Carson Slater

STA 5380: Intro to Statistical Methods I

Homework #0

August 29 2023

```
cancer <- read.table(file = "cancer.txt", header = T)
summary(cancer)
#    Country              colon            meat
#  Length:23          Min.   : 6.00   Min.   : 30.0
#  Class :character   1st Qu.: 9.00   1st Qu.: 93.5
#  Mode  :character   Median :11.00   Median :140.0
#                     Mean   :13.26   Mean   :142.4
#                     3rd Qu.:12.00   3rd Qu.:176.5
#                     Max.   :40.00   Max.   :315.0
```

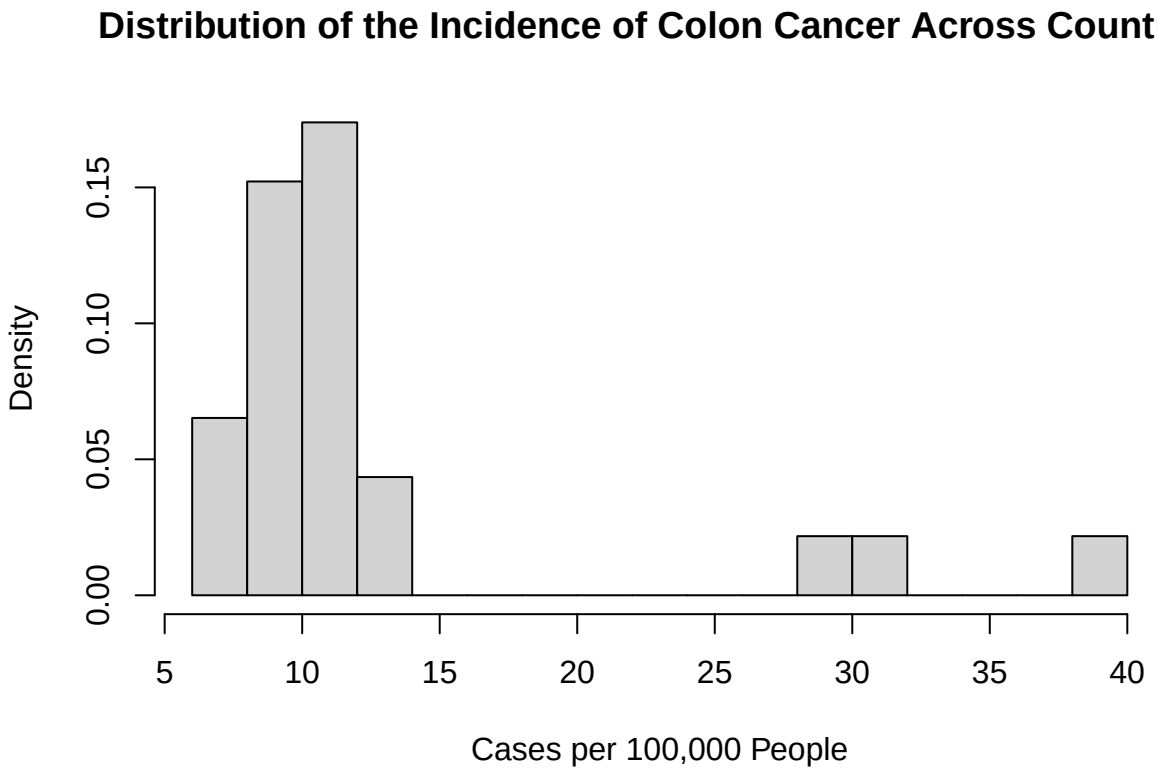
```
# Printing summary statistics of the data frame
summary(cancer)
#    Country              colon            meat
#  Length:23          Min.   : 6.00   Min.   : 30.0
#  Class :character   1st Qu.: 9.00   1st Qu.: 93.5
#  Mode  :character   Median :11.00   Median :140.0
#                     Mean   :13.26   Mean   :142.4
#                     3rd Qu.:12.00   3rd Qu.:176.5
#                     Max.   :40.00   Max.   :315.0
sd(cancer$colon)
# [1] 8.3094
sd(cancer$meat)
# [1] 71.71904
```

1. (a.) There are 23 observations of 3 variables in the data frame `cancer`. (b.) For average meat consumption, the sample $\mu = 142.4$ and sample $\sigma \approx 71.71$. As for colon cancer cases per 100,000 people, the sample $\mu = 13.26$ and the sample sigma is $\sigma \approx 8.31$.

