Math 189 HW 3: Data Analysis of USDA Women's Health Survey

Yunchun Pan, Siqing Lyu, Nathan Ng, Pudan Xu

1/25/2021

Introduction

In this homework, we examine the USDA Women's Health Survey dataset **nutrient.txt**[1]. This dataset contains five types of women's nutrient intakes which were measured from a random sample of 737 women aged 25-50 years in the United States. There are 5 nutrients in the dataset: **Calcium (mg)**, **Iron (mg)**, **Protein (g)**, **Vitamin A (ug)**, and **Vitamin C (mg)**.

The recommended intake amount of each nutrient is shown below:

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

We first analyze the distribution of each nutrient intake amount by calculating their sample means and sample standard deviations. Then, to test if the population mean of each nutrient intake equals to the recommended value, we perform a two-sided univariate t-test for each variable. However, this naive scheme to test multivariate means does not control family-wise error, hence we have a high chance to reject at least one null hypotheses when all null hypotheses are true. To deal with this problem, we use the Bonferroni and Holm's Methods to control FWER and analyze how these correction methods affect the results of 5 univariate tests. Then, we perform a one-sided univariate t-test for each nutrient to evaluate whether the US women meet the recommended nutrient intake amount. Based on the results, we make some suggestions to the public.

Our Work

We first load the USDA Women's Health Survey dataset into R to get the data we use throughout this assignment.

```
nutrient <- read.table("~/Desktop/ma189/Data/nutrient.txt")
nutrient$V1=NULL
colnames(nutrient)=c("Calcium (mg)", "Iron (mg)", "Protein (g)",
                     "Vitamin A (ug)", "Vitamin C (mg)")
head(nutrient)
```

```
##   Calcium (mg) Iron (mg) Protein (g) Vitamin A (ug) Vitamin C (mg)
## 1       522.29    10.188      42.561         349.13         54.141
```

[1]Source: The USDA Women's Health Survey dataset contains data extracted from the 1985 study commissioned by the USDA on women's nutrition.

```
## 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
## 6       607.58     6.800      45.785         165.68         13.050
```

After we load in our dataset, we calculate the sample mean and sample standard deviation of each variable. Then, we combine the results with recommended intake amount of each nutrient and save them into **RecomMeanSd**. By observing the table, we find that standard deviations of Calcium, Iron, and Protein intake are approximately 50% of their respective means. For Vitamin C intake, standard deviation is close to its respective mean, while the standard deviation of Vitamin A intake is almost two times larger than its respective mean. In general, standard deviations of these nutrients intake are large relative to their respective means. Hence, it indicates a high variablity among women nutrient intake, especially in Vitamin A and C. The reason might be that the recommended intake values are so small (i.e. 800 ug) and people find it hard to measure and intake the exact amount. Therefore, slightly more or less intake of each nutrient would cause the standard deviation to be large relative to the mean.

```
recom <- c(1000, 15, 60, 800, 75)
RecomMeanSd <- cbind(recom, apply(nutrient,2,mean), apply(nutrient,2,sd))
colnames(RecomMeanSd) <- c("Recommended Intake", "Sample Mean","Sample Std")
RecomMeanSd
```

```
##                Recommended Intake Sample Mean Sample Std
## Calcium (mg)                 1000   624.04925  397.27754
## Iron (mg)                      15    11.12990    5.98419
## Protein (g)                    60    65.80344   30.57576
## Vitamin A (ug)                800   839.63535 1633.53983
## Vitamin C (mg)                 75    78.92845   73.59527
```

