ASSIGNMENT 1

Brandon Lampe
STAT 527
Advanced Data Analysis I

September 8, 2014

1 Part I.

1. **Precip:**
Download monthly precipitation and sulphur concentration; data from Univ. of Stockholm. Data type changed from integer to numeric to allow for plotting.

```
library(ggplot2) #load ggplot
d1 <- read.csv("http://statacumen.com/teach/ADA1/ADA1_HW_01_F14-1.csv")

Month <- d1$Month
Precip <- as.numeric(d1$Precip)   # change type to numeric
Sulphur <- as.numeric(d1$Sulphur) # change type to numeric
```

(a) Make stem-and-leaf, histogram, and boxplot for Precip data:

```
stem(Precip, scale = 2)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   1 | 29
##   2 | 35
##   3 | 556
##   4 |
##   5 | 25
##   6 | 39
##   7 |
##   8 | 1
```

```
# histogram of Precip
Precip.hist <- ggplot(d1, aes(x = Precip))
Precip.hist <- Precip.hist + geom_histogram(binwidth = 5)
Precip.hist <- Precip.hist + labs(title = "Monthly Precipitation [mm]")
```

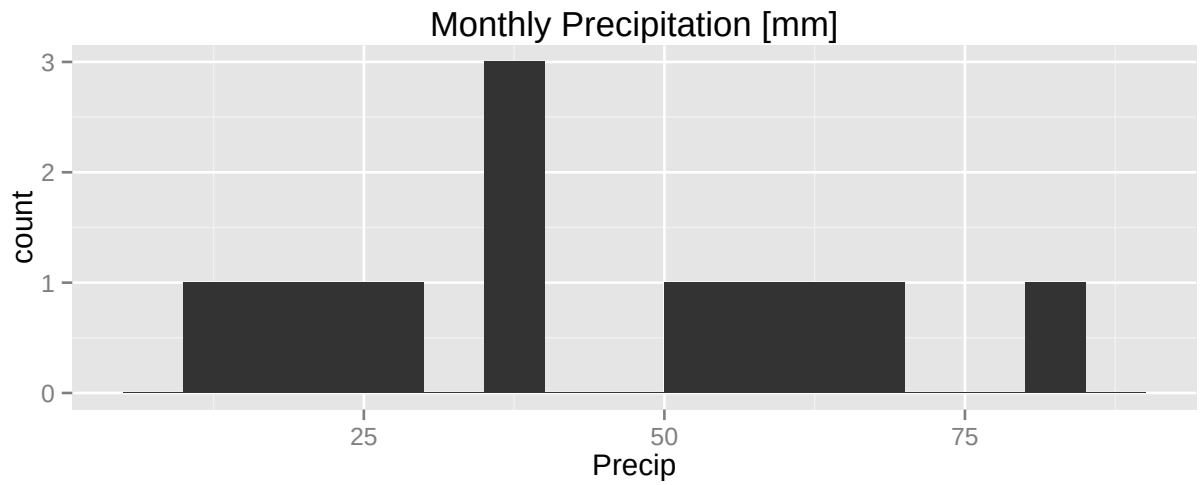

Figure 1: Histogram of Precip Data

```
Precip.box <- ggplot(d1, aes(x = "in", y = Precip)) # boxplot of Precip
Precip.box <- Precip.box + geom_boxplot()
Precip.box <- Precip.box + coord_flip()
Precip.box <- Precip.box + labs(title = "Monthly Precipitation [mm]")
```

