Advanced Data Analysis 1, Stat 427/527 Homework 1 Name: Answer Key
Fall 2014 Prof. Erik B. Erhardt

Part I. (60 points) Do all calculations in R. All R code for the assignment should be included with the part of the problem it addresses (for code and output use a fixed-width font, such as Courier). Code is used to calculate result. Text is used to report and interpret results. Do not report or interpret results in the code.

(15pts) **1. Precip:** The Dept of Meteorology at the University of Stockholm monitors chemical constituents of the atmosphere at several stations throughout Sweden. The chemicals are precipitated out of the atmosphere by rain and deposited on filters, from which the amount of chemical, in milligrams per square meter of filter surface, can be measured. The monthly sulphur (Sulphur) in mg/m^2 for each of the 12 months (Month) (1=Jan, 2=Feb, etc) and the monthly precipitation (Precip) in mm is given in the table below for one station.

Month	Precip	Sulphur
1	35	55
2	25	30
3	12	25
4	36	43
5	81	135
6	19	38
7	55	63
8	63	93
9	69	64
10	23	17
11	52	34
12	35	34

Read the data from the website with:

```
d1 <- read.csv("http://statacumen.com/teach/ADA1/ADA1_HW_01_F14-1.csv")
```

Refer to columns in the data using the "$" sign. `d1` is the dataframe, and `Precip` is a variable in the dataframe.

```
# Precip column
d1$Precip
# Sulphur column
d1$Sulphur
```

(a) (3 pts) Make a stem-and-leaf display, histogram, and boxplot for the Precip data.

Solution:

```
stem(d1$Precip, scale = 2)
```

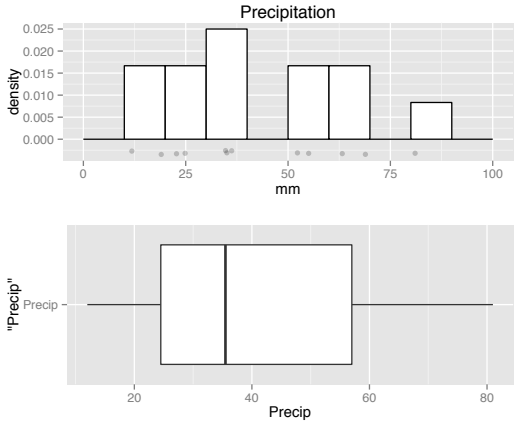
```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   1 | 29
##   2 | 35
##   3 | 556
##   4 |
##   5 | 25
##   6 | 39
##   7 |
##   8 | 1
```

```
library(ggplot2)
# Histogram overlaid with kernel density curve
p1 <- ggplot(d1, aes(x = Precip))
  # Histogram with density instead of count on y-axis
```

```
p1 <- p1 + geom_histogram(aes(y=..density..)
                          , binwidth=10
                          , colour="black", fill="white")
p1 <- p1 + geom_point(aes(y = -0.003)
                      , position = position_jitter(height = 0.0005)
                      , alpha = 1/5)
p1 <- p1 + labs(title = "Precipitation") + xlab("mm")

# boxplot
p2 <- ggplot(d1, aes(x = "Precip", y = Precip))
p2 <- p2 + geom_boxplot()
p2 <- p2 + coord_flip()
#p2

library(gridExtra)
## Loading required package:  grid
grid.arrange(p1, p2, ncol=1)
```



(b) (2 pts) Compute the mean, median, standard deviation, and interquartile range for the Precip data.

Solution:

```
mean(d1$Precip)
## [1] 42.08
median(d1$Precip)
## [1] 35.5
sd(d1$Precip)
## [1] 21.69
diff(fivenum(d1$Precip)[c(2,4)])
## [1] 35
```

(c) (4 pts) Is there much difference between the mean and median? Discuss, briefly, whether the size and the direction of the difference is sensible, given the graphical summaries.

Solution: $\bar{Y} - M = 42.1 - 35.5 = 6.6$ is a small difference relative to the variability of the sample.

(d) (6 pts) Using the graphical summaries, describe the shape of the Precip distribution. Discuss modality, presence/absence of outliers, whether skewness is present, and if so, in what direction, and whether it would be reasonable to assume that the Precip distribution is normal.

