Trace Tidwell

10/1/2016

STAT 603 – Statistical Methods – Project 1

1)

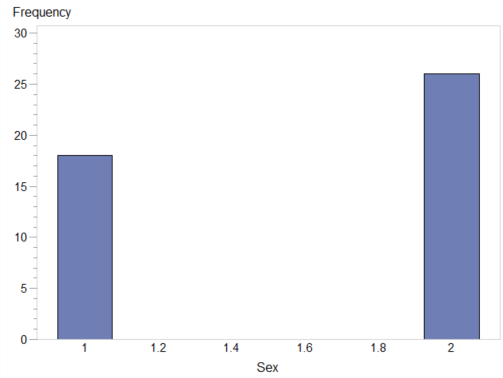


Figure 1.1 Bar Graph of Sex

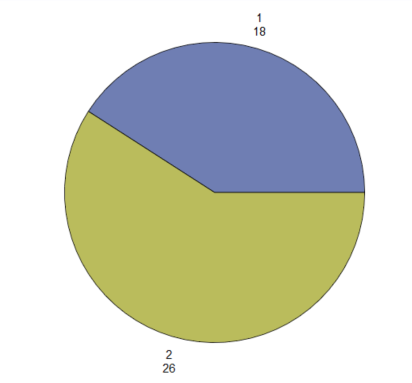


Figure 1.2 Pie Chart of Sex

From these two graphs, we can see that there were 18 girls (1 = girl) and 26 boys (2 = boy) born on December 21, 1977 at a Brisbane, Australia hospital.

2)

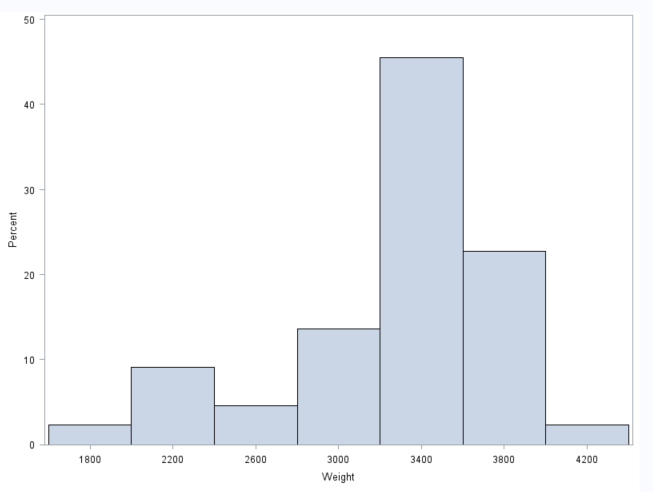


Figure 1.3 Histogram of Weight

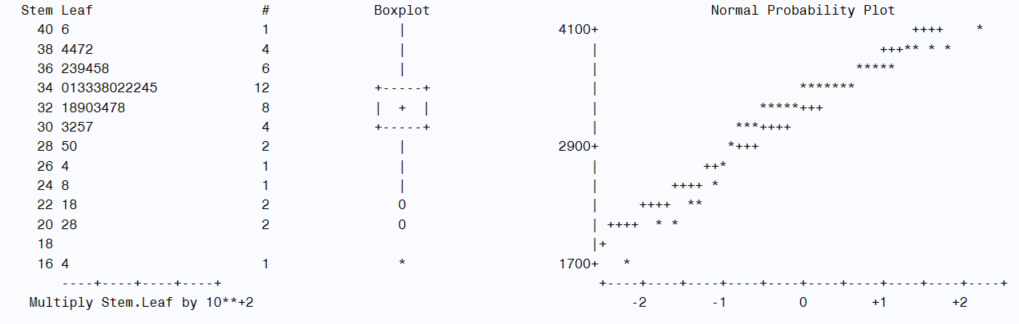


Figure 1.4 Boxplot of Weight

From Figures 1.3 we can see that the majority of the babies born fall into the bins between 3000 grams and 3800 grams. On the extreme ends, one child weighed only approximately 1800 grams, while one child weighed nearly 4200 grams. The median weight was approximately 3380 grams, while the mode was approximately 3400 grams, and the mean was approximately 3273 grams. The distribution is skewed to the left, with mode > median > mean. The distribution is also slightly bimodal, with another small peak around the 2200 gram bin. Lastly, from the histogram, we can estimate the variance as 281099 gram2 and the standard deviation as 530 grams.

These numbers are confirmed by Figure 1.4 Additionally, we see that the IQR is approximately 400 grams, with Q1 at 3200 grams and Q3 at 3600 grams. The max is around 4100 grams, and the min is around 1600 grams. Finally, we see mild outliers at around 2000 and 2200 grams, and we see the min is an extreme outlier at around 1600 grams.