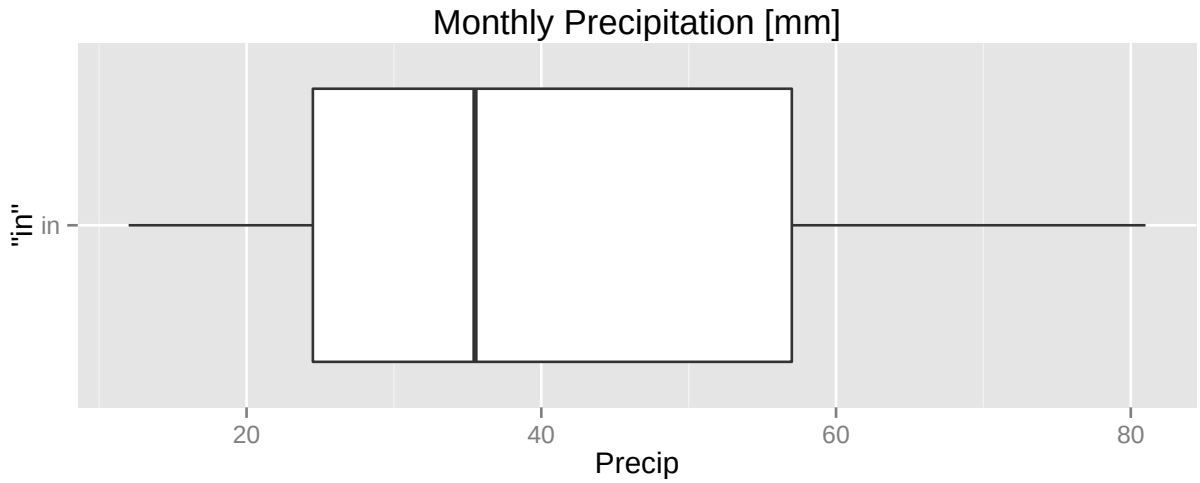

Figure 2: Boxplot of Precip Data

(b) Compute mean, median, standard deviation, and IQR for Precip data:

- mean: 42.08
- median: 35.5
- standard deviation: 21.69
- inter quartile range: 32.5

```
mean(Precip)    # mean of precip
```

```
## [1] 42.08
```

```r
median(Precip)  # median of precip
```

```
## [1] 35.5
```

```r
sd(Precip)      # standard deviation of precip
```

```
## [1] 21.69
```

```r
IQR(Precip)     # inter quartile range (range for middle half of data)
```

```
## [1] 32.5
```

(c) The mean is noticeably larger than the median. The boxplot makes the moment created by the maximum precipitation event evident. The tight grouping around the median except for the maximum value indicates that the mean would be larger than the median, as the mean is sensitive to outliers.

(d) The precipitation data are unimodal and skewed right (to the upper end) ; although no outliers exist. Based on a "Visual" test and an empirical inference of climate, these data are not normally distributed (although normality is difficult to assess with such a small data set). Additionally, no data are present outside of two standard deviations of the mean, whereas approximately four percent of the data should exist in theis region. This is made apparent when a population frequency curve is overlaid on a histogram of the population density, as shown in Figure 3.

