Homework #1 Solutions

Derek Sonderegger

First we'll load all the libraries we'll need for this homework.

```
library(ggplot2)
library(dplyr)
```

1. O&L 3.21. The ratio of DDE (related to DDT) to PCB concentrations in bird eggs has been shown to have had a number of biological implications. The ratio is used as an indication of the movement of contamination through the food chain. The paper "The ratio of DDE to PCB concentrations in Great Lakes herring gull eggs and its us in interpreting contaminants data" reports the following ratios for eggs collected at 13 study sites from the five Great Lakes. The eggs were collected from both terrestrial and aquatic feeding birds.

Source Type	DDE to PCB Ratio
Terrestrial	76.50, 6.03, 3.51, 9.96, 4.24, 7.74, 9.54, 41.70, 1.84, 2.5, 1.54
Aquatic	0.27, 0.61, 0.54, 0.14, 0.63, 0.23, 0.56, 0.48, 0.16, 0.18

   a) By hand, compute the mean and median separately for each type of feeder.

   The sorted Terrestrial observations are: `1.54 1.84 2.50 3.51 4.24 6.03 7.74 9.54 9.96 41.70 76.50` *and the middle (median) observation is 6.03. The sum of the Terrestrial observations is 165.1 and therefore the mean is* $165.1/11 = 15.01$.

   The sorted Aquatic observations are: `0.14 0.16 0.18 0.23 0.27 0.48 0.54 0.56 0.61 0.63` *and the average of the middle most observations (0.27 and 0.48) is the median 0.375. The sum of the Aquatic observations is* 3.8 *and therefore the mean is* $3.8/10 = 0.38$.

   b) Using your results from parts (a) and (b), comment on the relative sensitivity of the mean and median to extreme values in a data set.

   The mean is more sensitive to outliers than the median.

   c) Which measure, mean or median, would you recommend as the most appropriate measure of the DDE to PCB level for both types of feeders? Explain your answer.

   For the aquatic observations, the numbers are so similar it doesn't really matter, but for the terrestrial it certain will matter and we should use the median because of the strong influence of the 76.50 observation.

2. O&L 3.31. Consumer Reports in its June 1998 issue reports on the typical daily room rate at six luxury and nine budget hotels. The room rates are give in the following table.

Hotel Type	Nightly Rate
Luxury	\$175, \$180, \$120, \$150, \$120, \$125
Budget	\$50, \$50, \$49, \$45, \$36, \$45, \$50, \$50, \$40

   a) By hand, compute the means and standard deviations of the room rates for each class of hotel.

   Using similar calculates as in problem 1, it isn't hard to see that the mean for the Luxury hotels is $\bar{x}_L = 145$

x_i	$x_i - \bar{x}_L$	$(x_i - \bar{x})^2_L$
175	30	900
180	35	1225
120	-25	625
150	5	25
120	-25	625
125	-20	400
—		
		3800

so $s^2_L = \frac{1}{6-1}\sum(x_i - \bar{x}_{iL})^2 = \frac{1}{5}3800 = 760$ *and therefore* $s_L = \sqrt{(760)} = 27.57$.

For the Budget hotels is $\bar{x}_B = 46.11$.

x_i	$x_i - \bar{x}_B$	$(x_i - \bar{x}_B)^2$
50	3.889	15.123
50	3.889	15.123
49	2.889	8.345
45	-1.111	1.234
36	-10.111	102.234
45	-1.111	1.234
50	3.889	15.123
50	3.889	15.123
40	-6.111	37.345
—		
		210.889

so $s^2_B = \frac{1}{9-1}\sum(x_i - \bar{x}_B)^2 = \frac{1}{8}210.889 = 26.361$ *and therefore* $s_B = \sqrt{(26.361)} = 5.134$.

b) Give a reason why luxury hotels might have higher variability than the budget hotels.

Because luxury hotels are competing on a number of different areas, location, view, ammenities, etc, the competition on price is perhaps less. In contrast the budget hotesl are only competing on price, so you'd expect them to have prices that are much tighter.

3. Use R to confirm your calculations in problem 1 (the pollution data). *Show the code you used and the subsequent output. It will often be convenient for me to give you code that generates a data frame instead of uploading an Excel file and having you read it in. The data can be generated using the following commands:*

```
PollutionRatios <- data.frame(
  Ratio = c(76.50, 6.03, 3.51, 9.96, 4.24, 7.74, 9.54, 41.70, 1.84, 2.5, 1.54,
            0.27, 0.61, 0.54, 0.14, 0.63, 0.23, 0.56,  0.48, 0.16, 0.18       ),
  Type  = c( rep('Terrestrial',11), rep('Aquatic',10) ) )
# Print out some of the data to confirm what the column names are
head( PollutionRatios )
```

```
##   Ratio        Type
## 1 76.50 Terrestrial
## 2  6.03 Terrestrial
## 3  3.51 Terrestrial
## 4  9.96 Terrestrial
## 5  4.24 Terrestrial
## 6  7.74 Terrestrial
```