3)

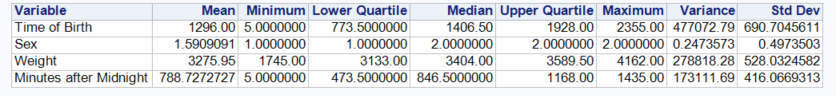


Figure 1.5 Descriptive Statistics

From the description data we see a few things. First, it should be noted that the time of birth should be ignored due to the method in which it was recorded. The pertinent measurement here is minutes after midnight. Using descriptive statistics on the sex variable can be done, but must be done knowing that sex is a qualitative variable. Knowing that 1 = girl and 2 = boy, we can quickly see that more boys were born that girls by looking at the mode (though the mean and median also indicate more boys).

We can also compare the statistics on weight with our grouped data statistics from part 2. We see that mean is 3276 grams (vs. 3273 above), and the median is 3404 grams (vs. 3380 above). Both estimates were pretty close. Q1, Q3, and the IQR are actually 3133, 3589, and 456 grams, respectively versus our estimates of 3200, 3600, and 400 grams, respectively. Again, these are pretty good estimates. Finally, we’ll compare the variance and standard deviation to the estimated values from the grouped data. The calculated variance for the sample is 278818 gram2 (vs. 281099 gram2), while the calculated standard deviation is 528 grams (vs. 530 grams). Applying the Empirical Rule to the Weight data, we can check the distribution. The intervals in question are 3276 ± 528 (68%), 3276 ± 1056 (95%), and 3276 ± 1584 (99.7%). The first interval actually contains 73% of the sample, the second interval contains 91% of the sample, and the third interval contains 100% of the sample. These numbers are close to the empirical rule, but the bimodal, right-skewing nature of the distribution causes them not to match up as closely as would a unimodal symmetric distribution.

Analyzing the Minutes after Midnight data, we can see that average time of birth was 1:08 PM, with the median time of birth occurring at 2:06 PM. The earliest birth occurred at 12:05 AM, and the latest birth occurred at 11:55 PM. Q1 occurred at 7:53 AM, Q3 occurred at 7:28 PM, and the IQR was 11 hours and 35 minutes. Based on the standard deviation of 6 hours 56 minutes and the mean time of birth of 1:08 PM, we see that the birth times are fairly uniformly distributed throughout the day, with the data slightly weighted towards later in the day.

4)

/\* --------------------------------------------------------------------

Code generated by a SAS task

Generated on Wednesday, September 28, 2016 at 6:01:59 PM

By task: Import Data Wizard

Source file: C:\Users\Trace\Documents\MS\STAT-603 Statistical

Methods\Project 1\babyboom.xlsx

Server: Local File System

Output data: WORK.babyboom

Server: Local

-------------------------------------------------------------------- \*/

/\* --------------------------------------------------------------------

This DATA step reads the data values from a temporary text file

created by the Import Data wizard. The values within the temporary

text file were extracted from the Excel source file.

-------------------------------------------------------------------- \*/

**DATA** WORK.babyboom;

LENGTH

'Time of Birth'n **8**

Sex **8**

Weight **8**

'Minutes after Midnight'n **8** ;

FORMAT

'Time of Birth'n BEST12.

Sex BEST12.

Weight BEST12.

'Minutes after Midnight'n BEST12. ;

INFORMAT

'Time of Birth'n BEST12.

Sex BEST12.

Weight BEST12.

'Minutes after Midnight'n BEST12. ;

INFILE 'C:\Users\Trace\AppData\Local\Temp\SEG8752\babyboom-7f85ea339bb74a4aab4481a6feb3dd7c.txt'

LRECL=**16**

ENCODING="WLATIN1"

TERMSTR=CRLF

DLM='7F'x

MISSOVER

DSD ;

INPUT

'Time of Birth'n : BEST32.

Sex : BEST32.

Weight : BEST32.

'Minutes after Midnight'n : BEST32. ;

**RUN**;

/\*Bar Graph of "Sex"\*/

**proc** **gchart** data=babyboom;

vbar sex;

**run**;

/\*Pie Chart of "Sex"\*/

**proc** **gchart** data=babyboom;

pie sex;

**run**;

/\*Histogram of "Weight"\*/

**proc** **univariate** data=babyboom;

histogram weight;

**run**;

/\*Boxplot of "Weight"\*/

**proc** **univariate** data=babyboom plot;

var Weight;

**run**;

/\*Descriptive Statistics\*/

**proc** **means** data=babyboom mean min q1 median q3 max var std;

**run**;

2) Per the provided documentation, “the purpose of the study was to determine if significant gender differences exist in the mean values of calcium, inorganic phosphorus, and alkaline phosphatase in subjects over age 65. A second purpose was to determine if analytical variation between laboratoreis (sic) would affect the mean values of the study variables.” The study contained 178 subjects over the age of 65 who had their calcium, inorganic phosphorous, and alkaline phosphatase measured. The measurements were done at 1 of 6 labs, and the subjects were identified as either male or female and placed into one of 5 age groups.