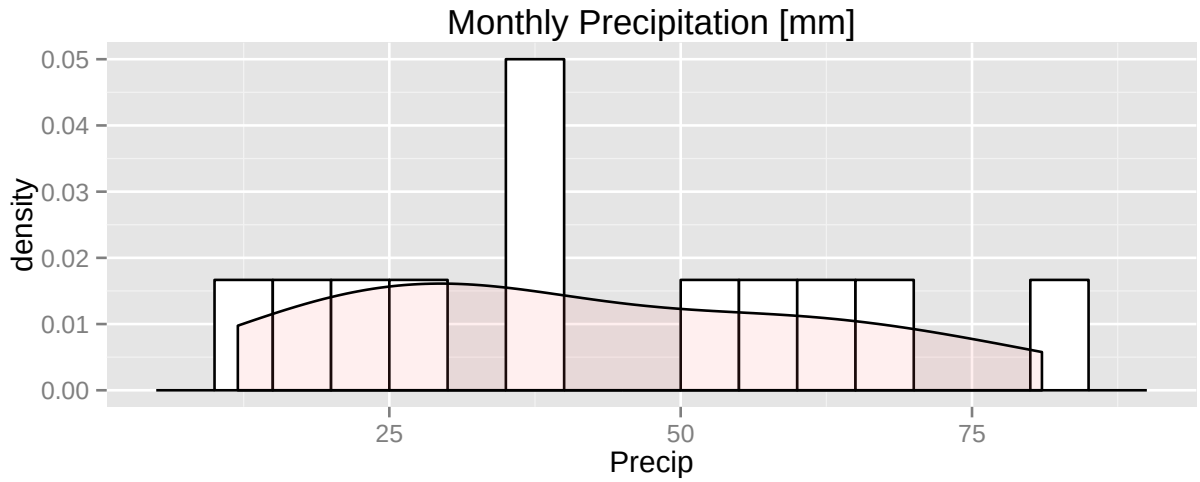

Figure 3: Histogram with population density

2. **Sulphur:**

(a) Make stem-and-leaf, histogram, and boxplot for Sulphur data:

```r
stem(Sulphur, scale = 4)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##    1 | 7
##    2 | 5
```

```
##    3 | 0448
##    4 | 3
##    5 | 5
##    6 | 34
##    7 |
##    8 |
##    9 | 3
##   10 |
##   11 |
##   12 |
##   13 | 5
```

```
# histogram of Sulphur
Sulphur.hist <- ggplot(d1, aes(x = Sulphur))
Sulphur.hist <- Sulphur.hist + geom_histogram(binwidth = 5)
Sulphur.hist <- Sulphur.hist + labs(title = "Sulphur Concentration [mg/m^2]")
```

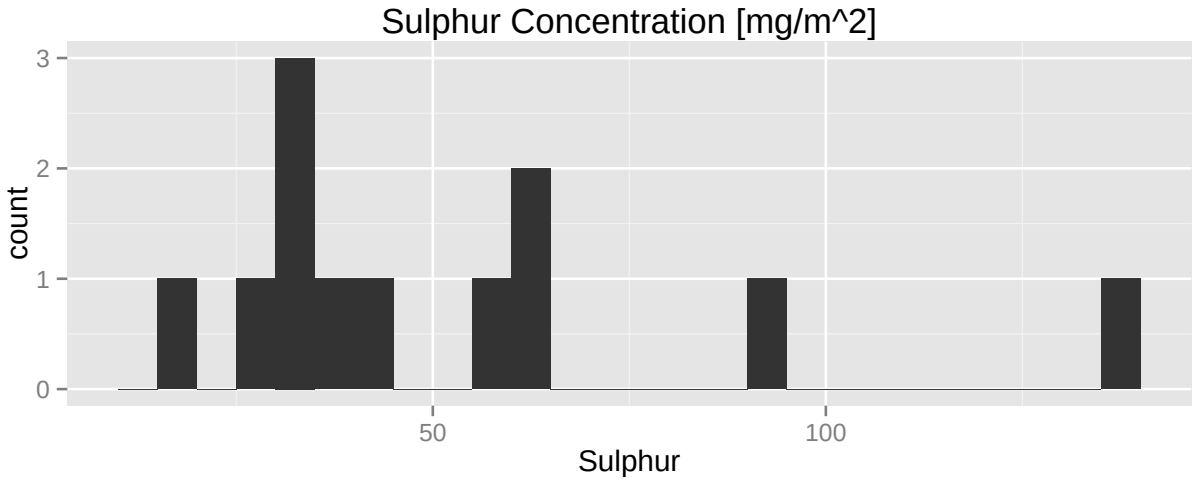

Figure 4: Histogram of Sulphur Data

```
Sulphur.box <- ggplot(d1, aes(x = "in", y = Sulphur)) # boxplot of Sulphur
Sulphur.box <- Sulphur.box + geom_boxplot()
Sulphur.box <- Sulphur.box + coord_flip()
Sulphur.box <- Sulphur.box + labs(title = "Sulphur Concentration [mg/m^2]")
```

