STATS 60 Summer 2020: HW4

Due August 14, 2020 - No late days allowed

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).

(b)

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

(c)

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

(d)

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

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.

(b)

What are the expected counts in the table under the null hypothesis?

(c)

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

(d)

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

(d)

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

(e)

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

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
```

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

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

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.

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.

(b)

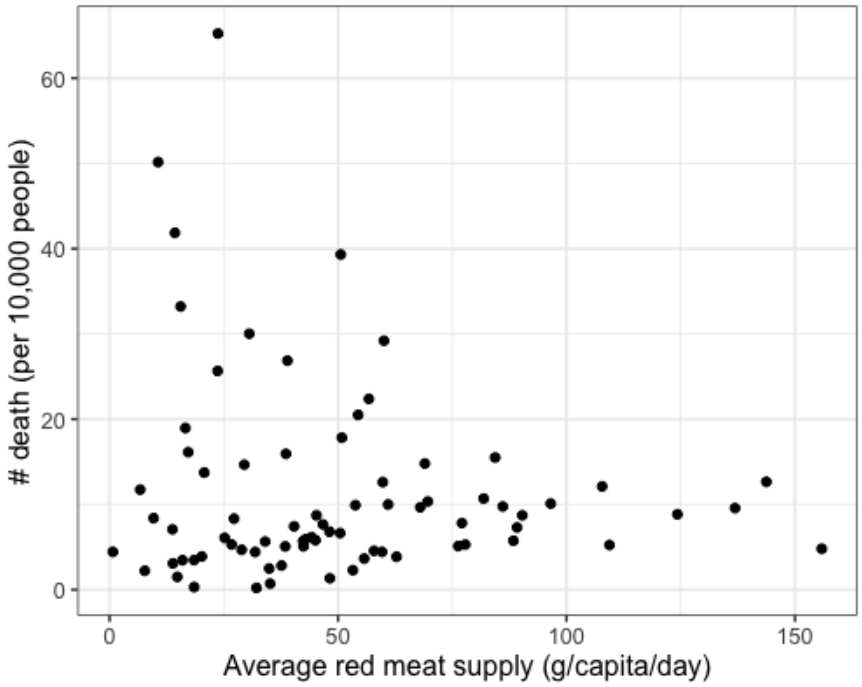
The population of San Francisco in the year 2015 is 860,000. Further, according to 2018 US Census Bureau estimates, San Francisco County's population was 40.0% Non-Hispanic White, 5.4% Hispanic White, 5.2% Black or African American, 34.3% Asian, 8.1% Some Other Race, 0.3% Native American and Alaskan 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.

(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?


60
40
20
0
death (per 10,000 people)
0
50
100
150
Average red meat supply (g/capita/day)