Upon loading the file, one of the first things we see are errors in the data. For example, three of the ages were input as 699, 730, and 771. These are very clearly wrong. There are also three values for sex that are neither 1 (male) or 2 (female), so these are also bad. There are only six labs, yet two of the records have the lab as 21 and 43. We see values for calcium (CAMMOL) in the twenties when the rest of the values are between 1.05 and 3.2. Finally, many of the observations are missing data. The next step in the analysis is to clean up the data, so that we can properly compare the good data.

By creating a second dataset of cleaned data, we can be much more certain that the analysis is representative of the study done and not skewed by poor data entry. We start by getting rid of all ages greater than 89, all sexes greater than 2, all labs greater than 6, and all calcium measurements greater than 3.2 (based on the initial analysis of values). We then remove all blank values, which leaves us with 152 records. This is what we’ll use to perform the analysis.

Since the purpose of the study is to compare differences in men and women, a pie chart was made to view the makeup of the study by sex. As shown in Figure 2.1, men make up nearly 54% of the subjects (82 of 152), while women make up the remaining 46% (70 of 152).

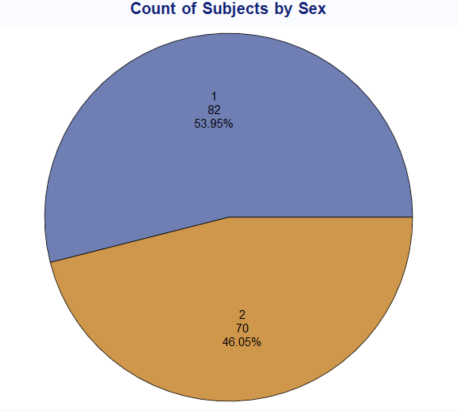


Figure 2.1 Count of Subjects by Sex

A secondary goal of the study is to compare differences between labs, so a bar graph was made to show how many subjects each lab saw. Figure 2.2 shows that over half of the subjects were seen at Lab 1, with Lab 2 seeing approximately 25%. The remaining four labs only saw 20% of the subjects, with Lab 3 seeing only 1.32%. This is not a very good distribution of subjects among the different labs, and if a goal of the study is to compare labs, it will be difficult to do with the small sample size at Labs 3-6.

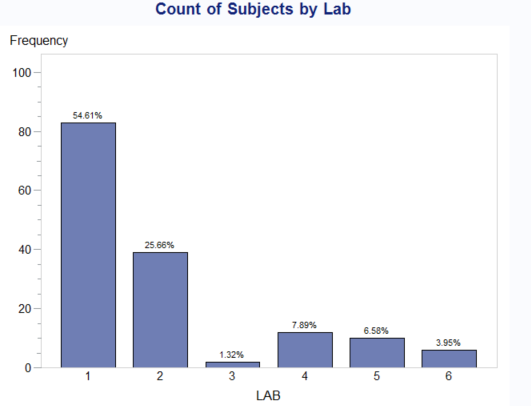


Figure 2.2 Count of Subjects by Lab

Finally, it might prove useful to view the breakdown of the subjects by age group. This is shown in Figure 2.3. From it, we see that most of the subjects fall between the ages of 65-79, with only 5% in the 80-84 Age Group and 2% in the 85-89 Age Group. While we would certainly expect fewer subjects in the higher age groups, we might not have enough subjects from Age Groups 4 & 5 to accurately describe these populations as a whole.

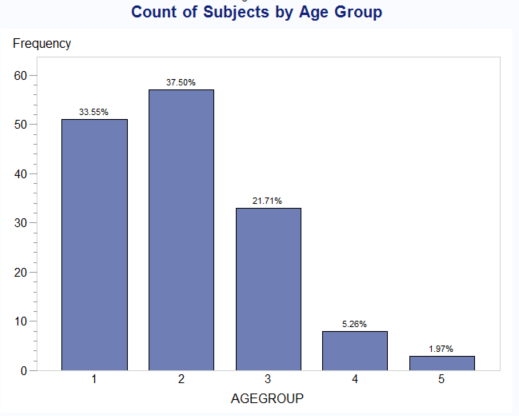


Figure 2.3 Count of Subjects by Age Group

With a better understanding of the makeup of the subjects by sex, age, and location of measurements, we can begin to compare the results of the measurements themselves. First we perform an overall comparison of male and female subjects. Figure 2.4 shows key descriptive statistics for calcium (CAMMOL), alkaline phosphatase (ALKPHOS), phosphorus (PHOSMMOL), and age.