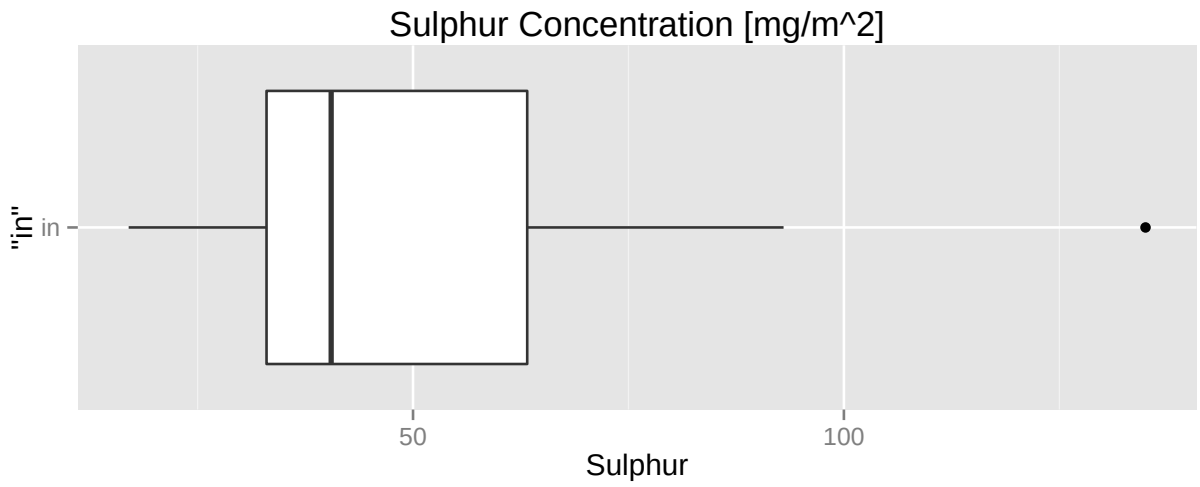

Figure 5: Boxplot of Sulphur Data

(b) Compute mean, median, standard deviation, and IQR for Sulphur data:

- mean: 52.58
- median: 40.5
- standard deviation: 33.31
- inter quartile range: 30.25

```
mean(Sulphur)    # mean of Sulphur
## [1] 52.58
median(Sulphur)  # median of Sulphur
## [1] 40.5
sd(Sulphur)      # standard deviation of Sulphur
## [1] 33.31
IQR(Sulphur)     # inter quartile range (range for middle half of data)
## [1] 30.25
```

(c) The mean again is noticeably larger than the median. An outlier exists in the upper end of the data; therefore, I expect the mean to again be larger than the median, as the mean is sensitive to skewed data.

(d) The sulphur concentration data are bimodal and skewed right (to the upper end) with one outlier. Based on a "Visual" test, these data are not normally distributed (although difficult to asses with limited data). These inferences are made apparent with a population frequency curve overlaid on a histogram of the population density, as shown in Figure 6.

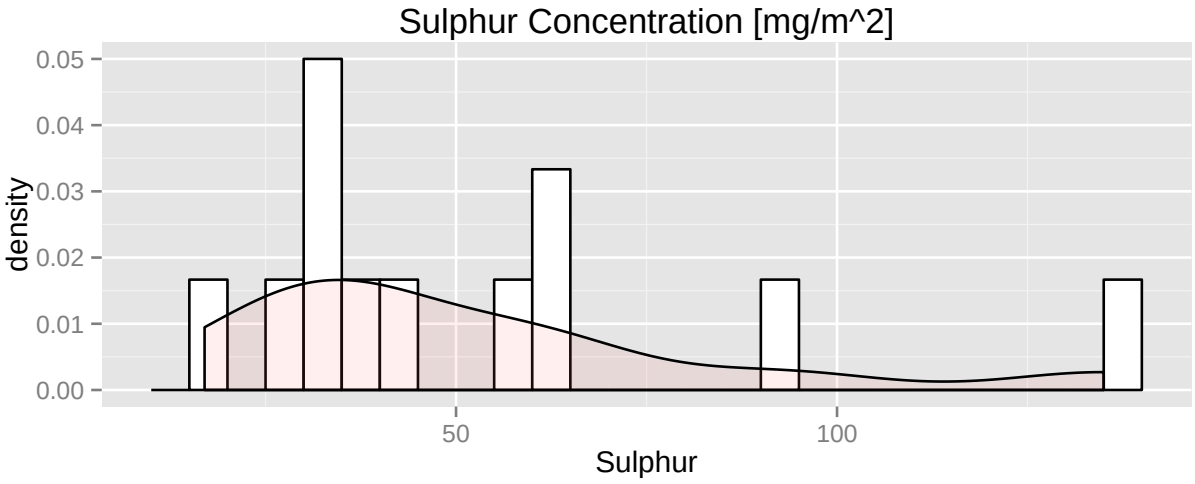

Figure 6: Histogram with population density

3. **Mammals:**

```
mass.df <- read.csv("http://statacumen.com/teach/ADA1/ADA1_HW_01_F14-2.csv")
mass <- as.numeric(mass.df$mass)
```

