MTH 427 - Spring 2023
Assignment #1
Due: Monday, January 30th 2023 (11:59PM)

1 Conceptual problems

1. For any random variables X, Y and any constants a, b, c and d, show that $\operatorname{Cov}(a + bX, c + dY) = bd \operatorname{Cov}(X, Y)$.

   Proof.

$$\begin{aligned}
\operatorname{Cov}(a+bX, c+dY) &= \mathrm{E}\left[(a+bX)(c+dY)\right] - \mathrm{E}\left[a+bX\right]\mathrm{E}\left[c+dY\right] \\
&= \mathrm{E}\left[ac + adY + bcX + bdXY\right] - \left(\mathrm{E}\left[a\right] + \mathrm{E}\left[bX\right]\right)\left(\mathrm{E}\left[c\right] + \mathrm{E}\left[dY\right]\right) \\
&= ac + ad\,\mathrm{E}\left[Y\right] + bc\,\mathrm{E}\left[X\right] + bd\,\mathrm{E}\left[XY\right] - \left(a + b\,\mathrm{E}\left[X\right]\right)\left(c + d\,\mathrm{E}\left[Y\right]\right) \\
&= ac + ad\,\mathrm{E}\left[Y\right] + bc\,\mathrm{E}\left[X\right] + bd\,\mathrm{E}\left[XY\right] - \left(ac + ad\,\mathrm{E}\left[Y\right] + bc\,\mathrm{E}\left[X\right] + bd\,\mathrm{E}\left[X\right]\mathrm{E}\left[Y\right]\right) \\
&= bd\,\mathrm{E}\left[XY\right] - bd\,\mathrm{E}\left[X\right]\mathrm{E}\left[Y\right] && \text{(“Lots of killing” - McAsey)} \\
&= bd\left(\mathrm{E}\left[XY\right] - \mathrm{E}\left[X\right]\mathrm{E}\left[Y\right]\right) \\
&= bd\operatorname{Cov}(X, Y)
\end{aligned}$$

□

2. Suppose that $\mathrm{E}(\widehat{\theta}_1) = \mathrm{E}(\widehat{\theta}_2) = \theta$, $\operatorname{Var}(\widehat{\theta}_1) = \sigma_1^2$, and $\operatorname{Var}(\widehat{\theta}_2) = \sigma_2^2$. Consider the estimator $\widehat{\theta}_3 = \alpha\widehat{\theta}_1 + (1-\alpha)\widehat{\theta}_2$

   (a) Show that $\widehat{\theta}_3$ is an unbiased estimator for θ.

   Proof.

$$\begin{aligned}
\mathrm{E}[\widehat{\theta}_3] &= \mathrm{E}[\alpha\widehat{\theta}_1 + (1-\alpha)\widehat{\theta}_2] && \text{(given)} \\
&= \alpha\,\mathrm{E}[\widehat{\theta}_1] + (1-\alpha)\,\mathrm{E}[\widehat{\theta}_2] && \text{(linearity)} \\
&= \alpha\theta + (1-\alpha)\theta && \text{(substitute with hypotheses)} \\
&= \theta\,(\alpha + 1 - \alpha) && \text{(factor)} \\
&= \theta && \text{(simplify)}
\end{aligned}$$

□

   (b) If $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are independent, how should α be chosen in order to minimize the variance of $\widehat{\theta}_3$?

   Solution: First derive the general form of the variance of the estimator $\widehat{\theta}_3$. That is,

$$\begin{aligned}
\operatorname{Var}[\,\widehat{\theta}_3\,] &= \operatorname{Var}[\,\alpha\widehat{\theta}_1 + (1-\alpha)\widehat{\theta}_2\,] && \text{(substitute)} \\
&= \alpha^2\operatorname{Var}[\,\widehat{\theta}_1\,] + (1-\alpha)^2\operatorname{Var}[\,\widehat{\theta}_2\,] + 2\alpha(1-\alpha)\operatorname{Cov}(\widehat{\theta}_1, \widehat{\theta}_2) && \text{(formula)} \\
&= \alpha^2\operatorname{Var}[\,\widehat{\theta}_1\,] + (1-\alpha)^2\operatorname{Var}[\,\widehat{\theta}_2\,] && \text{(By independence hypothesis)} \\
&= \alpha^2{\sigma_1}^2 + (1-\alpha)^2{\sigma_2}^2 && \text{(substitute)}
\end{aligned}$$

