Math 189: Homework 3

Alan Lui, Derek So, Xiangyu Wei

January 28, 2021

Introduction

This study looks into the recommended women's nutritional intake, sampling from data gathered in the United States Department of Agriculture (USDA) Women's Health Survey dataset (nutrient.txt). We use statistical hypothesis testing method, such as univariate t-test, and two familywise error rate (FWER) correction methods, such as Bonferroni Method and Holm's Method, to determine whether US women meet the recommended nutrient intake amount based on recommended levels given in the Dataset section. We determine whether each nutrient meet the recommended intake level, and make suggestions on the adjustment of nutrient intake to the public for future benefits.

Tasks & Analysis

Dataset

The dataset is from a study of women's nutrition commissioned by USDA in 1985. Nutrient intake of five nutritional components (Calcium, Iron, Protein, Vitamin A and Vitamin C) were collected from a random sample of 737 women aged 25-50 years. The units of these variables are measured in different magnitudes of grams. Calcium, Iron, and Vitamin C are measured in milligrams, Protein in grams, and Vitamin A in micrograms. The dataset is downloaded from https://github.com/tuckermcelroy/ma189/blob/main/Data/nutrient.txt at 2021-01-28 20:15:24 PST. The recommended levels of each nutrient are listed below.

- Calcium: 1000mg
- Iron: 15mg
- Protein: 60g
- Vitamin A: 800µg
- Vitamin C: 75mg

```
library(tidyverse)
nutrient <- read.table("Data/nutrient.txt")[,-1] # load data
col.names <- c("Calcium", "Iron", "Protein",
               "Vitamin A", "Vitamin C") # new column names
colnames(nutrient) <-  col.names # rename column
head(nutrient, 5)
```

```
##   Calcium   Iron Protein Vitamin A Vitamin C
## 1  522.29 10.188  42.561    349.13    54.141
## 2  343.32  4.113  67.793    266.99    24.839
## 3  858.26 13.741  59.933    667.90   155.455
## 4  575.98 13.245  42.215    792.23   224.688
## 5 1927.50 18.919 111.316    740.27    80.961
```

Methods

After some simple exploration by calculating the sample mean and variance, our methods were the t-test, along with two FWER correction methods (Bonferroni Correction and Holm's Method), using the significance level of $\alpha = .05$. The t-test compares the means of the sample to the recommended levels of each nutrient. Then we use the Bonferroni correction to compensate for multiple hypotheses being tested simultaneously by rejecting null hypotheses with p-values less than $\frac{\alpha}{n}$, along with the Holm's method which sorts the p-values to reject a true hypothesis of at most $\alpha = .05$.

Analysis

1. Calculate sample mean and sample standard deviation of each variable.

We use `colMeans()` to calculate the sample means, and `sd()` to calculate the sample standard deviation of each nutrient. Then, means and standard deviations are combined together using a `data.frame()` for display.

```
mean <- colMeans(nutrient)
sd <- c(sd(nutrient[,1]), sd(nutrient[,2]), sd(nutrient[,3]),
        sd(nutrient[,4]), sd(nutrient[,5]))
mu_mat <- data.frame(mean, sd)
colnames(mu_mat) <- c("Sample Mean", "Sample S.D.")
mu_mat
```

```
##           Sample Mean Sample S.D.
## Calcium     624.04925   397.27754
## Iron         11.12990     5.98419
## Protein      65.80344    30.57576
## Vitamin A   839.63535  1633.53983
## Vitamin C    78.92845    73.59527
```

From the table above, we can see that the central tendency (sample mean) and variation (sample standard deviation) for Calcium (M = 624.05(mg), SD = 397.28), Iron (M = 11.13(mg), SD = 5.98), Protein (M = 65.80(g), SD = 30.58), Vitamin A (M = 839.64(µg), SD = 1633.54), and Vitamin C (M = 78.83(mg), SD = 73.60).

2. The recommended intake amount of each nutrient is given in the table below. For each nutrient, apply a univariate t-test to test if the population mean of that variable equals the recommended value. Set the significance level at $\alpha = .05$.