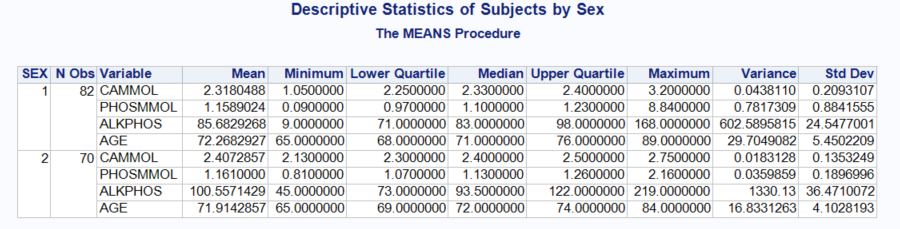


Figure 2.4 Descriptive Statistics of Subjects by Sex

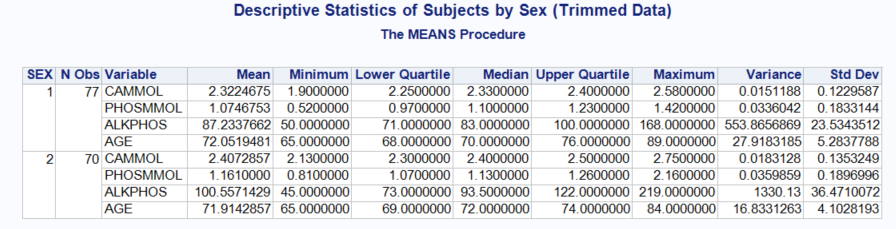


Figure 2.5 Descriptive Statistics of Subjects by Sex (Trimmed Data)

The above table contains very important information, but it can be rather cumbersome to make of sense. In order to see the results more clearly, we can compare the average values of the three minerals for men and women, as shown in Figures 2.6, 2.7, and 2.8.

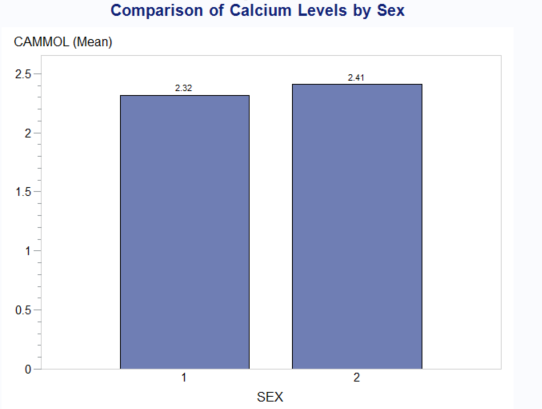


Figure 2.6 Average Calcium Levels by Sex

From Figure 2.6, we see that the average calcium value for males was 2.32 mmol and for females was 2.41 mmol. The absolute difference of 0.09 mmol seems rather small, but since the values themselves are rather small, the 0.09 mmol results in a 3.73% difference between the men and the women, which could warrant further examination. In comparing the values from Figure 2.4, we see that the median, Q1, and Q3 are all very similar between the sexes. However, the range of values for the males is wider than that of the females, resulting a larger standard deviation. I suspect that the minimum value of 1.05 mmol for the males is incorrect, as the next lowest value is 1.9 mmol. It is also possible that the maximum value of 3.2 mmol for males is incorrect, as the next highest value is 2.75 mmol for a female and 2.58 mmol for a male. Were we to remove these two extremes, the values for the males and females are quite so spread out, as shown in Figure 2.5. Though the mean doesn’t change much, the standard deviation becomes 0.1230 for males, which is much more in line with the 0.1353 for females.