Using the first derivative test wrt α, $\frac{\partial}{\partial\alpha}\left(\alpha^2{\sigma_1}^2 + (1-\alpha)^2{\sigma_2}^2\right) = 2\alpha({\sigma_1}^2+{\sigma_2}^2)+2{\sigma_2}^2$. Setting this equal to 0 and solving for α yields $\alpha = \frac{{\sigma_2}^2}{{\sigma_1}^2+{\sigma_2}^2}$.

3. Suppose that X_1, X_2, X_3 denote a random sample from an exponential distribution with density function

$$f(x) = \begin{cases} \left(\frac{1}{\theta}\right)\exp(-x/\theta) & x > 0 \\ 0 & \text{elsewhere} \end{cases}$$

consider the following four estimators of θ:

$$\widehat{\theta}_1 = X_1, \qquad \widehat{\theta}_2 = \frac{X_1 + 2X_2}{3}, \qquad \widehat{\theta}_3 = \bar{X}, \qquad \widehat{\theta}_4 = \min(X_1, X_2, X_3)$$

Which of these estimators are unbiased? (Show your work.)

Solution: Start by computing the expected value of the distribution of X,

$$\begin{aligned} \mathrm{E}[X] &= \int_{-\infty}^{\infty} x f(x)\mathrm{d}x \\ &= \frac{1}{\theta}\int_{-\infty}^{\infty} x e^{-x/\theta}\mathrm{d}x \\ &= \frac{1}{\theta}\int_{0}^{\infty} x e^{-x/\theta}\mathrm{d}x \\ \text{Let } u = x, \quad \mathrm{d}u &= \mathrm{d}x, \quad \mathrm{d}v = e^{-x/\theta}\mathrm{d}x, \quad v = -\theta e^{-x/\theta} \\ &= \frac{1}{\theta}\left(x(-\theta e^{-x/\theta}) - \int -\theta e^{-x/\theta}\mathrm{d}x\right)\Bigg|_0^{\infty} \\ &= \frac{1}{\theta}\left(x(-\theta e^{-x/\theta}) + \theta\int e^{-x/\theta}\mathrm{d}x\right)\Bigg|_0^{\infty} \\ &= \frac{1}{\theta}\left(x(-\theta e^{-x/\theta}) - \theta(\theta e^{-x/\theta})\right)\Bigg|_0^{\infty} \\ &= -e^{-x/\theta}(x+\theta)\Big|_0^{\infty} \\ &= -\left(\lim_{x\to\infty} e^{-x/\theta}(x+\theta) - e^0(0+\theta)\right) \\ &= -\left(\lim_{x\to\infty}\frac{x}{e^{x/\theta}} + \theta\lim_{x\to\infty} e^{-x/\theta} - \theta\right) \\ &= -\left(\lim_{x\to\infty}\frac{1}{\frac{1}{\theta}e^{x/\theta}} - \theta\right) \\ &= -(-\theta) \\ &= \theta \end{aligned}$$

$\widehat{\theta}_1$ **Solution:** $\mathrm{E}[\widehat{\theta}_1] = \mathrm{E}[X_1] = \mathrm{E}[X] = \theta \therefore$ unbiased.

$\widehat{\theta}_2$ **Solution:** $\mathrm{E}[\widehat{\theta}_2] = \mathrm{E}\left[\frac{X_1+2X_2}{3}\right] = \frac{1}{3}\mathrm{E}[X_1+2X_3] = \frac{1}{3}\mathrm{E}[3X] = \frac{3\theta}{3} = \theta \therefore$ unbiased.

$\widehat{\theta}_3$ **Solution:** $\mathrm{E}[\widehat{\theta}_3] = \mathrm{E}[\bar{X}] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[X\right] = \frac{n\theta}{n} = \theta \therefore$ unbiased.