To test if the population mean of each nutrient equals to the recommended value, we apply a two-sided univariate t-test for each variable. We first set the significance level at $\alpha = .05$ and calculate the critical value using the sample size **n** and the significance level **alpha**. In each test, we save the recommended value as **mu**, sample mean as **xbar**, and sample standard deviation as **sdev**. Then, we calculate the test statistic **t**, compare its absolute value with the critical value **crit**. If the absolute value of **t** is less than or equal to **crit** for all nutrients, we fail to reject the null hypothesis that the population mean of each nutrient equals to the recommended value. If absolute value of at least one test statistic is greater than **crit**, we reject the null hypothesis that the population mean of each nutrient equals the recommended values.

```
alpha <- .05
mu_mat <- RecomMeanSd[,1:2]
MeanSd <- RecomMeanSd[,2:3]
n <- dim(nutrient)[1]
crit <- qt(p=alpha/2, df=n-1, lower.tail=FALSE)
test <- NULL
tval <- NULL
for (i in 1:5) {
  mu <- mu_mat[i,1]
  xbar <- mu_mat[i,2]
  sdev <- MeanSd[i,2]
  t <- (xbar - mu)/(sdev/sqrt(n))
  tval <- c(tval, t)
  if (abs(t) > crit){test <- c(test,"reject")}
  else {test <- c(test,"fail to reject")}
```

```
}
cat("critical value:", crit)
```

```
## critical value: 1.963192
```

```
result = cbind(colnames(nutrient), abs(tval), test)
colnames(result) <- c("Nutrient", "Test Statistics", "Conclusions")
result
```

```
##      Nutrient         Test Statistics     Conclusions
## [1,] "Calcium (mg)"   "25.6903891016745"  "reject"
## [2,] "Iron (mg)"      "17.557010480273"   "reject"
## [3,] "Protein (g)"    "5.15278599421882"  "reject"
## [4,] "Vitamin A (ug)" "0.658698493121023" "fail to reject"
## [5,] "Vitamin C (mg)" "1.44912103621398"  "fail to reject"
```

Since absolute values of test statistics for Calcium, Iron, and Protein nutrients are greater than critical value 1.963192, we reject the null hypothesis that the population mean of each nutrient equals to the recommended value.

According to Lecture 8, if we do not control family-wise error, we have a high chance to reject at least one null hypotheses when they are all true and commit Type I error. Hence, we use the Bonferroni correction to control the FWER for the five univariate t-tests for evaluating if the population mean of each nutrient equals to the recommended value. Using the Bonferroni Correction, we change the significance level for each individual test to $\alpha/m = 0.05/5 = 0.01$ where **m** is the number of nutrients in the dataset. By doing so, we decrease the probability of rejecting the null, and thus decrease the probability of committing a Type I error. Then, in each test, we calculate the test statistic **t**, compute its p-value **pval**, and compare it with the adjusted significance level **sig**. If at least one **pval** is less than **sig**, we reject the null hypothesis that the population mean of each nutrient equals to the recommended value. Otherwise, we fail to reject the null hypothesis.

```
alpha <- .05
n <- dim(nutrient)[1]
m <- 5
sig <- alpha/m
tests <- NULL
pvals <- NULL
for (i in 1:5) {
  mu <- mu_mat[i,1]
  xbar <- mu_mat[i,2]
  sdev <- MeanSd[i,2]
  t <- (xbar - mu)/(sdev/sqrt(n))
  pval <- 2*pt(abs(t),df=n-1, lower.tail = FALSE)
  pvals <- c(pvals, pval)
  if (pval < sig){tests <- c(tests,"reject")}
  else {tests <- c(tests,"fail to reject")}
}
cat("significance level:", sig)
```

```
## significance level: 0.01
```

```
result = cbind(colnames(nutrient), pvals, tests)
colnames(result) <- c("Nutrient", "P-values", "Conclusions")

result
```

```
##      Nutrient         P-values                Conclusions
## [1,] "Calcium (mg)"   "2.10872717908103e-104" "reject"
## [2,] "Iron (mg)"      "6.65439129290145e-58"  "reject"
## [3,] "Protein (g)"    "3.29970631409601e-07"  "reject"
## [4,] "Vitamin A (ug)" "0.510295406450313"     "fail to reject"
## [5,] "Vitamin C (mg)" "0.147729707627124"     "fail to reject"
```

Since p-values of univariate t-tests for Calcium, Iron, and Protein are extremely low and less than the adjusted significance level 0.01, we reject the null hypothesis that the population mean of each nutrient equals to the recommended values. Even though we reduce the chance of a Type I error through the correction, our conclusion still leads us to reject the null hypothesis. Therefore, we are more confident that our conclusion is correct.