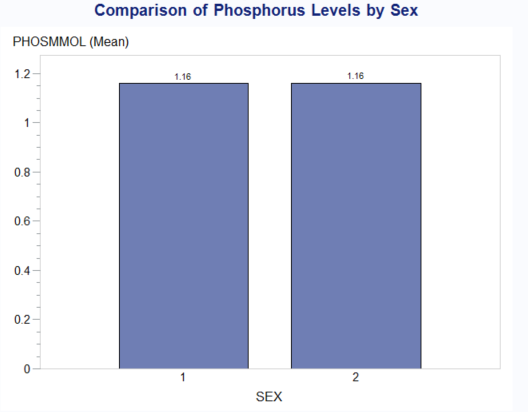


Figure 2.7 Average Phosphorus Levels by Sex

The average phosphorus levels for both men and women are the same at 1.16 mmol. Once again, the median, Q1, and Q3 are very similar, with the range and standard deviation showing the greatest differences. Here again, it seems that the minimum value of 0.09 mmol could be wrong, as the next highest value is 0.52 mmol. The maximum value of 8.84 is almost certainly an error, as the next highest value is for females is 2.16 mmol (which may also be an error) and for males is 1.58 mmol. Adjusting for the possible errors gives us a mean of 1.07 mmol and a standard deviation of 0.1833 mmol for the males, as shown in Figure 2.5. The corrected values would result in a 7.8% difference between the males and females, which could also be significant.

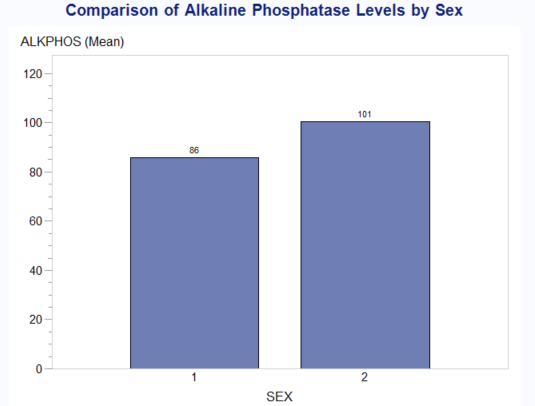
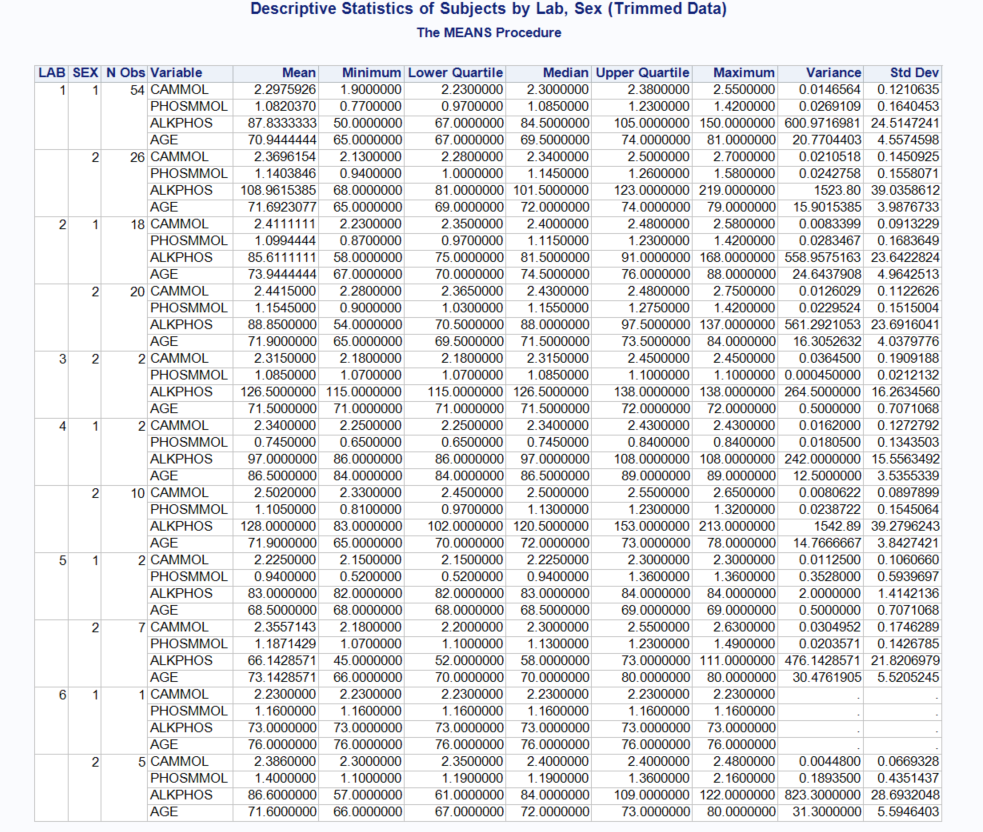


Figure 2.8 Average Alkaline Phosphatase

It is with the alkaline phosphatase that we see the greatest difference between males and females, with males having average values of 86 IU and females having average values of 101 IU from Figure 2.8 above. However, here again we see a possible error in the minimum value for males. The measure of 9 IU may be an error, since the next smallest measurement is 50 IU for males and 45 for females. In this case, though, removing the value doesn’t make a marked difference on the mean or the standard deviation, bringing them to 87.23 IU and 23.53 IU, respectively. This still represents a nearly 14% difference between the sexes. Once again, without knowing the desired values, all we can say is that there are considerable differences between males and females in alkaline phosphatase levels.