2. (a.)

```
# Creating a histogram of the incidence of colon cancer
hist(cancer$colon,
     main = "Distribution of the Incidence of Colon Cancer Across Countries",
     xlab = "Cases per 100,000 People",
     breaks = 20,
```

```
    freq = FALSE
    )
```



(b.) The histogram exhibits the highest density of countries having around ten to twelve cases of colon cancer per 100,000 people, with three outliers. Here they can be found with the `which()` command.

```
cancer$Country[which(colon >= 25)]
# Error in which(colon >= 25): object 'colon' not found
```

3.

```
library(ISwR)
data(malaria)
str(malaria)
# 'data.frame': 100 obs. of  4 variables:
#  $ subject: int  1 2 3 4 5 6 7 8 9 10 ...
#  $ age    : int  15 14 12 15 14 12 12 13 13 14 ...
#  $ ab     : int  546 268 284 38 827 252 24 1740 76 83 ...
#  $ mal    : int  0 0 0 0 0 0 1 0 0 0 ...
```

(a.) They data set has 100 observations of four variables: `subject`, `age`, `ab`, and `mal`.

```r
sum(malaria$mal != 1)
# [1] 73
```

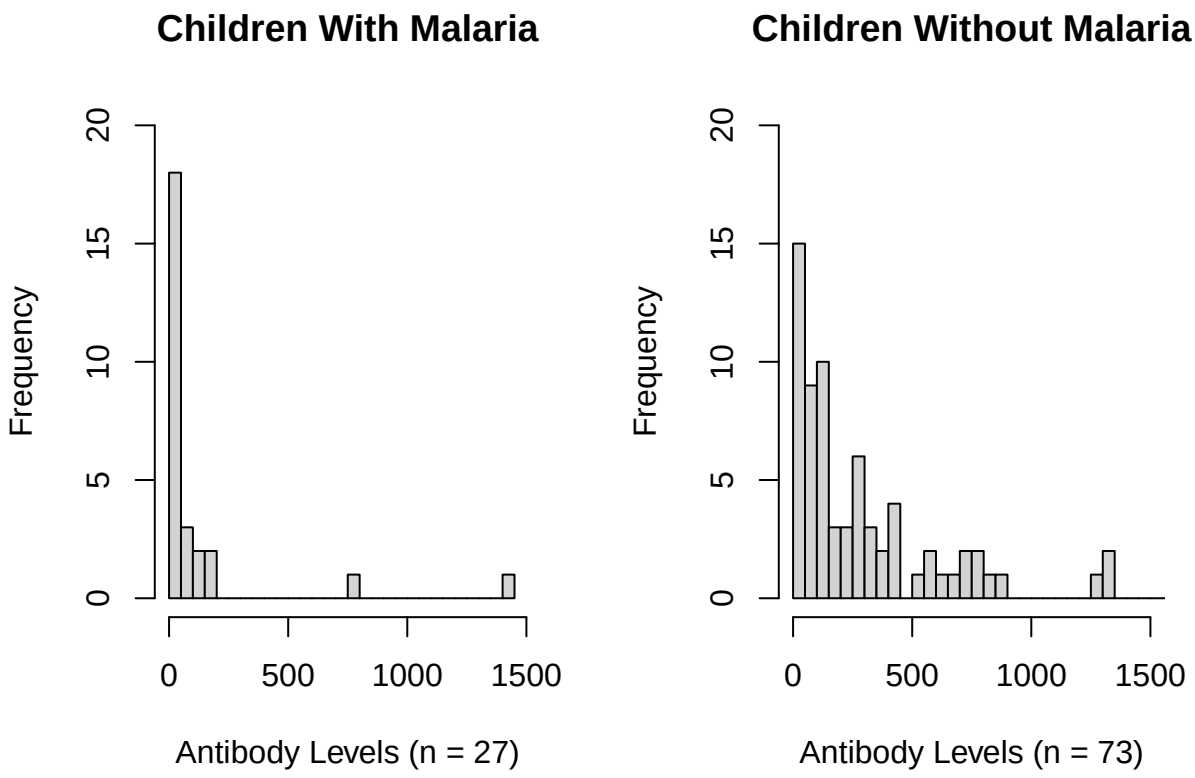
(b.) There are 73 children that do not have malaria after the 8 month period.

```r
# For children with malaria symptoms

mean(malaria$ab[which(malaria$mal == 1)])
# [1] 126.1111
sd(malaria$ab[which(malaria$mal == 1)])
# [1] 296.5646

# For children without malaria symptoms

mean(malaria$ab[which(malaria$mal != 1)])
# [1] 380.0411
sd(malaria$ab[which(malaria$mal != 1)])
# [1] 483.9842
```

(c.) For children with malaria symptoms the sample $\mu \approx 126.11$ and sample $\sigma \approx 296.56$. As for children without malaria symptoms, the sample $\mu \approx 380.04$ and the sample sigma is $\sigma \approx 483.98$.

