STATS 60 Summer 2020: HW4 Solutions

Due August 14, 2020 - No late days allowed

```
library(tidyverse)
```

Heads-up: We will cover materials you need to solve problem 2 and problem 3 part 2 - 3 next week.

Problem 1 Virus testing quality control

A virus test kit claims to be 99% sensitive [sensitivity is defined is Section 6.7 of the online companion]. To test this claim, you run the test on 10,000 infected samples and got 116 negatives. In this problem we want determine whether the test kit works as claimed.

(a)

State your null hypothesis and the alternative hypothesis (use one-tailed alternative for this problem).

Answer Null hypothesis: the sensitivity of the test kit is 99%.

Alternative hypothesis: the test kit is less sensitive than claimed.

(b)

Under the null hypothesis, what is the probability of obtaining at least 116 negative results?

Answer Under the null hypothesis, the number of negative test results is from a binomial distribution with parameters $n = 10,000$ and $p = 0.01$. Thus the probability is

$$\mathrm{P}(\mathrm{Binom}(n, p) \geq 116) = 0.062.$$

```
1 - pbinom(115, 10000, 0.01)
```

```
## [1] 0.06222326
```

(c)

Given our calculation, can you reject the null hypothesis at 5% level?

Answer No, we cannot reject the null hypothesis at 5% level.

(d)

What is the p-value of this test? How do you interpret the p-value?

Answer The p-value is 0.0622. If the null hypothesis is true, the chance of observing at least 116 negative results in 10,000 tests is 0.0622.

Problem 2 Lizard habitats

The following data table records the day time habits of a species of lizard, *grahami*. It contains the number of observed lizards at different times of day (early, mid-day or late) and whether the site was sunny or shaded. The question we want to answer is: do *grahami* lizards prefer sunny or shaded locations?

	Early	Mid-day	Late
Sun	47	20	22
Shade	94	205	43

(a)

State the null hypothesis and the alternative hypothsis.

Answer Null hypothesis: *grahami* lizards do not have a prefered location, i.e. probability of seeing a lizart at sunny and shady location are both 1/2.

Alternative hypothesis: *grahami* lizards prefer sunny location or shady location.

(b)

What are the expected counts in the table under the null hypothesis? (Hint: sum up each row to get a vector of length 2, i.e. number of lizards in sunny/shady location. Assuming the total number of lizards is fixed, compute the expected counts under the null for this vector)

Answer The following table is the expected counts and the observed counts: observed are 431 lizards in total, under the null hypothesis, we expect half of them in sunny / shaded location.

	Observed	Expected
Sun	89	215.5
Shade	342	215.5

(c)

Compute the pearson chi-squared statistics. What is the degree of freedom?

Answer The pearson chi-squared statistics is 148.5. The degree of freedom is $2 - 1 = 1$.

```
O <- c(89, 342) # observed counts
E <- c(215.5, 215.5) # expected counts
sum((O - E)^2 / E ) # chi-squared statistics
```

```
## [1] 148.5128
```

(d)

What is the p-value of this test? Can you reject the null hypothesis?

Answer

```r
1 - pchisq(148.51, df = 1) # p-value
```

```
## [1] 0
```

The p-value is 0. There's very strong evidence to reject the null hypothesis that grahami lizard have no location preference.

```r
# chi-squared test of goodness of fit
chisq.test(O)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  O
## X-squared = 148.51, df = 1, p-value < 2.2e-16
```

```r
# same as chisq.test(O, p = c(0.5, 0.5))
```

(e)

What is the odds that *grahami* lizard is in a sunny location compared to a shady location, when it is early in the day? What about mid-day?

Answer The odds that *grahami* lizard is in a sunny location early in the day is

$$\frac{\text{P(sunny and early in day)/P(early in day)}}{\text{P(shaded and early in day)/P(early in day)}} = \frac{47}{94} = 0.5.$$

At mid-day, the odds is 0.09. It seems lizard may prefer shaded area even more in the mid-day.

(f)

Is there evidence that the *grahami* lizards' prefered site does not depend on time of the day?

Answer We test the null hypothesis that the *grahami* lizards' prefered site is independent of time of the day. We use a chi-squared test for independence.

```r
x <- matrix(c(47, 20, 22,
              94, 205, 43), nrow = 2, byrow = TRUE)
chisq.test(x)
```

```
##
##  Pearson's Chi-squared test
##
## data:  x
## X-squared = 39.745, df = 2, p-value = 2.342e-09
```

The p-value is almost 0, and thus there's very strong evidence against the null hypothesis that the *grahami* lizards' prefered site does not depend on time of the day. Here's how to compute the chi-squared test of independence by hand. The following is the expected table under the null hypothesis.