Finally, we want to compare data between the labs. Since we determined some possible issues with the data in the steps above, we will use the trimmed data for this part of the analysis. Figure 2.9 below shows the descriptive statistics of the subjects split out by lab and sex.



Though this information is useful and necessary to the analysis, it is quite cumbersome and difficult to make sense of. In order to simply things we have put together bar charts comparing the average values of the three minerals broken down by sex and lab, as shown in Figures 2.10-2.12.

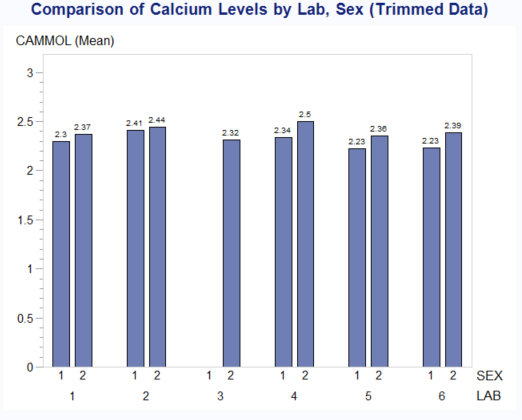


Figure 2.10 Comparison of Average Calcium Levels by Lab, Sex (Trimmed Data)

From Figure 2.10 we can see that there is definitely some variation in the measures between the labs. The range for males is 2.23 mmol to 2.41 mmol, and the range for females is 2.32 mmol to 2.5 mmol.

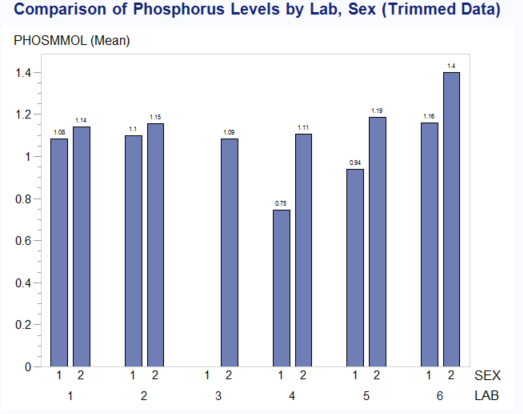


Figure 2.11 Comparison of Average Phosphorus Levels by Lab, Sex (Trimmed Data)

The average phosphorus levels are much more spread out than the calcium levels for both males and females. We see a range of 0.75 mmol to 1.16 mmol for males and a range of 1.09 mmol to 1.4 mmol for the females.

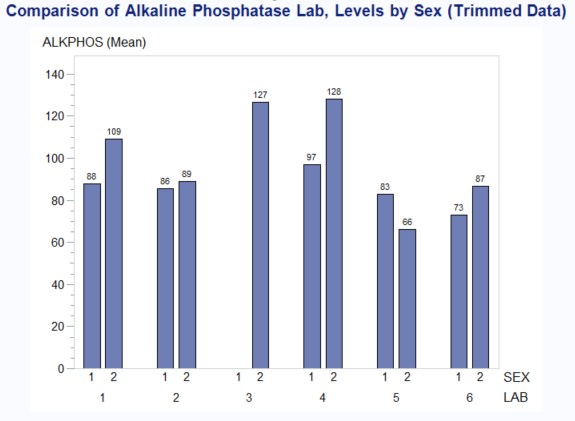


Figure 2.12 Comparison of Average Alkaline Phosphatase Levels by Lab, Sex (Trimmed Data)

Finally, we compare average results of alkaline phosphatase as shown in Figure 2.12. In this case, the female measurements show much more variability than the male measurements, with a range of 66 IU to 128 IU, versus only 73 IU to 97 for males.

