Stat 111 Homework 1

Kojin Oshiba

1. The CEO Golf and Stock Data

Descriptive Question

How is the stock rate distributed? What is its mean median, etc?

Predictive Question

Given the name of the CEO, the name of the company and the handicap score of the CEO, can we estimate what company's stock rate?

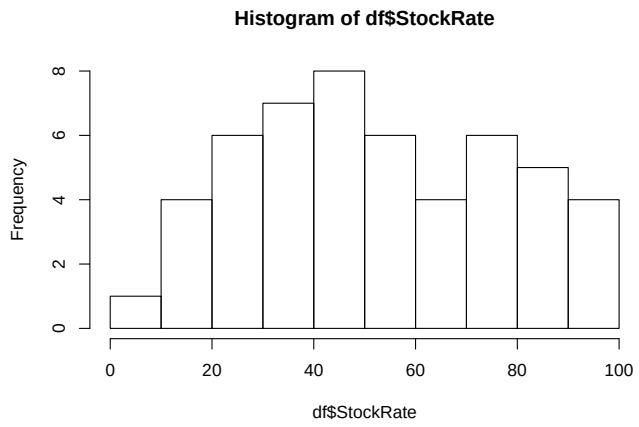
Causal Question

What is the causal effect of CEO's golf handicap on the company's stock rate? In other words, if we change the handicap score by 1 (holding everything else constant), what is the incremental change in the company's stock rate?

Adressing the Descriptive Question

The stock rate is distributed as below. The distribution is not very skewed, and has a mean of 52.47, median of 49.

```
df <- read.table("https://www.dartmouth.edu/~chance/teaching_aids/data/golf.txt",
                 header=TRUE)
hist(df$StockRate)
```



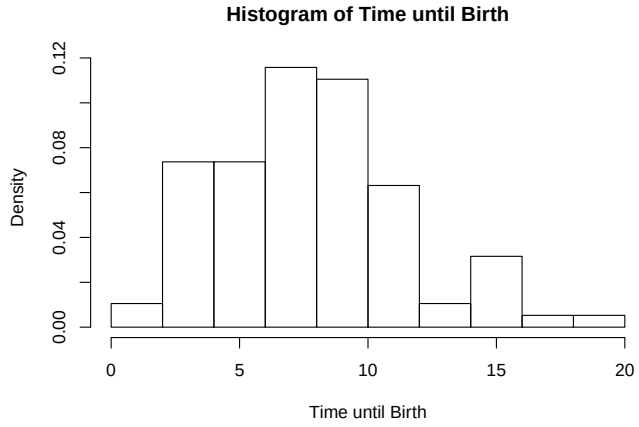

```
summary(df$StockRate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00   34.00   49.00   52.47   73.50   97.00
```

2. Histogram

(a) Compute the histogram for the durations of birth data.

```
df <- read.csv("Births/births.csv")
hist(df$time,xlab="Time until Birth",
     freq=FALSE,main="Histogram of Time until Birth")
```



(b) Derive the expectation of $H_{a,b}$.