Solution: Distribution is unimodal, symmetric, no outliers, approximately normal.

(15pts) **2. Sulphur:** Repeat with the Sulphur data in the previous problem.

Solution: (a)

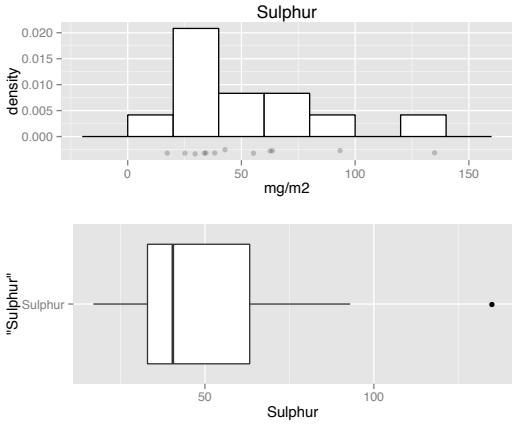
```
stem(d1$Sulphur, scale = 2)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##    0 | 7
##    2 | 50448
##    4 | 35
##    6 | 34
##    8 | 3
##   10 |
##   12 | 5
```

```
# Histogram overlaid with kernel density curve
p1 <- ggplot(d1, aes(x = Sulphur))
  # Histogram with density instead of count on y-axis
p1 <- p1 + geom_histogram(aes(y=..density..)
                          , binwidth=20
                          , colour="black", fill="white")
p1 <- p1 + geom_point(aes(y = -0.003)
                      , position = position_jitter(height = 0.0005)
                      , alpha = 1/5)
p1 <- p1 + labs(title = "Sulphur") + xlab("mg/m2")

# boxplot
p2 <- ggplot(d1, aes(x = "Sulphur", y = Sulphur))
p2 <- p2 + geom_boxplot()
p2 <- p2 + coord_flip()
#p2

library(gridExtra)
grid.arrange(p1, p2, ncol=1)
```



Solution: (b)

```
mean(d1$Sulphur)
```

```
## [1] 52.58
```

```
median(d1$Sulphur)
```

```
## [1] 40.5
```

```
sd(d1$Sulphur)
```

```
## [1] 33.31
```

```
diff(fivenum(d1$Sulphur)[c(2,4)])
```

```
## [1] 31.5
```

Solution: (c) $\bar{Y} - M = 52.6 - 40.5 = 12.1$ is fairly large relative to the variability of the sample, consistent with the skewness observed in the plots.

Solution: (d) Distribution is unimodal, skewed right, one outlier based on the boxplot fences, not normal.

(15 pts) **3. Mammals:** These data, from Holling (1982), are the body masses (mass) in grams for 36 species of boreal forest mammals found east of the Manitoba-Ontario border in pure or mixed conifer strands. Repeat the steps in Problem 1 in Part I.

mass(g)	
3.69	104.90
4.11	191.36
4.25	839.15
5.24	1224.70
5.50	1496.85
8.36	2642.17
13.18	3118.45
22.45	8504.85
22.96	9071.84
24.52	10149.12
26.93	12303.68
28.30	14061.35
28.35	23995.02
33.00	43204.64
42.53	86416.53
43.94	105686.93
44.51	169643.40
80.80	481828.10

Solution: (a)

```
d2 <- read.csv("http://statacumen.com/teach/ADA1/ADA1_HW_01_F14-2.csv")
stem(d2$mass, scale = 1)
```

```
##
##   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
```

```
stem(d2$mass[(d2$mass<9000)], scale = 1)
```

```
##
##   The decimal point is 3 digit(s) to the right of the |
##
##   0 | 00000000000000001128
##   1 | 25
##   2 | 6
##   3 | 1
##   4 |
##   5 |
##   6 |
```

```
##   7 |
##   8 | 5
```