After viewing the data from multiple different angles, it seems very clear that are major issues with data collection, the distribution of subjects by age and sex, and the assignment of the labs. The trimmed data set had only 147 records versus 178 for the original file. This means roughly 17% of the original data were deemed incorrect or, at the very least, questionable. The original data set was over 50% male, and the trimmed data set was 52% male. With women making up over 50% of the US population, and most likely an even higher percentage of the over 65 population, the sample should have probably been selected to reflect these observations. The distribution of subjects throughout the age groups is another possible cause for concern. I would not say that the distribution of subjects in the age groups need to necessarily match their proportions in the general population, but I would say that more subjects should have been tested in the higher age groups. Age Group 4 had only 8 subjects, while Age Group 5 had only 3 subjects. With such small sample sizes, it becomes much easier for an outlier to greatly influence the numbers for the group. Finally, in order to determine actual discrepancies among the labs, the subjects should have been randomly assigned to one of the labs in an even distribution. Lab 1 performed 56% of the measurements per the trimmed data, while Lab 3 performed only 2 measurements with both being female. Labs 4 and 5 each only had 2 males, and Lab 6 had only 1 male. Additionally, it would make sense to have an even distribution of ages across the labs.

In conclusion, though we can some trends, namely that male values are lower than those of females for all the measured minerals, the study itself was very poorly designed and conducted.

SAS Code

/\* --------------------------------------------------------------------

Code generated by a SAS task

Generated on Sunday, October 2, 2016 at 10:30:59 AM

By task: Import Data Wizard

Source file: C:\Users\Trace\Documents\MS\STAT-603 Statistical

Methods\Project 1\calcium.csv

Server: Local File System

Output data: WORK.CALCIUM\_0000

Server: Local

-------------------------------------------------------------------- \*/

**DATA** WORK.calcium;

LENGTH

OBSNO **8**

AGE **8**

SEX **8**

ALKPHOS **8**

LAB **8**

CAMMOL **8**

PHOSMMOL **8**

AGEGROUP **8** ;

FORMAT

OBSNO BEST3.

AGE BEST3.

SEX BEST2.

ALKPHOS BEST3.

LAB BEST2.

CAMMOL BEST4.

PHOSMMOL BEST4.

AGEGROUP BEST1. ;

INFORMAT

OBSNO BEST3.

AGE BEST3.

SEX BEST2.

ALKPHOS BEST3.

LAB BEST2.

CAMMOL BEST4.

PHOSMMOL BEST4.

AGEGROUP BEST1. ;

INFILE 'C:\Users\Trace\AppData\Local\Temp\SEG10944\calcium-6f24a9a9a05a459a98655b47bb8f1586.txt'

LRECL=**26**

ENCODING="WLATIN1"

TERMSTR=CRLF

DLM='7F'x

MISSOVER

DSD ;

INPUT

OBSNO : ?? BEST3.

AGE : ?? BEST3.

SEX : ?? BEST2.

ALKPHOS : ?? BEST3.

LAB : ?? BEST2.

CAMMOL : ?? COMMA4.

PHOSMMOL : ?? COMMA4.

AGEGROUP : ?? BEST1. ;

**RUN**;

/\* Remove the age records that are greater than 89 (since they are clearly wrong)

Remove the sex records that greater than 2

Remove the remaining records that are blank\*/

**data** calciumclean;

set calcium;

if age <= **89**;

if sex <= **2**;

if lab <= **6**;

if cammol <= **3.2**;

if age ^= **.**;

if lab ^= **.**;

if alkphos ^= **.**;

if cammol ^= **.**;

if phosmmol ^= **.**;

if agegroup ^= **.**;

**run**;

/\* Remove the age records that are greater than 89 (since they are clearly wrong)

Remove the sex records that greater than 2

Remove the remaining records that are blank\*/

**data** calciumclean2;

set calciumclean;

if cammol > **1.05**;

if cammol < **3.2**;

if alkphos > **9**;

if phosmmol > **0.09**;

if phosmmol < **8.84**;

**run**;

/\*Bar Graph of "Sex"\*/

**proc** **gchart** data=calciumclean;

title1'Count of Subjects by Sex';

vbar sex / levels=**2** inside=percent;

**run**;

/\*Pie Chart of "Sex"\*/

**proc** **gchart** data=calciumclean;

pie sex / percent=inside slice=inside value=inside;

**run**;

/\*Bar Graph of "Lab"\*/

**proc** **gchart** data=calciumclean;

title1'Count of Subjects by Lab';

vbar lab / levels=**6** outside=percent;

**run**;

/\*Pie Chart of "Lab"\*/

**proc** **gchart** data=calciumclean;

pie lab / levels=**6**;

**run**;

/\*Bar Graph of "Age Group"\*/

**proc** **gchart** data=calciumclean;

title1'Count of Subjects by Age Group';

vbar agegroup / levels=**5** outside=percent;

**run**;

/\*Pie Chart of "Age Group"\*/

**proc** **gchart** data=calciumclean;

pie agegroup;