(a) Make stem-and-leaf, histogram, and boxplot for mass data:

```
stem(mass)
```

```
##
##   The decimal point is 5 digit(s) to the right of the |
##
##   0 | 000000000000000000000000001111124
##   0 | 9
##   1 | 1
##   1 | 7
##   2 |
##   2 |
##   3 |
##   3 |
##   4 |
##   4 | 8
```

```
# histogram of mass
mass.hist <- ggplot(mass.df, aes(x = mass))
mass.hist <- mass.hist + geom_histogram(binwidth = 10000)
mass.hist <- mass.hist + labs(title = "mass [gram]")
```

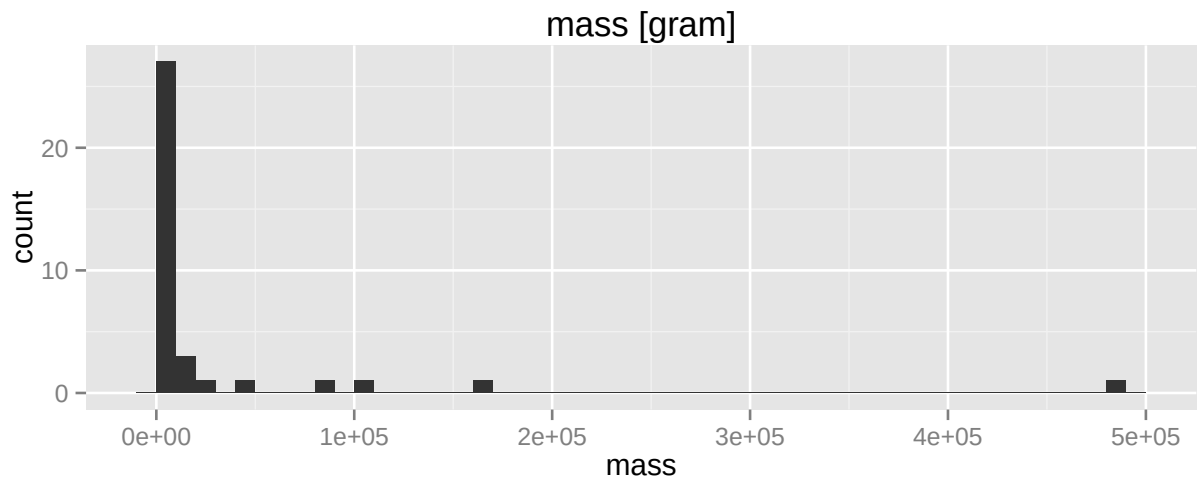

Figure 7: Histogram of mass Data

```
mass.box <- ggplot(mass.df, aes(x = "Mammals", y = mass)) # boxplot of mass
mass.box <- mass.box + geom_boxplot()
mass.box <- mass.box + coord_flip()
mass.box <- mass.box + labs(title = "mass [gram]")
```