The table below also shows how the Bonferroni Correction helps control the FWER compared to our naive scheme. As the number of variables that are being tested increases, there is a significant increase in the FWER from the naive scheme. Although the FWER is equal to the alpha level of 0.05 for 1 variable, as more variables are introduced, the FWER goes beyond the alpha level, reaching 0.2262191 when we only test five variables. On the other hand, the Bonferroni Correction, in the right column, yields a FWER of approximately 0.05. As the number of variables that we are testing increases, the FWER after Bonferroni Correction continues to stay approximately 0.05.

```
alpha <- .05
m <- seq(0,5,1)
fwer_naive <- 1 - (1-alpha)^m
fwer_bon <- 1 - (1 - alpha/m)^m
cbind(m,cbind(fwer_naive,fwer_bon))
```

```
##      m fwer_naive   fwer_bon
## [1,] 0  0.0000000 0.00000000
## [2,] 1  0.0500000 0.05000000
## [3,] 2  0.0975000 0.04937500
## [4,] 3  0.1426250 0.04917130
## [5,] 4  0.1854938 0.04907029
## [6,] 5  0.2262191 0.04900995
```

Another method to control family-wise error rate is to adjust our significance levels according to Holm's Method. Following this method, we sort the p-values calculated from our data in ascending order, and set the significance level for each individual test to $sig = \alpha/(m - j + 1)$, where m is the number of nutrients and j is the order of the corresponding p-value. j ranges from 1 to m, and the lowest p-value has j = 1, while the highest p-value has j = m. Starting with the lowest p-value, we compare the p-value with the adjusted significance level, rejecting the null hypothesis if the p-value is lower than the significance level and failing to reject when the p-value is greater than the significance level. This process stops when we encounter the first fail to reject, and we fail to reject all null hypotheses after. Holm's method also controls family-wise error rate like the Bonferroni correction, but Holm's method has a lower increase of Type II error risk, or not rejecting when null hypothesis is not true.

```r
alpha <- .05
m <- 5
t <- (mu_mat[,2]-mu_mat[,1])/(MeanSd[,2]/sqrt(n))
pvals <- 2*pt(abs(t),df=n-1, lower.tail = FALSE)
pvals_sort <- sort(pvals)
index <- seq(1,m)
crits <- alpha/(m - index +1)
nulls <- NULL
j <- 1
while(j <= m)
{
  if(pvals_sort[j] <= crits[j])
  {
    nulls <- c(nulls,j)
  } else { j <- m+1 }
  j <- j+1
}
# length(nulls)
comps <- pvals_sort <= crits
testH <- ifelse(comps == TRUE, "Reject", "Fail to Reject")
holm_method_df <- data.frame("Index"=index,"Sorted_P_Values"=pvals_sort,
                             "Significance_Level"=crits,"Conclusion"=testH)
holm_method_df
```

```
##                Index Sorted_P_Values Significance_Level     Conclusion
## Calcium (mg)       1   2.108727e-104         0.01000000         Reject
## Iron (mg)          2    6.654391e-58         0.01250000         Reject
## Protein (g)        3    3.299706e-07         0.01666667         Reject
## Vitamin C (mg)     4    1.477297e-01         0.02500000 Fail to Reject
## Vitamin A (ug)     5    5.102954e-01         0.05000000 Fail to Reject
```

As the table shows, variables are organized in ascending order of p-values. The lowest p-value is on the top and the highest p-value is on the bottom. Using Holm's method, we adjust the significance level according to the order of the p-value. As the index of each variable we test increases, the significance level for that test also increases. The significance level ranges from 0.01, which is the one we use for Bonferroni's Correction, to 0.05, which is the significance level of our naive scheme.

The p-values for Calcium, Iron, and Protein are all below the adjusted significance level, and thus we reject null hypotheses for those variables. Since we encounter a fail to reject when we test for Vitamin C, we also fail to reject the null hypothesis for Vitamin A. To sum up, after using Holm's method to control FWER and limit the chance of committing Type I error, we still reject the null hypothesis that the population mean of each nutrient equals to the recommended levels.