$\widehat{\theta}_4$ **Solution:** Let $T := \min\left(X_1, X_2, X_3\right) = \widehat{\theta}_4$. Then for some t,

$$\begin{aligned}\Pr\left(T > t\right) &= \prod_{i=1}^{3} \Pr\left(X_i > t\right)\\ &= \Pr\left(X > t\right)^3\\ &= \left[e^{-t/\theta}\right]^3\\ &= e^{-3t/\theta}\end{aligned}$$

Then the CDF $F(t) = \Pr\left(T \leq t\right) = 1 - e^{-3t/\theta}$. Hence the PDF $f(t) = F'(T) = \frac{3}{\theta}e^{-3t/\theta}$. Then $\mathrm{E}[\widehat{\theta}_3] = \mathrm{E}[f(t)] = \mathrm{E}\left[\frac{3}{\theta}e^{-3t/\theta}\right] = \frac{\theta}{3} \neq \theta \therefore$ biased.

2 Applied problems - R

1. This exercise relates to the "Credit" dataset, which can be found as "Credit.csv" in Canvas.

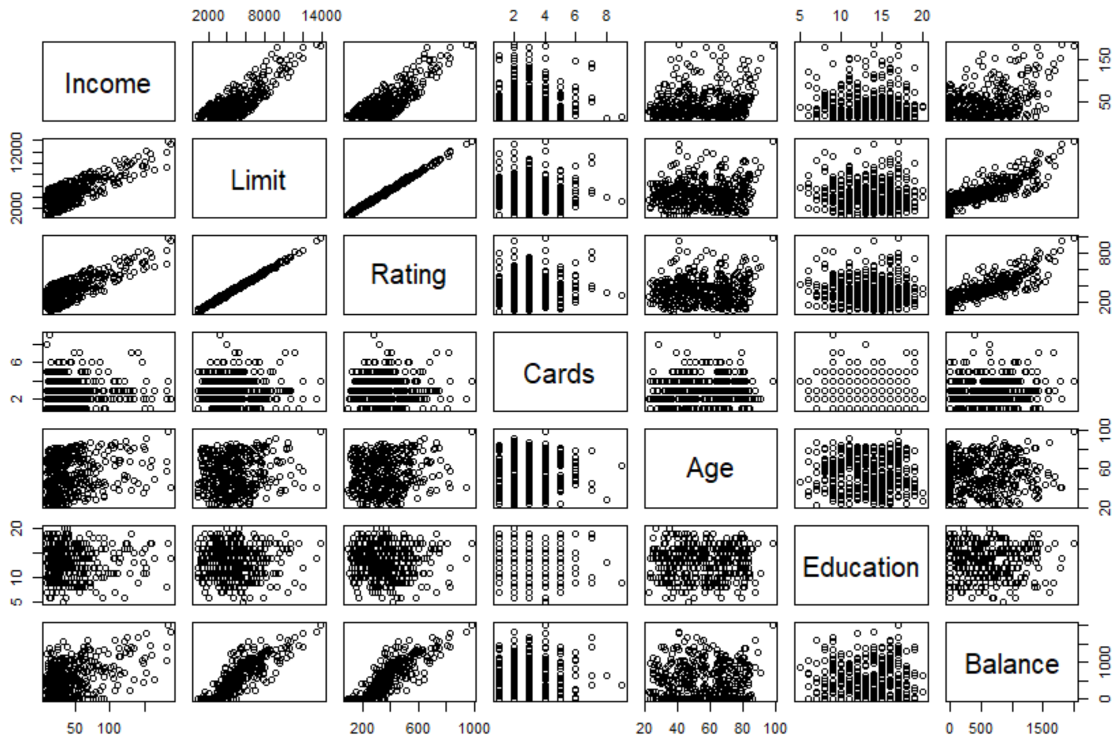
   (a) Use the appropriate function in R to produce a numerical summary of the quantitative variables in the data.

```r
library(readr)

credit_data = read.csv("O:/Arr Matey/Credit.csv", header=T, na.strings="?")
credit_data = na.omit(credit_data)

quantitative_credit_data = credit_data[, c(1:6, 11)]
summary(quantitative_credit_data)
```

```
> summary(quantitative_credit_data)
     Income           Limit           Rating          Cards            Age       
 Min.   : 10.35   Min.   :  855   Min.   : 93.0   Min.   :1.000   Min.   :23.00  
 1st Qu.: 21.01   1st Qu.: 3088   1st Qu.:247.2   1st Qu.:2.000   1st Qu.:41.75  
 Median : 33.12   Median : 4622   Median :344.0   Median :3.000   Median :56.00  
 Mean   : 45.22   Mean   : 4736   Mean   :354.9   Mean   :2.958   Mean   :55.67  
 3rd Qu.: 57.47   3rd Qu.: 5873   3rd Qu.:437.2   3rd Qu.:4.000   3rd Qu.:70.00  
 Max.   :186.63   Max.   :13913   Max.   :982.0   Max.   :9.000   Max.   :98.00  
   Education        Balance       
 Min.   : 5.00   Min.   :   0.00  
 1st Qu.:11.00   1st Qu.:  68.75  
 Median :14.00   Median : 459.50  
 Mean   :13.45   Mean   : 520.01  
 3rd Qu.:16.00   3rd Qu.: 863.00  
 Max.   :20.00   Max.   :1999.00  
> 
```