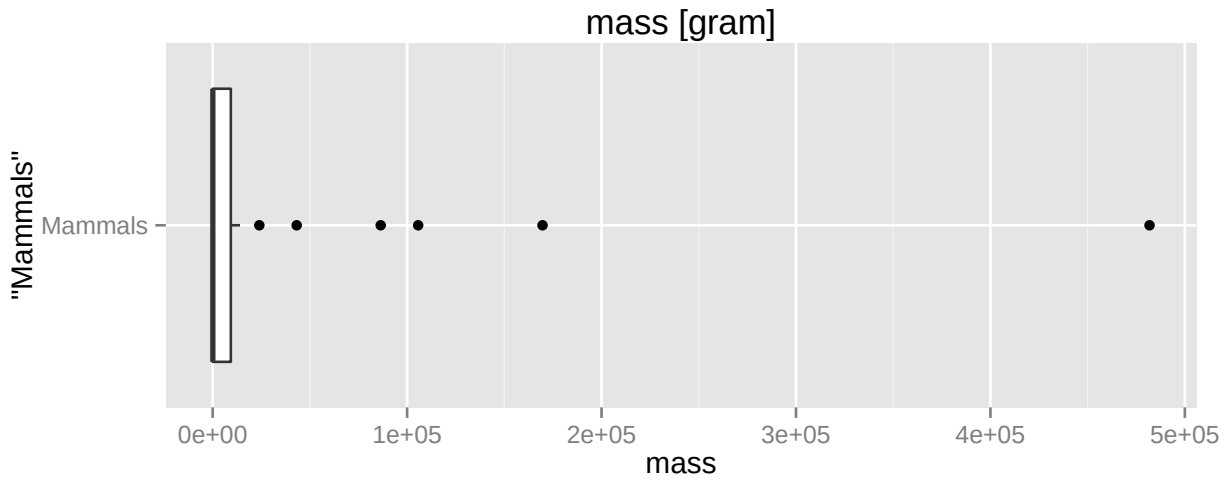

Figure 8: Boxplot of mass Data

(b) Compute mean, median, standard deviation, and IQR for mass data:

- mean: 27,081
- median: 92.85
- standard deviation: 85,564
- inter quartile range: 9,317

```
mean(mass)    # mean of mass

## [1] 27081
```

```
median(mass)  # median of mass
```

```
## [1] 92.85
```

```
sd(mass)      # standard deviation of mass
```

```
## [1] 85564
```

```
IQR(mass)     # inter quartile range (range for middle half of data)
```

```
## [1] 9317
```

(c) The mean again is noticeably way larger than the median. Multiple outliers exist in the upper end of the data; therefore, I expect the mean to again be larger than the median, as the mean is sensitive to skewed data.

(d) The mass data are multimodal and skewed right (to the upper end) with numerous outliers. Based on a "Visual" test, these data are not normally distributed. Also, the population density is much too low in the IQR to be considered normally distributed. These inferences are made apparent with a population frequency curve overlaid on a histogram of the population density, as shown in Figure 9.

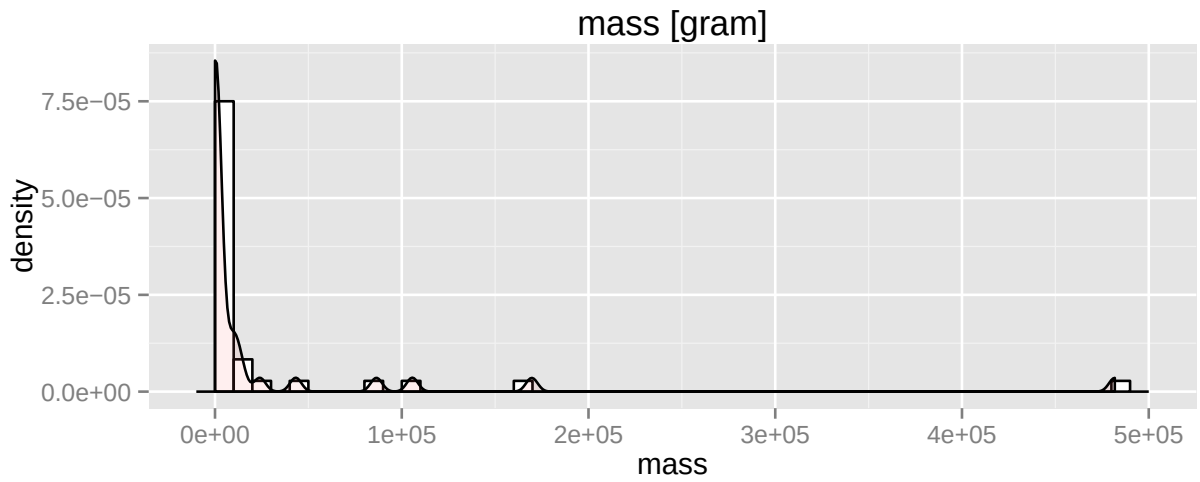

Figure 9: Histogram with population density

4. **log(Mammals):**

```
logmass.df <- data.frame(log(mass.df$mass))
colnames(logmass.df) <- "logmass"
logmass <- log(mass)
```