To test whether the US women meet the recommended nutrient intake amount, we apply a left-sided univariate t-test for each variable. Moreover, we use the Bonferroni Correction to control FWER and reduce the chance of committing Type I error. We change the significance level to $\alpha/m = 0.05/5 = 0.01$ where **m** is the number of nutrients in the dataset. Then, in each test, we calculate the test statistic **t**, compute its p-value **pval**, and compare it with the adjusted significance level **sig**. If at least one **pval** is less than **sig**, we reject the null hypothesis that the US women meet the recommended nutrient intake amount. Otherwise, we fail to reject this null hypothesis.

```r
alpha <- .05
n <- dim(nutrient)[1]
m <- 5
```

```r
sig <- alpha/m
tests <- NULL
pvals <- NULL
mu_mat <- RecomMeanSd[,1:2]
MeanSd <- RecomMeanSd[,2:3]
for (i in 1:5) {
  mu <- mu_mat[i,1]
  xbar <- mu_mat[i,2]
  sdev <- MeanSd[i,2]
  t <- (xbar - mu)/(sdev/sqrt(n))
  pval <- pt(t, df=n-1, lower.tail = TRUE)
  pvals <- c(pvals, pval)
  if (pval < sig){tests <- c(tests,"reject")}
  else {tests <- c(tests,"fail to reject")}
}
cat("significance level:", sig)
```

```
## significance level: 0.01
```

```r
result = cbind(colnames(nutrient), pvals, tests)
colnames(result) <- c("Nutrient", "P-values", "Conclusions")
result
```

```
##      Nutrient         P-values                Conclusions
## [1,] "Calcium (mg)"   "1.05436358954052e-104" "reject"
## [2,] "Iron (mg)"      "3.32719564645072e-58"  "reject"
## [3,] "Protein (g)"    "0.999999835014684"     "fail to reject"
## [4,] "Vitamin A (ug)" "0.744852296774844"     "fail to reject"
## [5,] "Vitamin C (mg)" "0.926135146186438"     "fail to reject"
```

Since p-values of left-sided univariate t-tests for Calcium and Iron are extremely low and less than the adjusted significance level 0.01, we reject the null hypotheses that the US women meet the recommended intake of Calcium and Iron. Therefore, we reject the null hypothesis that the US women meet the recommended nutrient intake amount, and we would like to suggest US women to intake more Calcium and Iron nutrients.

Conclusion:

After exploring sample means and sample deviations of each nutrient intake, we find that in general, standard deviations of nutrients intake are large relative to their respective means, especially Vitamin A and Vitamin C. This indicates a high variablity among women nutrient intake.

To test if the population mean of each nutrient equals to the recommended value, we first perform two-sided univariate t-tests for each variable. To control FWER for five tests, we apply Bonferroni Correction and Holm's Method. The results of all three methods lead us to reject the null hypothesis that the population mean of each nutrient equals to the recommended value. Moreover, we conclude that Calcium, Iron, and Protein intakes are significantly different from the recommended levels. Since Bonferroni Correction and Holm's Method both adjust the significance level to reduce the chance of committing to Type I error, it should be difficult to reject null hypotheses when they are all true. Therefore, we are confident that there is a difference between the average levels of nutrient intake by all United States women aged 25-50 years and the recommended levels of nutrient intake.

To evaluate if the US women meet the recommended nutrient intake amount, we perform left-sided univariate t-tests and adjust significance level by Bonferroni Correction. Since the p-values of tests for Calcium and

Iron are extremely low and less than significance level, we reject the null hypothesis that the US women meet the recommended nutrient intake amount. Because Bonferroni Correction reduces the chance of committing Type I error, we believe this conclusion is correct. Moreover, observing sample means and recommended levels of nutrients intake, we find that US women, on average, have too little Calcium and Iron and too much Protein. Based on our observations and results of all tests, we would suggest that US women have more Calcium and Iron, less Protein, and continue to intake the same amount of Vitamin A and C.