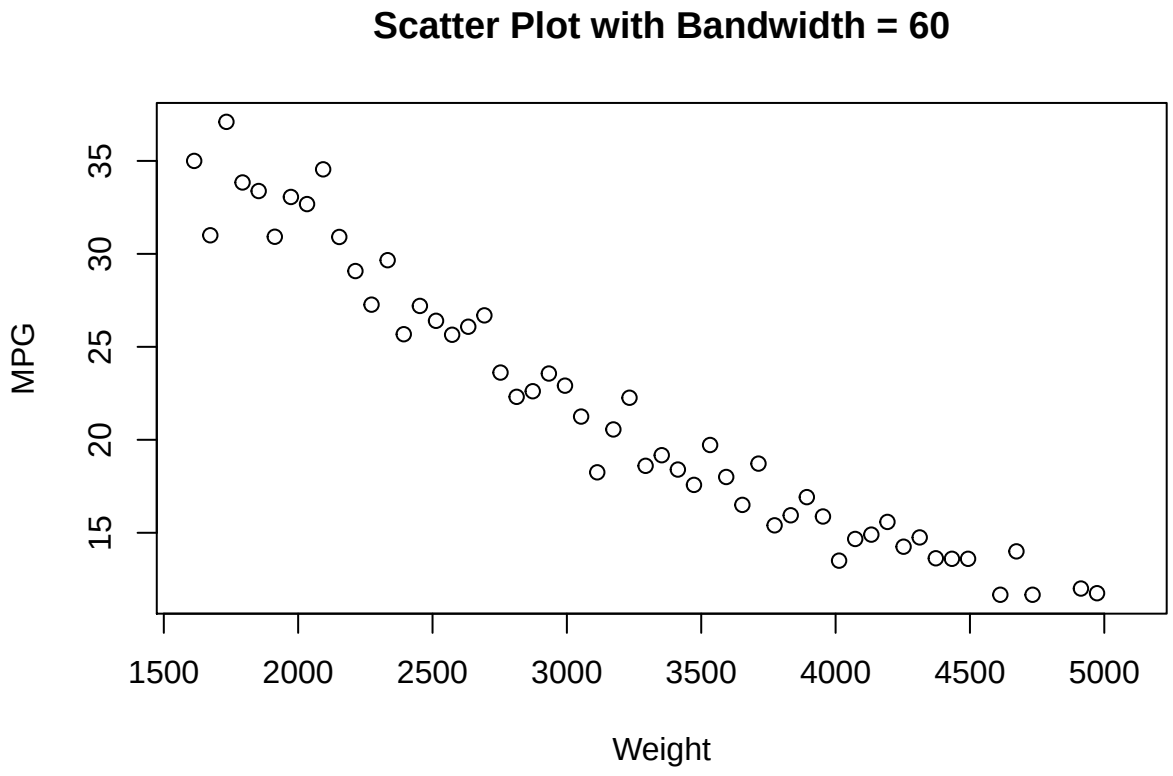
$$
\begin{aligned}
E(H_{a,b}) &= \frac{1}{n(b-a)} \sum_{i=1}^{n} E(\mathbf{1}_{y_i \in (a,b]}) \quad (\because \text{linearity}) \\
&= \frac{1}{n(b-a)} \sum_{i=1}^{n} P(y_i \in (a,b]) \quad (\because \text{fundamental bridge, from Stat110}) \\
&= \frac{1}{n(b-a)} \sum_{i=1}^{n} F(b) - F(a) \quad (\because \text{definition of CDF}) \\
&= \frac{F(b) - F(a)}{b-a}
\end{aligned}
$$

(c) Derive the variance of $H_{a,b}$.