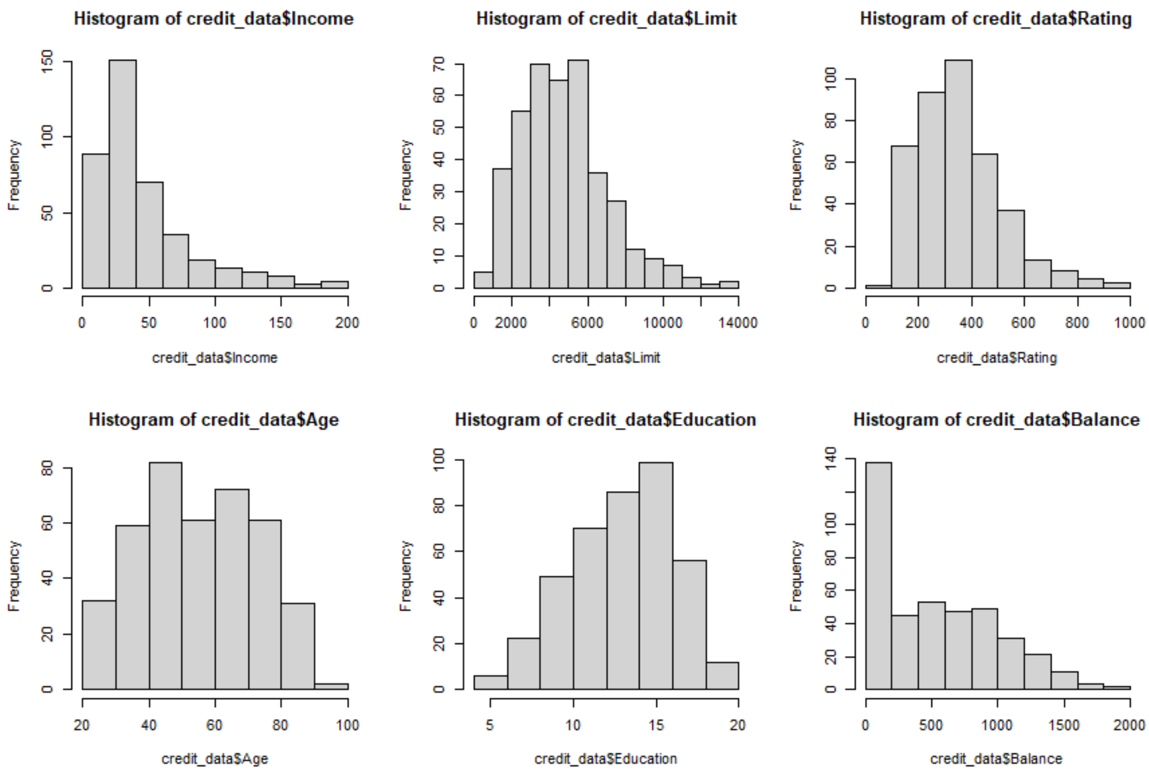
(b) Display a scatter plot matrix between quantitative variables in the data set.

```
# Question 1b
pairs(quantitative_credit_data)
```