Hint: for computing the means and medians for each type of feeder separately, there is a very convenient command `group_by()`

```
PollutionRatios %>%
  summarise( mean(Ratio), median(Ratio))  # without regard for Type
```

```
##   mean(Ratio) median(Ratio)
## 1    8.042857          1.54
```

```
PollutionRatios %>% group_by(Type) %>%
  summarise( mean(Ratio), median(Ratio))
```

```
## # A tibble: 2 × 3
##          Type `mean(Ratio)` `median(Ratio)`
##        <fctr>         <dbl>           <dbl>
## 1     Aquatic       0.38000           0.375
## 2 Terrestrial      15.00909           6.030
```

4. Use R to confirm your calculations in problem 2 (the hotel data). Show the code you used and the subsequent output. The data can be loaded into a data frame using the following commands Show the code you used and the subsequent output:

```
Hotels <- data.frame(
  Price = c(175, 180, 120, 150, 120, 125, 50, 50, 49, 45, 36, 45, 50, 50, 40),
  Type  = c( rep('Luxury',6),  rep('Budget', 9) ) )

# Print out some of the data to confirm what the column names are
head( Hotels )
```

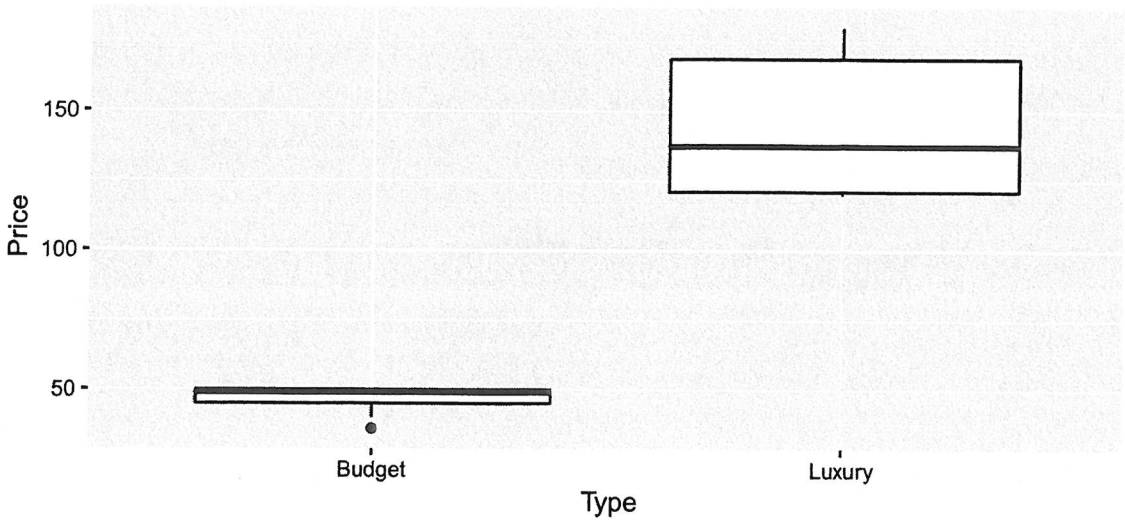
```
##   Price   Type
## 1   175 Luxury
## 2   180 Luxury
## 3   120 Luxury
## 4   150 Luxury
## 5   120 Luxury
## 6   125 Luxury
```

```
library(dplyr)
Hotels %>% group_by(Type) %>%
  summarise(mean = mean(Price),
            s2   = var(Price),
            s    = sd(Price))
```