(a) Make stem-and-leaf, histogram, and boxplot for log(mass) data:

```
stem(logmass)
```

```
##
##   The decimal point is at the |
```

```
##
##    0 | 34477
##    2 | 161123335888
##    4 | 473
##    6 | 7139
##    8 | 001246
##   10 | 1746
##   12 | 01
```

```
# histogram of logmass
logmass.hist <- ggplot(logmass.df, aes(x = logmass))
logmass.hist <- logmass.hist + geom_histogram(binwidth = 1)
logmass.hist <- logmass.hist + labs(title = "logmass [gram]")
```

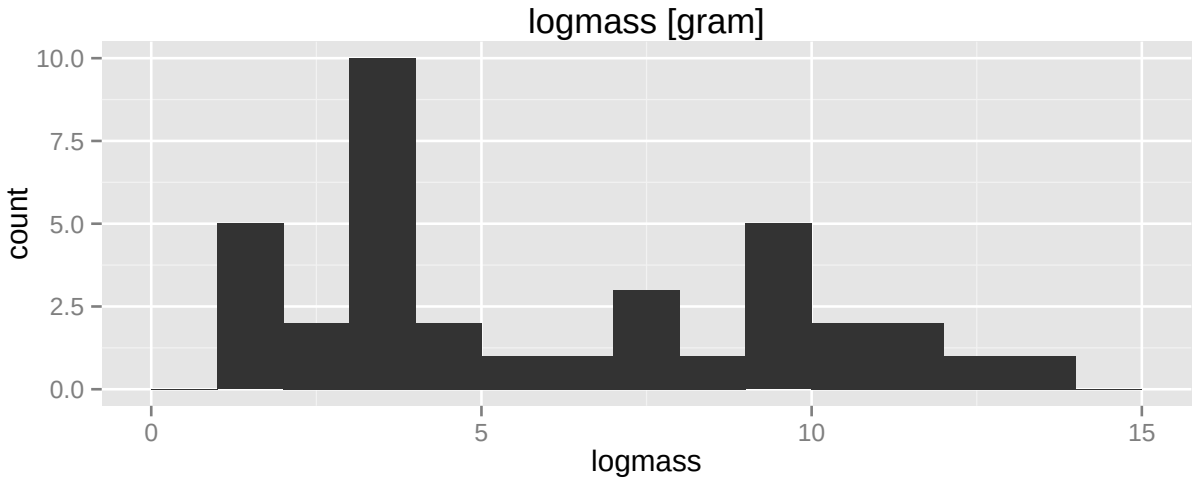

Figure 10: Histogram of log(mass) Data

```
logmass.box <- ggplot(logmass.df, aes(x = "Mammals", y = logmass)) # boxplot of logmass
logmass.box <- logmass.box + geom_boxplot()
logmass.box <- logmass.box + coord_flip()
logmass.box <- logmass.box + labs(title = "logmass [gram]")
```