Expected	Early	Mid-day	Late	Total
Sun	29	46	13	89
Shade	112	179	52	342
Total	141	225	65	431

Here the first cell is calculated as

$$\frac{141}{431} \times \frac{89}{431} \times 431 = \frac{141 \times 89}{431} = 29.11.$$

The chi-squared statistics is

$$\sum_{i,j} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} = 40.32,$$

and the degree of freedom is $(3-1) * (2-1) = 2$.

```
O <- c(47 , 20, 22 , 94, 205, 43) # observed counts
E <- c(29, 46, 13, 112, 179, 52) # expected counts
sum((O-E)^2 / E)
```

```
## [1] 40.32592
```

```
1-pchisq(40.32, df = 2) # pvalue
```

```
## [1] 1.756399e-09
```

Problem 3 Heart health

Does consuming red meat increase the risk of heart disease? We will think about this question from two perspectives in this problem.

Part 1

It is conjectured that red meat increase the risk of heart disease because it contains a substance called choline, which is used to produce trimethylamine (TMA). The liver converts TMA into TMAO (trimethylamine N-oxide). It is known that high blood levels of TMAO with a higher risk for both cardiovascular disease and early death from any cause (see this website)

A group of researchers recruited 30 volunteers, and randomly assign half of them to baseline diets and the other half a diet high in red meat. After four weeks, plasma TMAO level was measured for each volunteer. The measurements are recorded below. Does this study provide evidence that a diet high in red meat increase the plasma level of TMAO? State the null hypothesis and compute the p-value.

```
value_redmeat <- c(11.50, 4.39, 11.79, 11.14, 18.14, 11.40, 15.15, 7.16,
                   5.49, 7.92, 5.10, 9.92,19.55,1.49,3.78) # red meat group
value_baseline <- c(4.42, 4.83 ,2.49, 5.39, 4.41, 5.20, 3.07, 4.34, 3.41,
                    6.11 ,5.06 ,4.87, 3.77, 6.02, 2.08) # baseline group
```

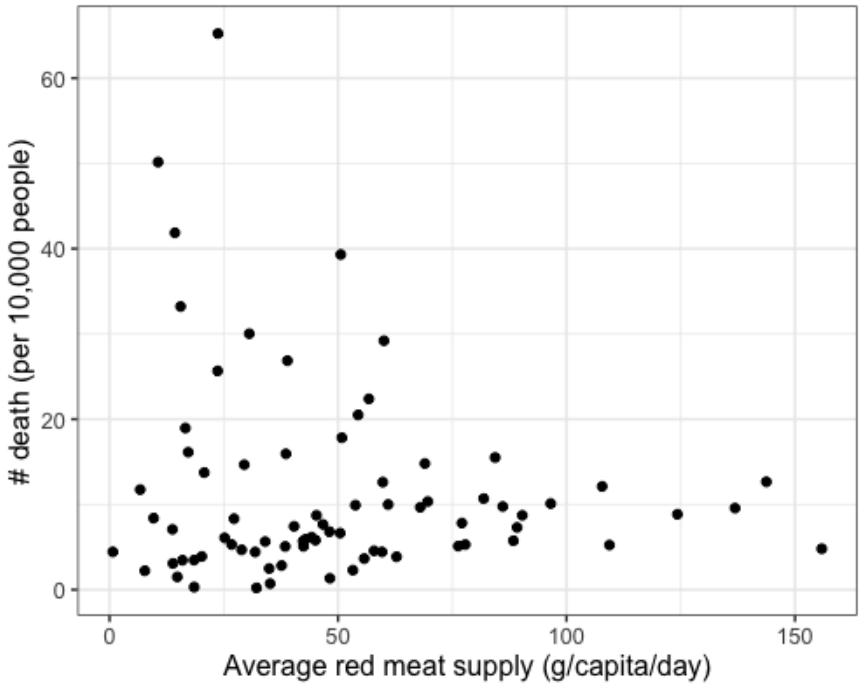
Answer The null hypothesis is the mean TMAO level is the same for baseline and red meat group. The alternative hypothesis is mean TMAO level is higher in the red meat group.

We use a one-sided t-test.

```
t.test(value_baseline, value_redmeat, var.equal = FALSE, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  value_baseline and value_redmeat
## t = -3.7442, df = 15.446, p-value = 0.0009337
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -2.785931
## sample estimates:
## mean of x mean of y
##  4.364667  9.594667
```

The p-value is 0.0009, which provides evidence that the red meat group has higher average level of TMAO in the plasma.



Part 2

The plot on the last page shows the average red meat supply and the number of death from ischaemic heart diseases (per 10,000 people) in a number of countries in the year 2012. The countries are not complete due to missing data.

In particular, here are the data from 5 countries. Compute the Pearson correlation, and Spearman correlation between red-meat supply and the number of death for these 5 countries.

Country	Red-meat supply	# death
Japan	25.2	6.08
Fiji	45.3	8.72
Armenia	60.0	29.2
Colombia	48.1	6.80
Croatia	38.9	26.9

Answer

```
supply <- c(25.2, 45.3, 60.0, 48.1, 38.9)
mortality <- c(6.08, 8.72, 29.2, 6.80, 26.9)
cor(supply, mortality, method = "pearson") # pearson correlation
```

```
## [1] 0.5004048
```

```
cor(supply, mortality, method = "spearman") # spearman correlation
```

```
## [1] 0.6
```

Part 3

(a) If I tell you that the Pearson correlation between # death and total red meat supply is -0.138 in the plot shown in part (2) and the spearman correlation is 0.105, does it provide evidence that consuming red meat is related to the risk of heart disease?