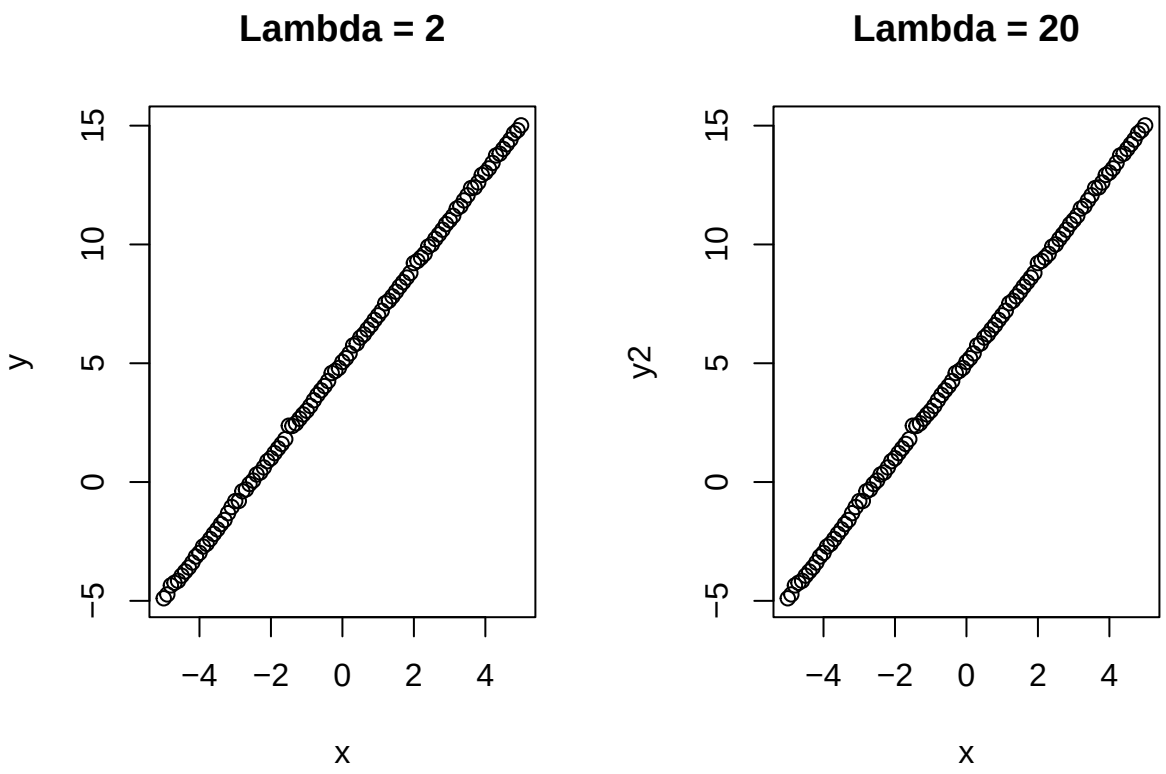
(d.)

```r
par(mfrow = c(1,2))
hist(malaria$ab[which(malaria$mal == 1)],
     main = "Children With Malaria",
     xlab = "Antibody Levels (n = 27)",
     xlim = c(0, 1500),
     ylim = c(0,20),
     breaks = 50)
hist(malaria$ab[which(malaria$mal != 1)],
     main = "Children Without Malaria",
     xlab = "Antibody Levels (n = 73)",
     xlim = c(0, 1500),
     ylim = c(0,20),
     breaks = 50)
```



4.

```r
set.seed(78)
# Creating the first data frame
x <- seq(-5, 5, length = 101)
y <- 5 + 2*x + rexp(n = 101, rate = 2)

#Creating the second data frame
y2 <- y <- 5 + 2*x + rexp(n = 101, rate = 15)


# Plotting the data
par(mfrow = c(1,2))
plot(x, y, main = "Lambda = 2")
plot(x, y2, main = "Lambda = 20")
```



To be completely honest, the rate does not appear to change the scatter with the exception of a few deviations from $y = 2x + 5$ located at different x-values than the other plot.