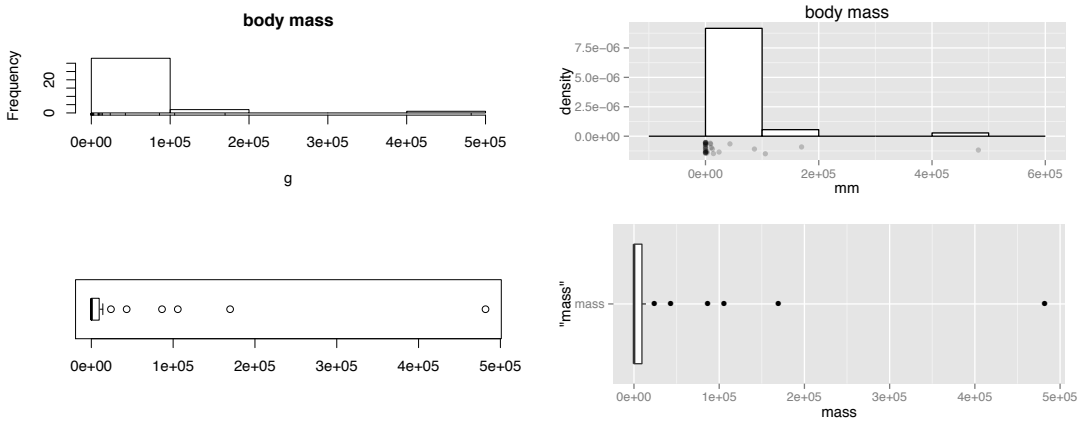
```
par(mfrow=c(2,1))
# Histogram overlaid with kernel density curve
hist(d2$mass, breaks = 5, main = "body mass", xlab = "g")
rug(d2$mass)

# boxplot
boxplot(d2$mass, horizontal=TRUE)

library(ggplot2)
# Histogram overlaid with kernel density curve
p1 <- ggplot(d2, aes(x = mass))
  # Histogram with density instead of count on y-axis
p1 <- p1 + geom_histogram(aes(y=..density..)
                          , binwidth=100000
                          , colour="black", fill="white")
p1 <- p1 + geom_point(aes(y = -0.000001)
                      , position = position_jitter(height = 0.0000005)
                      , alpha = 1/5)
p1 <- p1 + labs(title = "body mass") + xlab("mm")

# boxplot
p2 <- ggplot(d2, aes(x = "mass", y = mass))
p2 <- p2 + geom_boxplot()
p2 <- p2 + coord_flip()
#p2

library(gridExtra)
grid.arrange(p1, p2, ncol=1)
```



Solution: (b)

```
mean(d2$mass)
```

```
## [1] 27081
```

```
median(d2$mass)
```

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

```
sd(d2$mass)
```

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

```
diff(fivenum(d2$mass)[c(2,4)])
```

```
## [1] 9587
```

Solution: (c) $\bar{Y} - M = 2.7081 \times 10^4 - 92.8 = 2.6989 \times 10^4$ is substantial. The mean is expected to be much larger here because of the extreme right skewness observed in all of the graphical summaries.

Solution: (d) Distribution is unimodal, extremely skewed right, many outliers, not normal.

(15$^{\text{pts}}$) **4. log(Mammals):** Repeat with the natural logarithm of the data in the previous problem. You can create a new variable with something like: `logmass <-log(mass)`.

Solution: The natural log pulled in the extreme observations and substantially reduced the skewness.
(a)

```
d2$logmass <- log(d2$mass)
stem(d2$logmass, scale = 1)
```

```
##
##   The decimal point is at the |
##
##    0 | 34477
##    2 | 161123335888
##    4 | 473
##    6 | 7139
##    8 | 001246
##   10 | 1746
##   12 | 01
```

Solution: (b)

```
mean(d2$logmass)
```

```
## [1] 5.918
```

```
median(d2$logmass)
```

```
## [1] 4.522
```

```
sd(d2$logmass)
```

```
## [1] 3.583
```

```
diff(fivenum(d2$logmass)[c(2,4)])
```

```
## [1] 6.002
```

Solution: (c) $\bar{Y} - M = 5.9 - 4.5 = 1.4$ is relatively small. The distribution is slightly skewed to the right, so we expect the mean to be slightly larger than the median, so the direction of the difference between these two measures of location is as expected.

Solution: (d) Distribution is bi/tri/quad-modal, slightly skewed right, no outliers based on the boxplot fences. The multiple modes (more so than the skewness which is not extreme) make it unreasonable to assume that the log transformed body mass distribution is normal.