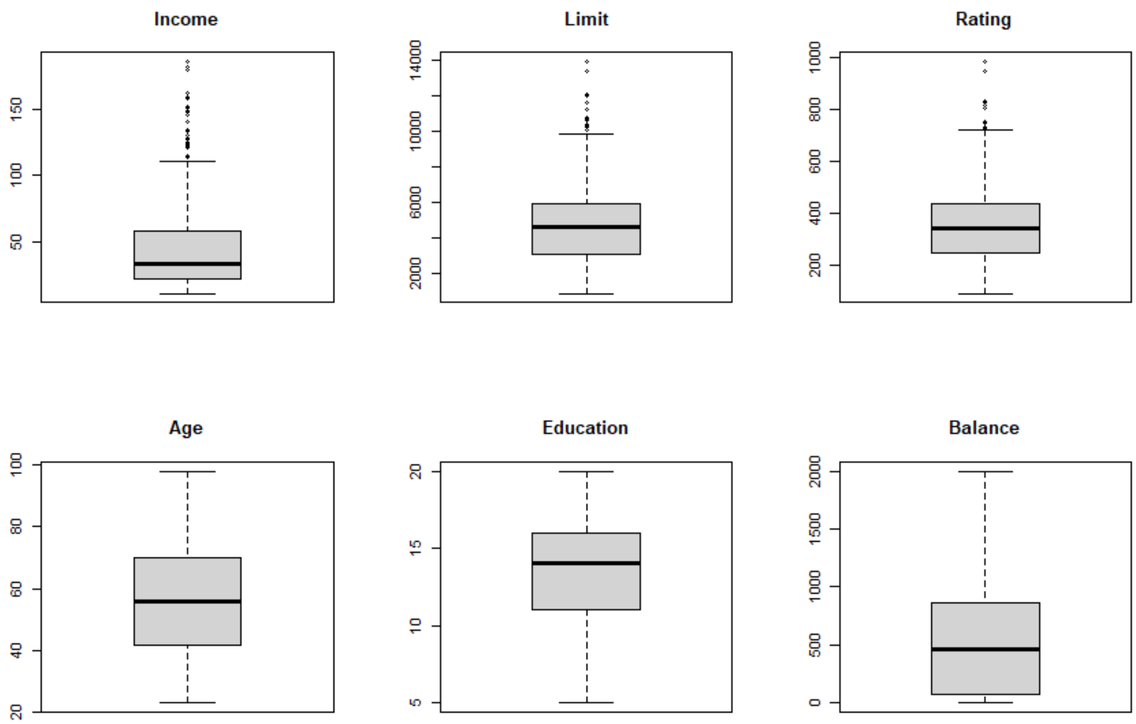
(c) Display histograms of all quantitative variables in one graph except "cards" (side by side). You may use different colors.

```
# Question 1c
par(mfrow=c(2,3))
hist(credit_data$Income)
hist(credit_data$Limit)
hist(credit_data$Rating)
hist(credit_data$Age)
hist(credit_data$Education)
hist(credit_data$Balance)
```



(d) Display box-plots of all quantitative variables only in one graph except "cards" (side by side). Make sure to label them.

```
# Question 1d
par(mfrow=c(2,3))
boxplot(credit_data$Income, main="Income")
boxplot(credit_data$Limit, main="Limit")
boxplot(credit_data$Rating, main="Rating")
boxplot(credit_data$Age, main="Age")
boxplot(credit_data$Education, main="Education")
boxplot(credit_data$Balance, main="Balance")
```



2. This exercise relates to the "Hwk-data1" dataset, which can be found in Canvas.

   Operators of gasoline-fueled vehicles complain about the price of gasoline in gas stations. According to the American Petroleum Institute, the federal gas tax per gallon is constant (18.4 cents as of January 13, 2005), but state and local taxes vary from 7.5 cents to 32.10 cents for $n = 18$ key metropolitan areas around the country.