5. The `subset()` function filters the observations in the data frame that meets a specified criterion. The `transform()` function takes a column vector within a data frame and is able to perform an operation on the entire column.

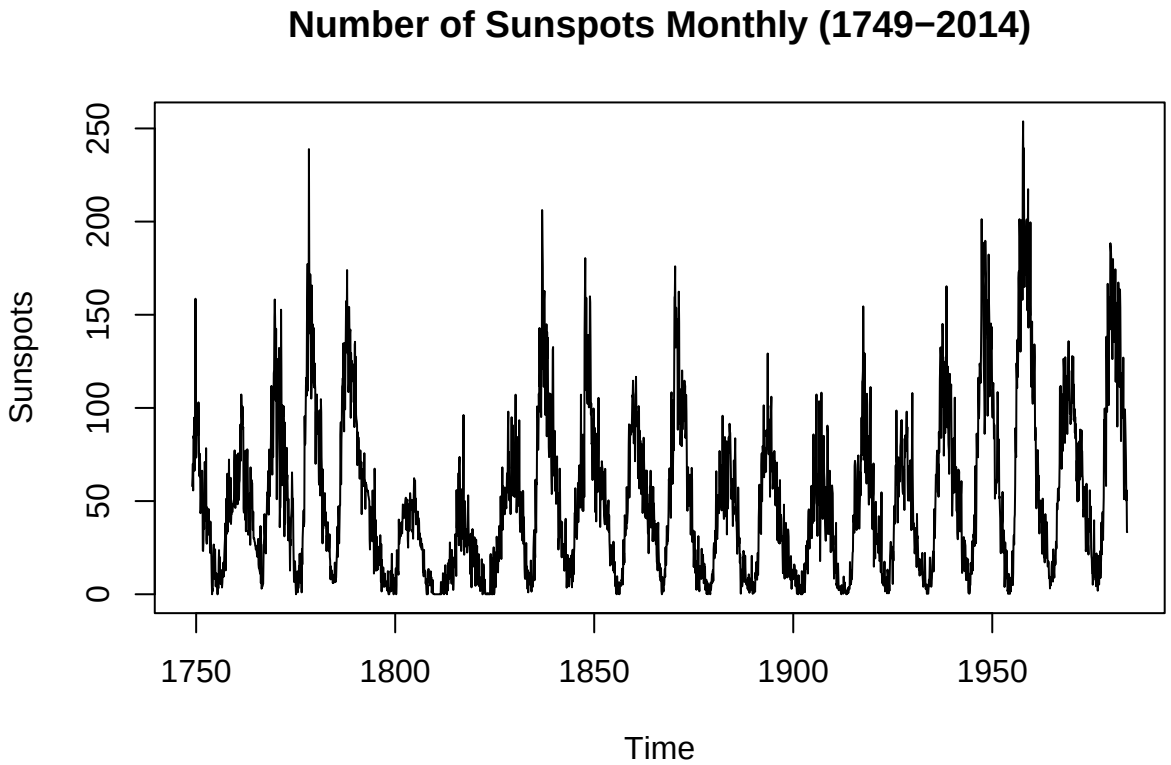
```
# An example of the subset() function
subset(cancer$colon, colon < 10)
# Error in subset.default(cancer$colon, colon < 10): object 'colon' not found

# An example of the transform() function
transform(cancer, colon = (colon - mean(colon))/sd(colon))
#     Country       colon meat
# 1   Nigeria -0.87381395   30
# 2     Japan -0.63312268   38
# 3   Jamaica -0.27208578   62
# 4       Yug -0.51277705   76
# 5  Columbia -0.63312268   80
# 6     Chile -0.57294987   82
# 7   Romania -0.51277705  105
# 8   Finland -0.42853511  105
# 9    Norway -0.21191296  110
# 10       PR -0.51277705  130
```

```
# 11    Israel -0.27208578  133
# 12    Poland -0.45260423  140
# 13   Hungary -0.51277705  141
# 14    Sweden -0.15174015  141
# 15   Netherl -0.15174015  150
# 16       DDP -0.33225860  170
# 17   Denmark  0.08895112  173
# 18       FDP -0.15174015  180
# 19   Iceland -0.17580927  198
# 20   England  0.08895112  202
# 21    Canada  1.89413564  235
# 22       USA  2.07465409  280
# 23 NZealand  3.21793761  315
```

6. Loading the data.

```
library(datasets)
data("sunspot.month")
```

```
par(mfrow = c(1,1))
plot(sunspots,
     main = "Number of Sunspots Monthly (1749-2014)",
     ylab = "Sunspots")
```



(a.)