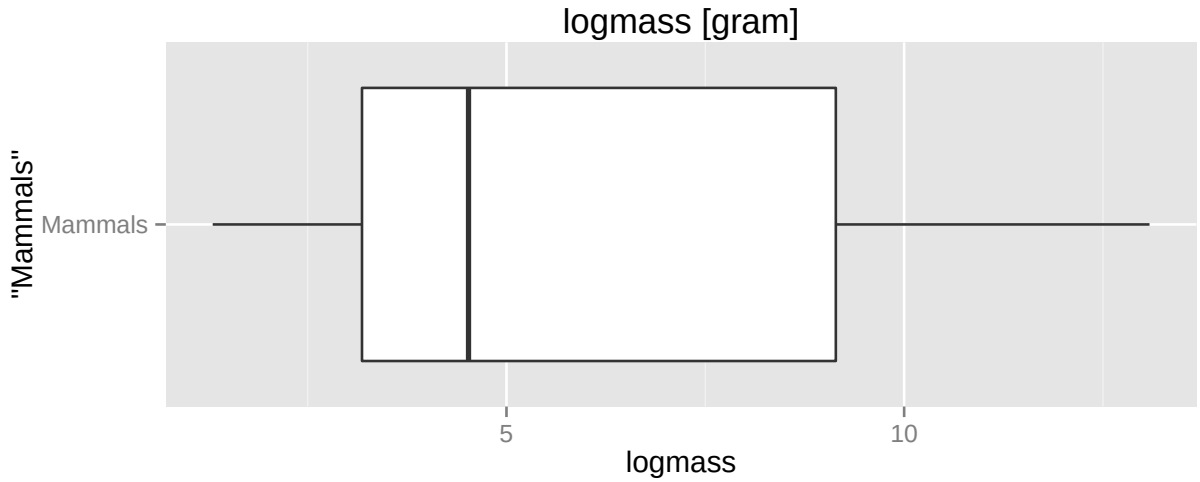

Figure 11: Boxplot of logmass Data

(b) Compute mean, median, standard deviation, and IQR for log(mass) data:

- mean: 5.918
- median: 4.522
- standard deviation: 3.583
- inter quartile range: 5.958

```
mean(logmass)    # mean of logmass
## [1] 5.918
median(logmass)  # median of logmass
## [1] 4.522
sd(logmass)      # standard deviation of logmass
## [1] 3.583
IQR(logmass)     # inter quartile range (range for middle half of data)
## [1] 5.958
```

(c) The mean again is noticeably larger than the median. No exist in the data; therefore. I expect the mean to again be larger than the median, as the mean is sensitive to skewed data and these data are skewed to the right.

(d) The log(mass) data are bimodal and skewed right (to the upper end) with no outliers (a value would need to be extremely large to be considered an outlier in log space e.g., e^{20}). Based on a "Visual" test, these data are not normally distributed, as they are bimodal. These inferences are made apparent with a population frequency curve overlaid on a histogram of the population density, as shown in Figure 12.

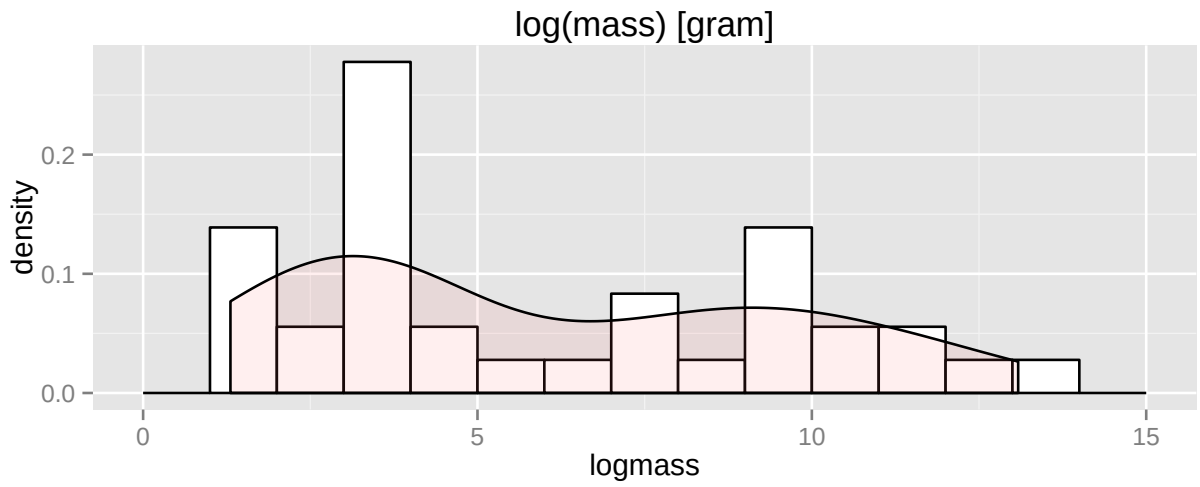

Figure 12: Histogram with population density