hw4_sol_draft

Qian Zhao

7/29/2020

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

```r
library(tidyverse)
library(lubridate)
```

Problem 1 Virus testing quality control

A virus test kit claims to be 99% sensitive. 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 The null hypothesis is that the test kit is 99% sensitive. The alternative is 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 The number of negative test results is from a binomial distribution with parameters $n = 10,000$ and $p = 0.01$. Thus the probability is given by

$$\text{prob}(\text{Binom}(n, p) \geq 116) = 0.062.$$

The probability is 0.0622.

```r
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. This means if the null hypothesis were true, the chance of observing at least 116 negative results in 10,000 tests is at 0.0622.

Problem 2 Lizard habitats

The following data table records the day time habits of a species of lizard, *grahami*. It records 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* 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 The null hypothesis is grahami lizard do not prefer sunny location over shady location. The alternative hypothesis is otherwise.

(b)

If the null hypothesis were true, what are the expected counts?

Answer If the null hypothesis were true, then regardless of what time of day it is, we should expect half the lizards at sunny location and half at shady location. We compute the marginal totals for each column and mulply by 0.5.

Expected	Early	Mid-day	Late
Sun	70.5	112.5	32.5
Shade	70.5	112.5	32.5
Total	141	225	65

(c)

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

Answer Compare with the observed table, the pearson chi-squared statistics is 174.56. The degree of freedom is 3.

```r
O <- c(47 , 20, 22 , 94, 205, 43) # observed counts
E <- c(70.5, 112.5, 32.5, 70.5, 112.5, 32.5) # expected counts
sum((O - E)^2 / E ) # chi-squared statistics
```

```
## [1] 174.5624
```

(d)

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

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

```r
1 - pchisq(174, df = 3)
```

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

(d)

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

Answer The odds ratio early in the day is

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

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

(e)

Is there evidence that the grahami lizard's preference of sites does not depend on time of the day?

Answer Use a chi-squared test for independence. Compute the marginal probabilities and the expected counts under the null.

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

The chi-squared statistics is 40.32, and the degree of freedom is $(3-1)*(2-1) = 2$, so we have very strong evidence to reject the null hypothesis.

```r
E <- c(29, 46, 13, 112, 179, 52)
sum((O-E)^2 / E)
```

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

Problem 3 Heart health

Does consuming red meat increase the risk of heart disease? In this problem, we think about this question from two perspectives.

Part 1

(a) 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.

```r
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 You can use a one-sided t-test. There is strong evidence that the red meat group has higher level of TMAO in the plasma.

```r
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
```

You can also use a permutation test. The p-value is almost the same as two-sample t-test in this example. Note we generate the samples from a normal distribution, so p-value from the t-test is correct.

```r
B <- 1000
x <- c(value_redmeat, value_baseline)
t_perm <- numeric(B)
for(b in 1:B){
  a <- sample(1:30, 10, replace = FALSE) # randomly assign baseline label
  # compute t-statistics
  t_stat <- (mean(x[-a]) - mean(x[a])) / sqrt(var(x[-a])/15 + var(x[a])/15)
  t_perm[b] <- t_stat
}
mean(t_perm>3.74)
```

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

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 for these 5 countries.

Country	Red-meat supply	# death
Japan	25.2	6.08

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

Answer

Part 3

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

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

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.

```
sf_stopping <- read.csv("~/Desktop/sf_stopping.csv", sep="")
```

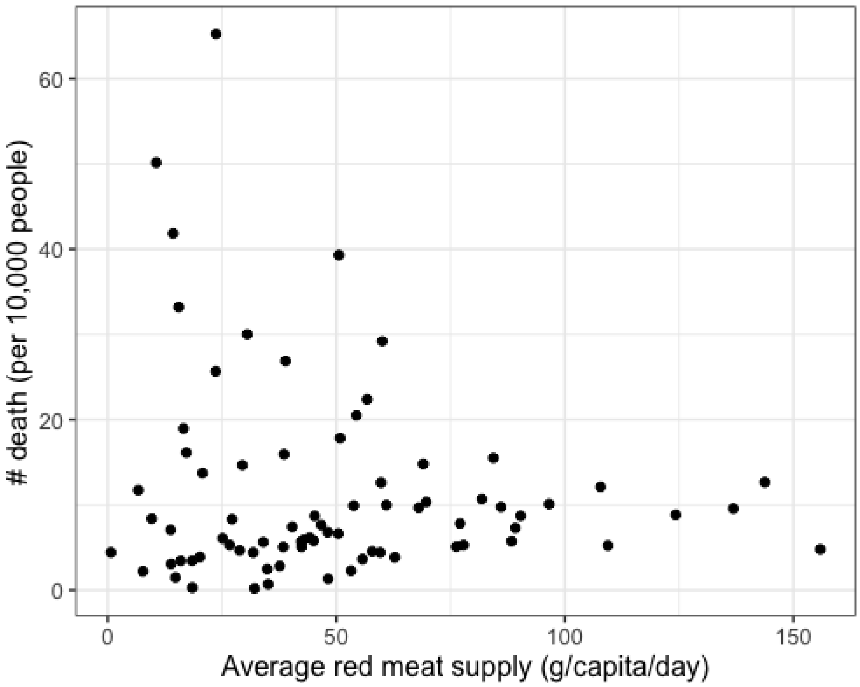
(a)

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. For simplicity, we group hispanics with other racial group, also 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.

```
# observed counts
observed <- sf_stopping_race$n
observed[4] <- observed[3] + observed[4]
observed <- observed[-3]
# expected counts
expected <- sum(sf_stopping_race$n) * c(0.345, 0.052,0.195 ,0.408)
# chi-squared statistics
chi2 <- sum((expected - observed)^2 / expected)
# what's the degree of freedom
df <- 3
# what's the deviation
1 - pchisq(chi2, df)
```

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

```
observed - expected # stopping much fewer asian and white, more black and other race
```

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

(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 For example, perhaps the proportion of drivers that violates traffic rules are different in each racial group.