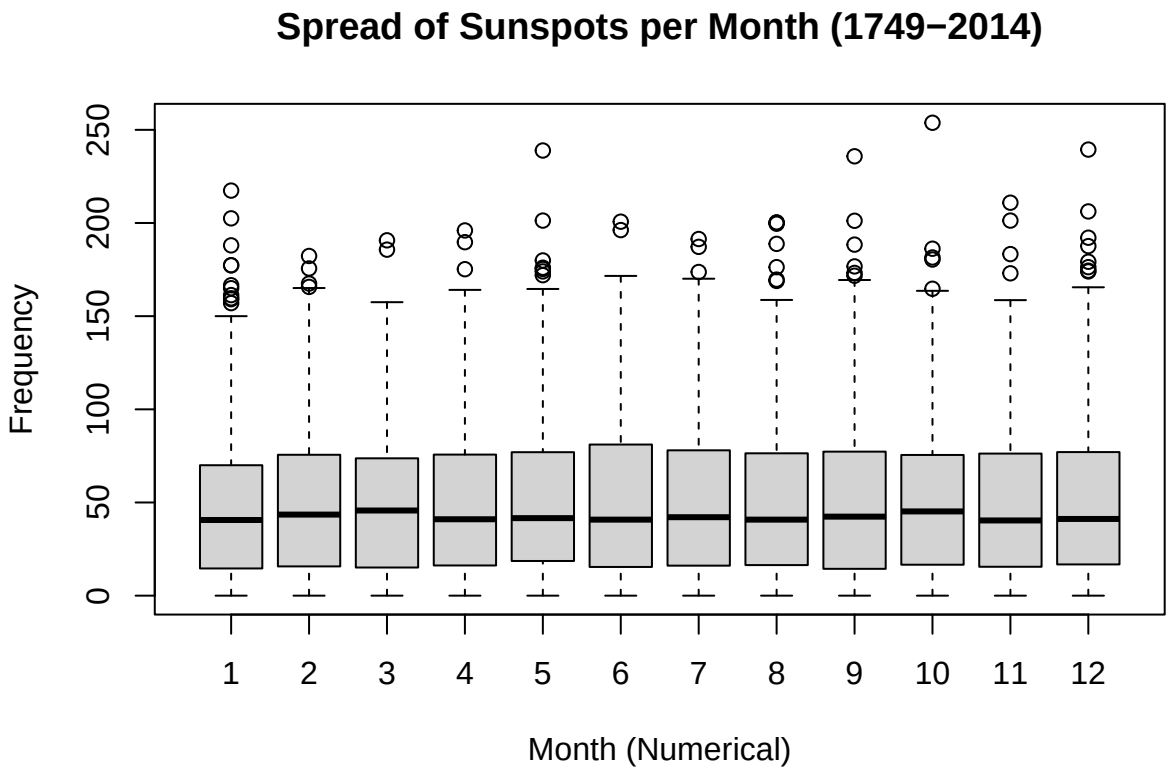
There appears to be oscillating behavior for frequency of sunspots, with each peak and trough occurring almost every ten years.

(b.)

```
# Creating the month variable
month <- rep(c(1:12), times = 265)
```

(c.)

```
# Creating the boxplot
boxplot(sunspot.month ~ month[1:3177],
        main = "Spread of Sunspots per Month (1749-2014)",
        ylab = "Frequency",
        xlab = "Month (Numerical)")
```




```
sunspots.df <- na.omit(as.data.frame(cbind(month[1:3177], sunspot.month)))

# Create an array
averages <- array()

# For loop takes average of each month and stores it in array
for (i in 1:12) {
  obs <- na.omit(subset(sunspots.df$sunspot.month, month == i))
```

```
  averages[i] <- mean(obs)
}
```

```
# Overlaying the array of averages on the boxplot
averages.df <- as.data.frame(averages)
boxplot(sunspot.month ~ month[1:3177],
        main = "Spread of Sunspots per Month (1749-2014)",
        ylab = "Frequency",
        xlab = "Month (Numerical)")
points(averages, col = "blue", pch = 19)
```