We loop through all the columns in `nutrient` and compare each nutrient intake distribution with the recommended intake level using `t.test()`. All five t-statistics and p-values are saved to `t.stats` and `p.values` respectively. Then, the t-statistics and p-values are combined together using a `data.frame()` for display. See more in **Appendix » T-Test** for detailed explanation of the method.

```
# declare variables
alpha <- .05
null_mu <- c(1000,15,60,800,75)
t.stats <- rep(NA, length(null_mu))
p.values <- rep(NA, length(null_mu))
```

```r
# loop through each nutrient
for (i in 1:length(null_mu)){
  t_test <- t.test(nutrient[,i], # each nutrient data
                   mu = null_mu[i], # recommended level
                   alternative = "two.sided") # two-side test
  t.stat <- t_test$statistic # t-stats
  p_value <- t_test$p.value # p values
  t.stats[i] <- t.stat
  p.values[i] <- p_value
}

# format output
sig.mat <- data.frame(col.names, t.stats, p.values, p.values <= alpha)
colnames(sig.mat) <- c("Nutrient", "T-stats", "P-value", "Significance (alpha = .05)")
sig.mat
```

```
##    Nutrient     T-stats       P-value Significance (alpha = .05)
## 1   Calcium -25.6903891 2.108727e-104                       TRUE
## 2      Iron -17.5570105  6.654391e-58                       TRUE
## 3   Protein   5.1527860  3.299706e-07                       TRUE
## 4 Vitamin A   0.6586985  5.102954e-01                      FALSE
## 5 Vitamin C   1.4491210  1.477297e-01                      FALSE
```

3. Repeat step 2, now using the Bonferroni and Holm's Methods to control the FWER for the five tests. How does this affect the results?

Bonferroni Correction. Bonferroni correction is first applied by deviding the α by the number of hypothesis testing (5) that we are making. Then, `data.frame()` is used to combine and display the results. See more in **Appendix » Family Wise Error Rate (FWER)** for why we need error rate correction. And see more in **Appendix » The Bonferroni Correction** for detailed explanation of the method.

```r
alpha.bonferroni <-  alpha / length(null_mu)
sig.mat.bonf <- data.frame(col.names, t.stats, p.values, alpha.bonferroni,
                           p.values <= alpha.bonferroni) # result dataframe
colnames(sig.mat.bonf) <- c("Nutrient", "T-stats", "P-value", "Alpha (Bonf.)",
                            "Significance (Bonf.)") # rename columns
sig.mat.bonf
```

```
##    Nutrient     T-stats       P-value Alpha (Bonf.) Significance (Bonf.)
## 1   Calcium -25.6903891 2.108727e-104          0.01                  TRUE
## 2      Iron -17.5570105  6.654391e-58          0.01                  TRUE
## 3   Protein   5.1527860  3.299706e-07          0.01                  TRUE
## 4 Vitamin A   0.6586985  5.102954e-01          0.01                 FALSE
## 5 Vitamin C   1.4491210  1.477297e-01          0.01                 FALSE
```

Holm's Method Holm's method is then applied to adjust each α value used for sorted p-value hypothesis testing, with α being at most .05. Then, `data.frame()` is used to combine and display the results. See more in **Appendix » Family Wise Error Rate (FWER)** for why we need error rate correction. And see more in **Appendix » The Holm's Method** for detailed explanation of the method.

```r
# declare variables
sorted.pvals <- sort(p.values)
alpha.holms <- rep(NA, length(sorted.pvals))

# loop through each nutrient
for (j in 1:length(sorted.pvals)){
  alpha.holm <- alpha / (length(sorted.pvals) - j + 1)
  alpha.holms[j] <- alpha.holm
}

# format output
sig.mat.holm <- data.frame(col.names, t.stats, p.values)# result dataframe
sig.mat.holm <- arrange(sig.mat.holm, p.values) %>% # arrange
  mutate(alpha = alpha.holms, # add alpha values
         sig = p.values <= alpha.holms) # add significance
colnames(sig.mat.holm) <- c("Nutrient", "T-stats", "P-value", "Alpha (Holm's)",
                            "Significance (Holm's)") # rename columns
sig.mat.holm
```

```
##    Nutrient     T-stats       P-value Alpha (Holm's) Significance (Holm's)
## 1   Calcium -25.6903891 2.108727e-104     0.01000000                  TRUE
## 2      Iron -17.5570105  6.654391e-58     0.01250000                  TRUE
## 3   Protein   5.1527860  3.299706e-07     0.01666667                  TRUE
## 4 Vitamin C   1.4491210  1.477297e-01     0.02500000                 FALSE
## 5 Vitamin A   0.6586985  5.102954e-01     0.05000000                 FALSE
```

From both the Bonferroni Correction result table and the Holm's Method result table, we can see that both result remain the same as when we are just using $\alpha = .05$, with signficant result of Calcium ($t(736) = -25.690, p < 0.001$), Iron ($t(736) = -17.557, p < 0.001$) and Protein ($t(736) = 5.153, p < 0.001$), and non-significant result of Vitamin A ($t(736) = 1.449, p = 0.148$), and Vitamin C ($t(736) = 0.659, p = 0.510$).