**run**;

/\*Descriptive Statistics of Subjects by Sex\*/

**proc** **means** data=calciumclean mean min q1 median q3 max var std;

title1'Descriptive Statistics of Subjects by Sex';

class sex;

var cammol phosmmol alkphos age;

**run**;

/\*Bar Graph of Average Calcium Level by Sex\*/

**proc** **gchart** data=calciumclean;

title1'Comparison of Calcium Levels by Sex';

vbar sex / levels=**2** type=mean sumvar=cammol mean;

**run**;

/\*Bar Graph of Average Phosphorus Level by Sex\*/

**proc** **gchart** data=calciumclean;

title1'Comparison of Phosphorus Levels by Sex';

vbar sex / levels=**2** type=mean sumvar=phosmmol mean;

**run**;

/\*Bar Graph of Average Alkaline Phosphatase Level by Sex\*/

**proc** **gchart** data=calciumclean;

title1'Comparison of Alkaline Phosphatase Levels by Sex';

vbar sex / levels=**2** type=mean sumvar=alkphos mean;

**run**;

/\*Descriptive Statistics of Subjects by Lab, Sex\*/

**proc** **means** data=calciumclean mean min q1 median q3 max var std;

class lab sex;

var cammol phosmmol alkphos age;

**run**;

/\*Bar Graph of Average Calcium Level by Lab, Sex\*/

**proc** **gchart** data=calciumclean;

title1'Comparison of Calcium Levels by Sex';

vbar sex / levels=**2** type=mean sumvar=cammol mean group=lab;

**run**;

/\*Bar Graph of Average Phosphorus Level by Lab, Sex\*/

**proc** **gchart** data=calciumclean;

title1'Comparison of Phosphorus Levels by Sex';

vbar sex / levels=**2** type=mean sumvar=phosmmol mean group=lab;

**run**;

/\*Bar Graph of Average Alkaline Phosphatase Level by Lab, Sex\*/

**proc** **gchart** data=calciumclean;

title1'Comparison of Alkaline Phosphatase Levels by Sex';

vbar sex / levels=**2** type=mean sumvar=alkphos mean group=lab;

**run**;

/\*Descriptive Statistics of Subjects by Sex (Trimmed Data)\*/

**proc** **means** data=calciumclean2 mean min q1 median q3 max var std;

title1'Descriptive Statistics of Subjects by Sex (Trimmed Data)';

class sex;

var cammol phosmmol alkphos age;

**run**;

/\*Bar Graph of Average Calcium Level by Sex (Trimmed Data)\*/

**proc** **gchart** data=calciumclean2;

title1'Comparison of Calcium Levels by Sex (Trimmed Data)';

vbar sex / levels=**2** type=mean sumvar=cammol mean;

**run**;

/\*Bar Graph of Average Phosphorus Level by Sex (Trimmed Data)\*/

**proc** **gchart** data=calciumclean2;

title1'Comparison of Phosphorus Levels by Sex (Trimmed Data)';

vbar sex / levels=**2** type=mean sumvar=phosmmol mean;

**run**;

/\*Bar Graph of Average Alkaline Phosphatase Level by Sex (Trimmed Data)\*/

**proc** **gchart** data=calciumclean2;

title1'Comparison of Alkaline Phosphatase Levels by Sex (Trimmed Data)';

vbar sex / levels=**2** type=mean sumvar=alkphos mean;

**run**;

/\*Descriptive Statistics of Subjects by Lab, Sex (Trimmed Data)\*/

**proc** **means** data=calciumclean2 mean min q1 median q3 max var std;

title1'Descriptive Statistics of Subjects by Lab, Sex (Trimmed Data)';

class lab sex;

var cammol phosmmol alkphos age;

**run**;

/\*Bar Graph of Average Calcium Level by Lab, Sex (Trimmed Data)\*/

**proc** **gchart** data=calciumclean2;

title1'Comparison of Calcium Levels by Lab, Sex (Trimmed Data)';

vbar sex / levels=**2** type=mean sumvar=cammol mean group=lab;

**run**;

/\*Bar Graph of Average Phosphorus Level by Lab, Sex (Trimmed Data)\*/

**proc** **gchart** data=calciumclean2;

title1'Comparison of Phosphorus Levels by Lab, Sex (Trimmed Data)';

vbar sex / levels=**2** type=mean sumvar=phosmmol mean group=lab;

**run**;

/\*Bar Graph of Average Alkaline Phosphatase Level by Lab, Sex (Trimmed Data)\*/

**proc** **gchart** data=calciumclean2;

title1'Comparison of Alkaline Phosphatase Lab, Levels by Sex (Trimmed Data)';

vbar sex / levels=**2** type=mean sumvar=alkphos mean group=lab;

**run**;