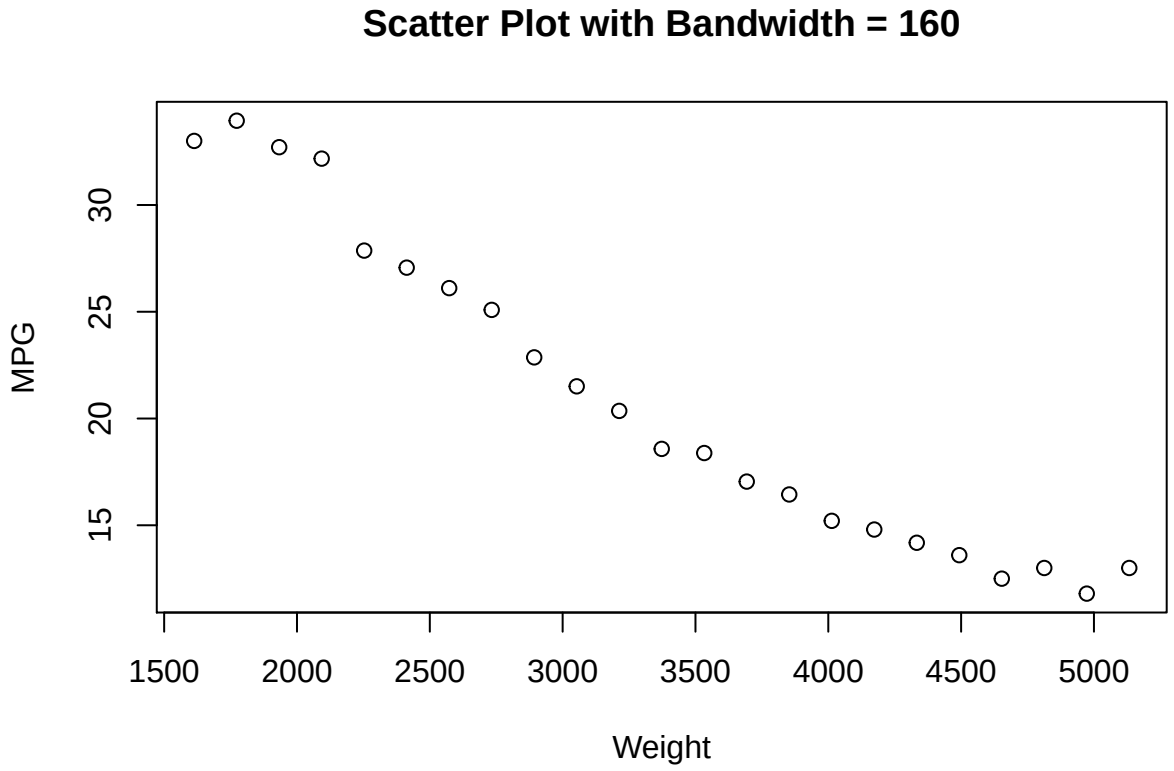
$$
\begin{aligned}
Var(H_{a,b}) &= \frac{1}{(n(b-a))^2} Var(\sum_{i=1}^{n} \mathbf{1}_{y_{i\in(a,b]}}) \quad (\because \text{constant comes out squared}) \\
&= \frac{1}{(n(b-a))^2} \sum_{i=1}^{n} Var(\mathbf{1}_{y_{i\in(a,b]}}) \quad (\because Y_i \text{ is iid}) \\
&= \frac{1}{(n(b-a))^2} \sum_{i=1}^{n} E(\mathbf{1}^2_{y_{i\in(a,b]}}) - E(\mathbf{1}_{y_{i\in(a,b]}})^2 \quad (\because \text{definition of variance}) \\
&= \frac{1}{(n(b-a))^2} \sum_{i=1}^{n} E(\mathbf{1}_{y_{i\in(a,b]}}) - E(\mathbf{1}_{y_{i\in(a,b]}})^2 \\
&= \frac{1}{(n(b-a))^2} \sum_{i=1}^{n} (F(b) - F(a)) - (F(b) - F(a))^2 \quad (\because \text{(b)}) \\
&= \frac{(F(b) - F(a)) - (F(b) - F(a))^2}{n(b-a)^2}
\end{aligned}
$$

3. AutoMPG Data

```
df <- read.table(paste0("https://archive.ics.uci.edu/ml",
                        "/machine-learning-databases/auto-mpg/auto-mpg.data"))
colnames <- c('mpg','cylinders','displacement','horsepower',
              'weight','acceleration','model year','origin','car name')
colnames(df) <- colnames

calc_ybar <- function(x,bandwidth) {
  # select elements within the bin
  bin_df <- df[(df$weight > x - bandwidth / 2)
               & (df$weight < x + bandwidth / 2),]
  # take the mean mpg value of each bin
  ybar <- mean(bin_df$mpg)
  ybar
}

plot_ybar <- function(bandwidth) {
  # select center points of bins
  xs <- seq(min(df$weight),max(df$weight),bandwidth)
  # calculate ybar for each bin
  ybars <- sapply(xs,calc_ybar,bandwidth=bandwidth)
  plot(xs, ybars, main = paste0("Scatter Plot with Bandwidth = ", bandwidth),
              xlab = "Weight",
              ylab = "MPG")
}

plot_ybar(60)
```



```
plot_ybar(160)
```



4. Census Income Data

The purpose of the question is to identify if there is an additive (or decremental) effect on income caused by divorce. This question can be useful, for example, when making advice to couples. It can also be useful for academic purposes, say for sociologists or economists who would like to understand the impact of divorce. The key point here is that we are not measuring the correlation between income and the rate of divorce. For example, if there exists a confounding factor, than correlation would not explain the impact of divorce on income. Hence, we need to ask a causal question.

