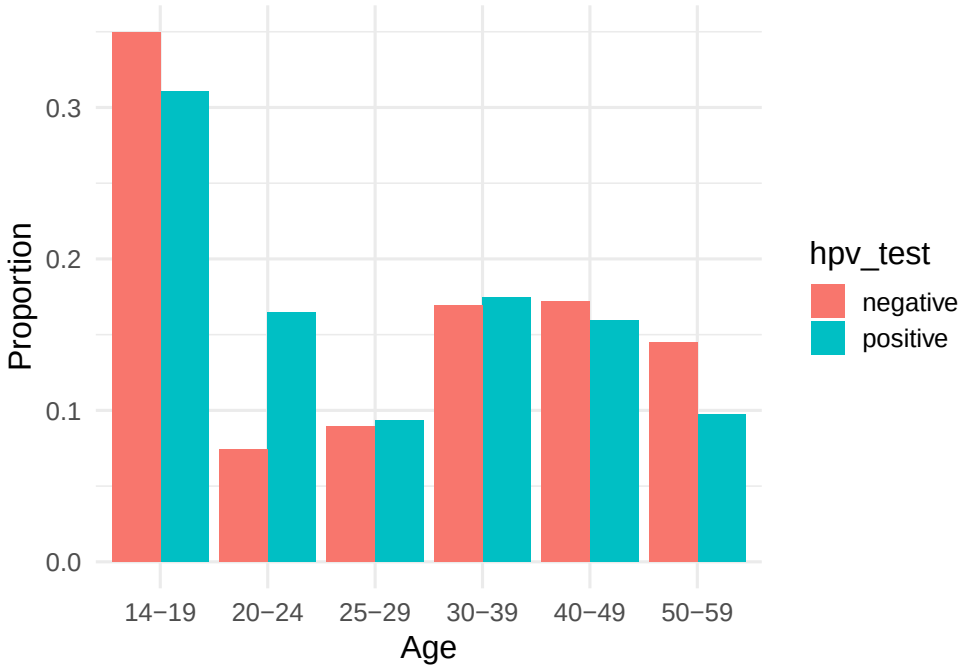
```
## [1] "Checking: Correct answer choice"
## Test passed
```

**9. 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.

**10. 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? Explain how you know.**




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. 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. 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 |

| 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. Calculate the test statistic by hand. Round your answer to 2 decimal places.**

```
. = " # BEGIN PROMPT
p12 <- NULL # YOUR CODE HERE
p12
" # END PROMPT


# BEGIN SOLUTION
p12 <- 40.55
# END SOLUTION
```

```
test_that("p12a", {
  expect_true(all.equal(p12, 40.55, tol = 0.01))
  print("Checking: Correct test statistic")
})
```

```
## [1] "Checking: Correct test statistic"
## 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. Calculate the p-value for your test statistic. Round your answer to 2 decimal places.**

```
. = " # BEGIN PROMPT
p13 <- NULL # YOUR CODE HERE
p13
" # END PROMPT


# BEGIN SOLUTION
p_value_p13 <- round(pchisq(q = 40.55353, df = 5, lower.tail = F), 2)
p13 <- 0.00
# END SOLUTION
```

```
test_that("p13a", {
  expect_true(all.equal(p13, 0.00, tol = 0.01))
  print("Checking: Correct p-value")
})
```

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

**14. Assess whether or not there is evidence against the null in favor of the alternative. Assign the letter of your answer choice (e.g., "a") to `p14`.**

a) in favor of the null
b) against the null

```
. = " # BEGIN PROMPT
p14 <- NULL # YOUR ANSWER CHOICE HERE
p14
" # END PROMPT


# BEGIN SOLUTION
p14 <- "b"
# END SOLUTION
```

```
test_that("p14a", {
  expect_true(p14 == "b")
  print("Checking: Correct answer choice")
})
```

```
## [1] "Checking: Correct answer choice"
## Test passed
```

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.

**15. Fill in the blanks. Assign p15 to a vector of the words (in quotations) to fill in the blanks.**

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.

```
. = " # BEGIN PROMPT
p15 <- NULL # YOUR CODE HERE
p15
" # END PROMPT


# BEGIN SOLUTION
p15 <- c("confidence intervals", "hypothesis tests", "with", "same", "histogram", "2.5", "97.5")
# END SOLUTION
```

```
test_that("p15a", {
  expect_true(p15[1] == "confidence intervals")
  print("Checking: confidence intervals")
})
```

```
## [1] "Checking: confidence intervals"
## Test passed
```

```
test_that("p15b", {
  expect_true(p15[2] == "hypothesis tests")
  print("Checking: hypothesis tests")
})
```

```
## [1] "Checking: hypothesis tests"
## Test passed
```

```
test_that("p15c", {
  expect_true(p15[3] == "with")
  print("Checking: with")
})
```

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

```
test_that("p15d", {
  expect_true(p15[4] == "same")
  print("Checking: same")
})
```

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

```
test_that("p15e", {
  expect_true(p15[5] == "histogram")
  print("Checking: histogram")
})
```

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

```
test_that("p15f", {
  expect_true(p15[6] == "2.5")
  print("Checking: 2.5")
})
```

```
## [1] "Checking: 2.5"
## Test passed
```

```
test_that("p15g", {
  expect_true(p15[7] == "97.5")
  print("Checking: 97.5")
})
```

```
## [1] "Checking: 97.5"
## Test passed
```