(b) Does the study in part (1) provide evidence that red meat consumption is related to higher risk of heart disease?

If the answer is yes, briefly explain why. If the answer is no, explain what concerns you have or what other information you need.

Answer (a) No, it does not provide evidence that consuming red meat is related to the risk of heart disease. Pearson correlation shows some negative correlation, while Spearman correlations shows positive correlation between the ranks, so it is inconclusive whether there is a positive correlation or not. The correlations we observed may be due purely to chance, and a test may provide more confidence. Finally, we need to beware that correlation does not imply causation.

(b) Yes. The randomized study in part 1 showed that consuming red meat leads to a higher level of TMAO on average. Because of the relationship between TMAO and heart health, we can conclude consuming red meat is related to heart health. Note the study does not answer how red meat consumption affects TMAO level for any particular individual.

Problem 4 Open Policing data

This problem looks at possible policing bias in traffic stops. The dataset is from the open policing project link. In this problem we are just scratching the surface, but the website link has many cool statistical analysis that are also quite accessible, so feel free to take a look if you are interested. You can access dataset from course website. We are interested in vehicle stopping in San Francisco in the year 2015, and only for MPC, moving violation or non-moving violation. Now let's load the data.

```
# change file to the location of the data
sf_stopping <- read.csv(file = "sf_stopping.csv", sep="")
```

(a)

For each racial group, count the number of drivers who are stopped.

Answer

```
sf_stopping_race <- sf_stopping %>%
  group_by(subject_race) %>%
  count() # count the number of drivers in each race
sf_stopping_race
```

```
## # A tibble: 5 x 2
## # Groups:   subject_race [5]
##   subject_race               n
##   <fct>                  <int>
## 1 asian/pacific islander 15451
## 2 black                  14792
## 3 hispanic               11803
## 4 other                  13594
## 5 white                  29677
```

(b)

The population of San Francisco in the year 2015 is 860,000. Further, according to 2018 US Census Bureau estimates, San Francisco County's population was 40.0% Non-Hispanic White, 5.4% Hispanic White, 5.2% Black or African American, 34.3% Asian, 8.1% Some Other Race, 0.3% Native American and Alaskan Native, 0.2% Pacific Islander and 6.5% from two or more races Wiki link. For simplicity, we group hispanics with other racial group, and we assume the demographics is the same in 2018 and 2015, thus the population is 40.8% white, 34.5% Asian, 5.2% black and 19.5% others.

From this information, test the null hypothesis that, in the year 2015, the demographics of drivers stopped for MPC and moving/non-moving violations is the same as population decomposition.

Answer We use a chi-squared test, and the p-value is essentially zero, so we reject the null hypothesis that the demographics of drivers stopped for MPC and moving/non-moving violations is the same as population decomposition.

```
# observed counts for each race
observed <- c(15451, 14792, 25397, 29677)
# expected counts
expected <- sum(sf_stopping_race$n) * c(0.345, 0.052,0.195 ,0.408)
# chi-squared statistics
chi2 <- sum((expected - observed)^2 / expected)
```

```r
# degree of freedom
df <- 3
# p-value
1 - pchisq(chi2, df)
```

```
## [1] 0
```

```r
observed - expected # fewer asian and white and more black and other race are stopped
```

```
## [1] -13983.365  10355.516   8760.185  -5132.336
```

```r
# use chisq.test function
chisq.test(observed, p = c(0.345, 0.052,0.195 ,0.408))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  observed
## X-squared = 36184, df = 3, p-value < 2.2e-16
```

```r
# if you use sf_stopping_small
chisq.test(c(860, 890, 1472, 1778), p = c(0.345, 0.052,0.195 ,0.408))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  c(860, 890, 1472, 1778)
## X-squared = 2247.3, df = 3, p-value < 2.2e-16
```

(c)

Based on your calculations, can you conclude that traffic stopping is biased? We would like you to think about other reasons that may explain the observed discrepancy even if the police is not biased.

Optional Based on your answer in part (c), can you think of a method, or what data you will need, to figure out whether stopping is biased?

Answer From part b, the racial decomposition is different for drivers who are stopped at San Francisco in 2015 for MPC and moving/non-moving violations and the general population.

There are many possible explanations, other than policing bias, to account for the discrepancy. For example,

1. The driving population who drive may be different from the general population. For example, some groups take public transportation more.
2. The proportion of drivers that violates traffic rules may be different in each group.
3. Perhaps the police are not biased, but the number of police are different in different neighborhoods.
4. Perhaps some drivers in San Francisco are not residents at San Francisco.
5. The general population demography may be different in 2015 and 2018.

We can control for other factors that may explain the discrepancy such as stopping location, driving population at a location, overall offense rate, type of cars etc..