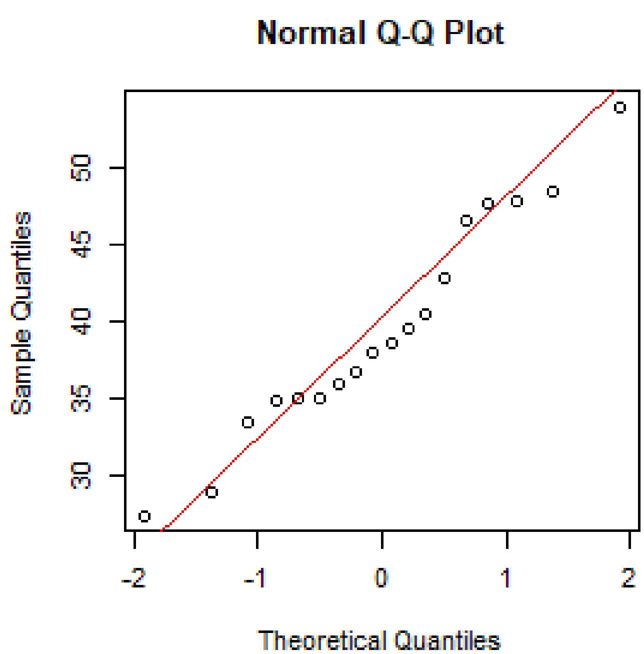
   Use R for part (a), (b), and (c)

   (a) Check whether the data are normally distributed by using the Shapiro test or by looking at the QQ plot. (Make sure to display your results).

```
library(readr)

data = read.csv("O:/Arr Matey/Hwk-data1.csv", header=T, na.strings="?")
data = na.omit(data)

tax_per_gal = data$Tax_per_gallon;

qqnorm(tax_per_gal)
qqline(tax_per_gal, col='red')

shapiro.test(tax_per_gal)
```

```
> shapiro.test(tax_per_gal)

        Shapiro-Wilk normality test

data:  tax_per_gal
W = 0.96231, p-value = 0.6469
```

   Since $p = 0.6469 > \alpha$ (assuming $\alpha = 0.05$), the data are normally distributed.



(b) Use the appropriate R function to find a 90% confidence interval for the average per gallon gas tax in the U.S. (Make sure to display your code and the corresponding result.)

```
> # Question 2b
> t.test(data, conf.level=0.90)$"conf.int"
[1] 36.62947 42.48275
attr(,"conf.level")
[1] 0.9
> |
```

The confidence interval $I := (36.62947, 42.48275)$.

(c) Is there sufficient evidence to claim that the average gas tax is less that 45.2 cents? (Make sure to specify hypotheses and the p-value).

Solution: $H_0 := \mu = 45.2$ and $H_a := \mu < 45.2$.

```
> # Question 2c
> t.test(data, conf.level=0.90, mu=45.2)

        One Sample t-test

data:  data
t = -3.3547, df = 17, p-value = 0.003758
alternative hypothesis: true mean is not equal to 45.2
90 percent confidence interval:
 36.62947 42.48275
sample estimates:
mean of x
 39.55611
```

Since $p = 0.003758 \leq 0.1 = \alpha$, we reject the null hypothesis (H_0). Hence, there is sufficient evidence that the true mean of the gas tax is less than 45.2 cents.

(d) Compute (by hand) a 98% confidence interval for the average per gallon gas tax in the U.S. Compare the length of this interval and the one in part (b). (Hint: the sample standard deviation $s = 7.138$)

Solution: $\bar{X} \approx 39.556$, df $= 18 - 1 = 17$, and $\alpha = 0.02$. Then $t_{\alpha/2}(17) = t_{0.01}(17) = 2.567$.

Then the CI is $39.556 \pm 2.567\left(\frac{7.138}{\sqrt{18}}\right) \equiv (35.237, 43.875)$. The measure of this interval is 8.638 cents, whereas in part (b) the measure was 5.85328 cents. Higher confidence levels require more of the domain since $\lim_{\alpha \to 0^+} \equiv D$.