The data might help in addressing this causal question only if we know how to use it. First, the data is likely observational, not randomized. Hence, a naive comparison of the average income in those who got divorced or not will not provide a valid causal effect. Applying methods like propensity score matching can overcome this issue.

5. Canonical Exponential Family

(a) Show that normal is in canonical exponential family.

$$\begin{aligned} f(y|\eta) &= \frac{1}{\sqrt{2\pi c^2}} exp(-\frac{(y-\eta)^2}{2c^2}) \\ &= \frac{1}{\sqrt{2\pi c^2}} exp(-\frac{1}{2c^2}(y^2 - 2y\eta + \eta^2)) \\ &= \frac{1}{\sqrt{2\pi c^2}} exp(-\frac{1}{2c^2}y^2) exp(-\frac{y\eta}{c^2} - \frac{\eta^2}{2c^2}) \end{aligned}$$

Hence,

$$\begin{aligned} h(y) &= \frac{1}{\sqrt{2\pi c^2}} exp(-\frac{1}{2c^2}y^2) \\ T(y) &= \frac{y}{c^2} \\ A(\eta) &= -\frac{\eta^2}{2c^2} \end{aligned}$$

(b) Show that binomial is in canonical exponential family.

$$\begin{aligned} f(y|\eta) &= \binom{n}{y} p^y (1-p)^(1-y) \\ &= \binom{n}{y} exp(ylogp + (n-y)log(1-p)) \\ &= \binom{n}{y} exp(ylog(\frac{p}{1-p} + nlog(1-p)) \end{aligned}$$

Let $\eta = log(\frac{p}{1-p})$. Then,

$$\begin{aligned} \eta &= log(\frac{p}{1-p}) \\ \Leftrightarrow e^\eta &= \frac{p}{1-p} \\ \Leftrightarrow \frac{1}{e^\eta} &= \frac{1}{p} - 1 \\ \Leftrightarrow p &= \frac{e^\eta}{e^\eta + 1} \end{aligned}$$

Hence,

$$\begin{aligned} h(y) &= \binom{n}{y} \\ T(y) &= y \\ A(\eta) &= nlog(1-p) = -nlog(e^\eta + 1) \end{aligned}$$

(c)

$$\begin{aligned}
E(e^{\delta T(Y)}|\eta) &= \int_{-\infty}^{\infty} e^{\delta T(y)} f(y|\eta) dy \\
&= \int_{-\infty}^{\infty} e^{\delta T(y)} h(y) e^{\eta T(y) - A(\eta)} dy \\
&= \int_{-\infty}^{\infty} e^{(\delta+\eta) T(y) - A(\eta)} h(y) dy \\
&= (e^{A(\eta+\delta) - A(\eta)}) \int_{-\infty}^{\infty} e^{(\delta+\eta) T(y) - A(\delta+\eta)} h(y) dy \\
&= (e^{A(\eta+\delta) - A(\eta)}) \int_{-\infty}^{\infty} f(y|\delta+\eta) dy \\
&= e^{A(\eta+\delta) - A(\eta)}
\end{aligned}$$

Hence,

$$E(e^{\delta T(Y)}|\eta) = e^{A(\eta+\delta) - A(\eta)}$$

From the definition of MGF,

$$E(e^{\delta T(Y)}|\eta) = M_{T(Y)|\eta}(\delta)$$

Since mean is the first moment,

$$E(T(Y)|\eta) = M^{(1)}_{T(Y)|\eta}(0)$$

Let's evaluate $M^{(1)}_{T(Y)|\eta}(\delta)$.

$$\begin{aligned}
M^{(1)}_{T(Y)|\eta}(\delta) &= \frac{\partial}{\partial \delta} e^{A(\eta+\delta) - A(\eta)} \\
&= (\frac{\partial}{\partial \delta} A(\eta+\delta)) e^{A(\eta+\delta) - A(\eta)} \\
&= (\frac{\partial}{\partial \eta} A(\eta+\delta)) e^{A(\eta+\delta) - A(\eta)} \quad (\because \text{symmetry of } \eta \text{ and } \delta)
\end{aligned}$$

Hence,

$$E(T(Y)|\eta) = M^{(1)}_{T(Y)|\eta}(0) = \frac{\partial}{\partial \eta} A(\eta)$$