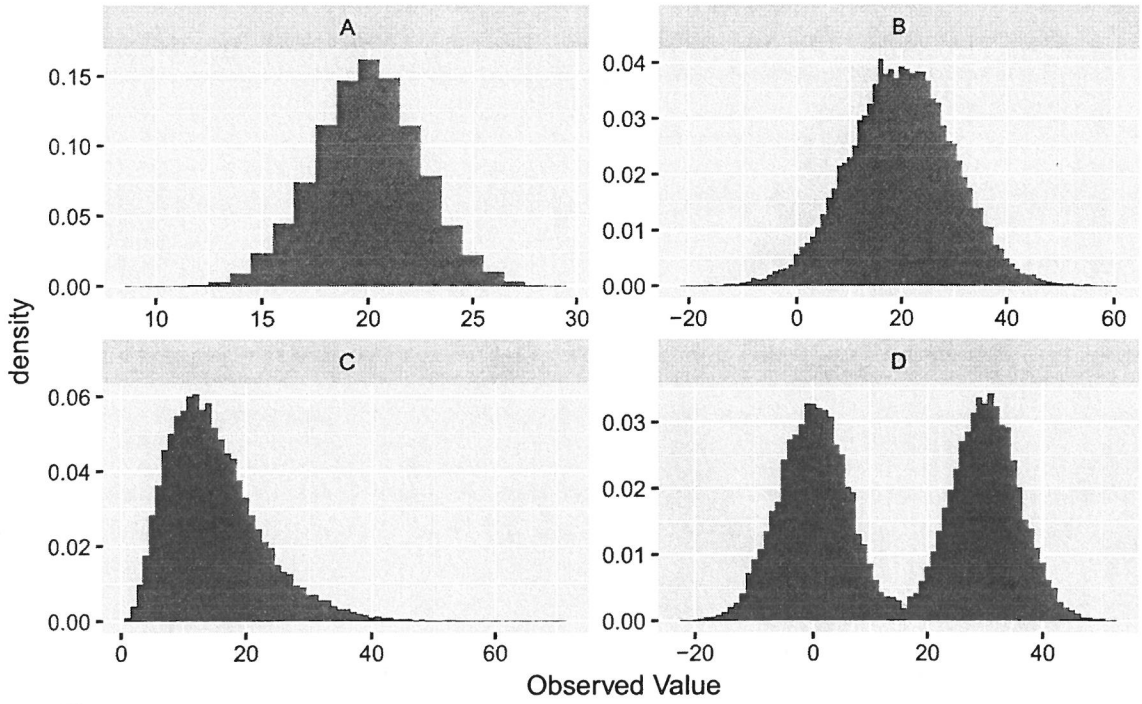
```
## # A tibble: 2 × 4
##     Type      mean        s2         s
##   <fctr>     <dbl>     <dbl>     <dbl>
## 1 Budget  46.11111  26.36111  5.134307
## 2 Luxury 145.00000 760.00000 27.568098
```

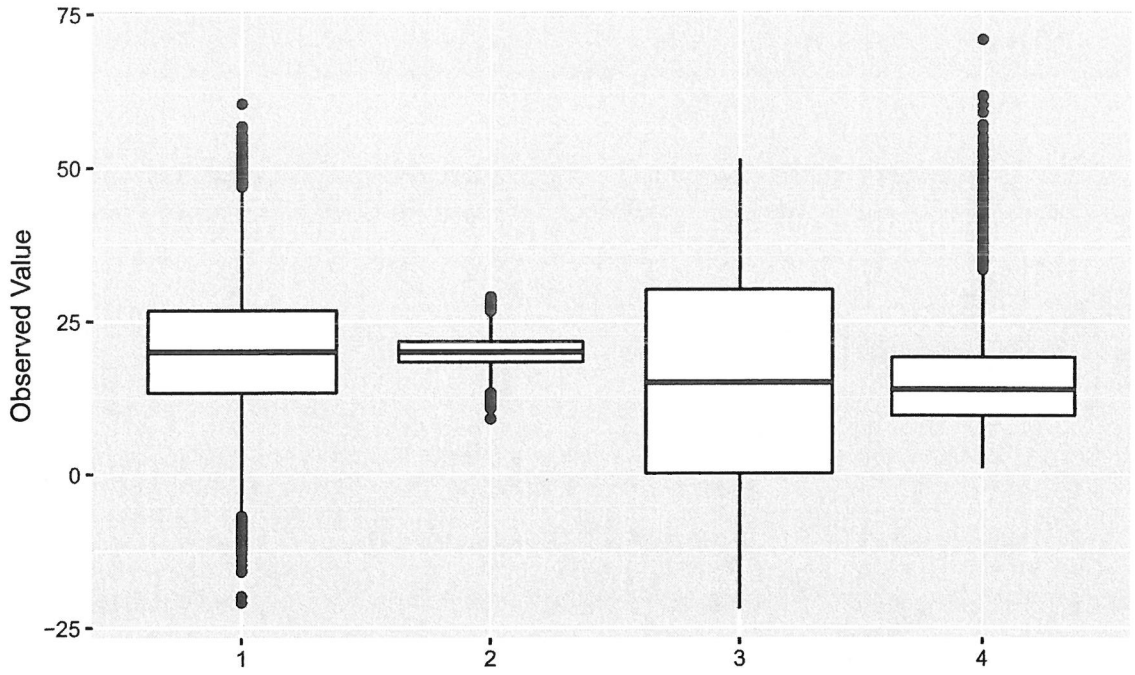
5. For the hotel data, create side-by-side box-and-whisker plots to compare the prices.

```
ggplot(Hotels, aes(x=Type, y=Price)) + geom_boxplot()
```


150
100
50
Price
Budget
Luxury
Type

6. Match the following histograms to the appropriate boxplot.





a) Histogram A goes with Boxplot _____2_____

b) Histogram B goes with Boxplot _____1_____

c) Histogram C goes with Boxplot _____4_____

d) Histogram D goes with Boxplot _____3_____

7. Twenty-five employees of a corporation have a mean salary of \$62,000 and the sample standard deviation

of those salaries is \$15,000. If each employee receives a bonus of \$1,000, does the standard deviation of the salaries change? Explain your reasoning.

No. The entire distribution of salaries will be shifted up by \$1,000 so the mean salary will change, but the SPREAD of the distribution will remain the same. So the standard deviation (a measure of spread) will be unchanged.