4. Based on the results you obtained in steps 2 and 3, how would you interpret your test results? Do you think the US Women meet the recommended nutrient intake amount? If not, what would you suggest to the public?

Based on the results from step 2 & 3, we would conclude that we can reject the null hypothesis that Calcium, Iron, and Protein intake are no different than their respective recommended level, while we will still retain the null hypothesis that Vitamin A and Vitamin C nutritional intake are the same as their recommended level. The results suggest that US women probably meet their recommended nutrient intake amount for Vitamin A and Vitamin C, but needs adjustment for Calcium, Iron, and Protein since these nutritions are signficiant different from the recommended level. More specifically, based on their respective t-statistics, we would recommend increasing the Calcium and Iron intake (negative t-statistics indicating lower than recommended level intake) and decreasing the Protein intake (positive t-statistics indicating higher than recommended level intake).

Conclusion

Using just the t-test, we can comfortably reject the null hypothesis regarding US women's Calcium, Iron, and Protein intake being no different than the recommended level, due to p-values of extremely small magnitudes. However, in regards to Vitamin A and Vitamin C, we cannot reject the null hypothesis. For those that we could reject, the t-statistics were significantly extreme, with the average US women's Calcium and Iron intake

falling far below average, and Protein intake being above average. However, for Vitamin A and Vitamin C, the differences were much more mild.

In conclusion, using the t-test results, we recommend to the US women that: intake of Calcium and Iron should be increased, the intake of Protein should be decreased, while Vitamin A and Vitamin C may remain as is.

Given the Bonferroni correciton, all the α of .05 is divided by the number of comparisons, giving us .01. This changes nothing of significance for any of our p-values, and thus the recommendation stays the same. The same is true with the Holm's method, as while new α were generated, the same null hypotheses were rejected and not rejected as before using the same p values, and again the recommendation stays the same.

Discussion

Note that the dataset consists of 737 women. Considering that the population of the data set is the United States, with 50 states and 5 major territories, this leaves approximately 15 individuals per state or territory, assuming equal sampling from each. It is dangerous to generalize the conclusions of these significance tests given such small sample sizes for each relatively unique region of the United States. Some of the possibly underrepresented features may be ethnicity and socioeconomic status.

Given the extreme p-values of the t-tests for the variables that we rejected, it is unsurprising that the two other methods did not change our conclusion, along with the high p-values for the two variables we did not reject. The limitation of those methods is that we can only change the α around the values we calculated. Further exploration would use other methods we have not yet learned, to see if other conclusions could be gathered or whether this conclusion should be further reinforced.

Appendix

T-Test

A test statistic named the t-statistic are used to address the hypothesis testing problem in the study because the population standard deviation is unknown.

$$T = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}.$$

The t-statistic measures the ratio between the deviation of sample mean $\overline{x}$ from the hypothesized value μ_0 and its standard error $s/\sqrt{n}$.

Family Wise Error Rate (FWER)

When multiple significance testings are conducted at once, the rate at which we committee Type-I errors increases. The Family Wise Error Rate is defined as the probability of having at least 1 Type-1 Error with p number of significance tests:

$$FWER = 1 - (1 - \alpha)^p$$

In order to account for this substantially increasing function, we correct the significance value α to reject less tests. Keep in mind that decreasing the rate of rejecting tests, also decreases the power of the test and increases the beta of the test.

The Bonferroni Correction

A popular method to correct FWER is called the Bonferroni Correction, where the significance level, α, is divided by the number of significance tests conducted, m:

$$\alpha_{bonf} = \frac{\alpha}{m}$$

Note that the Bonferroni Correction required the assumption that the significance tests be independent to reach a FWER of α.

The Holm's Method

Another method of correcting FWER is called Holm's Method. It is a simple sequential method to correct the FWER by ranking the calculated p-values in ascending order, and creating a robustly adjusted significance level for testing. The sorted p-values may be in the form:

$$p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(j)}$$

To calculate the adjusted significance level, we utilize the following formula, where j is the rank of the p-value, and m is the number of significance tests:

$$\alpha_{holm} = \frac{\alpha}{m - j + 1}$$

One drawback of utilizing the Holm's Method is that it becomes difficult to define a confidence interval because the significance level of the test α, varies per test, and as a result, the confidence of the interval will also change.