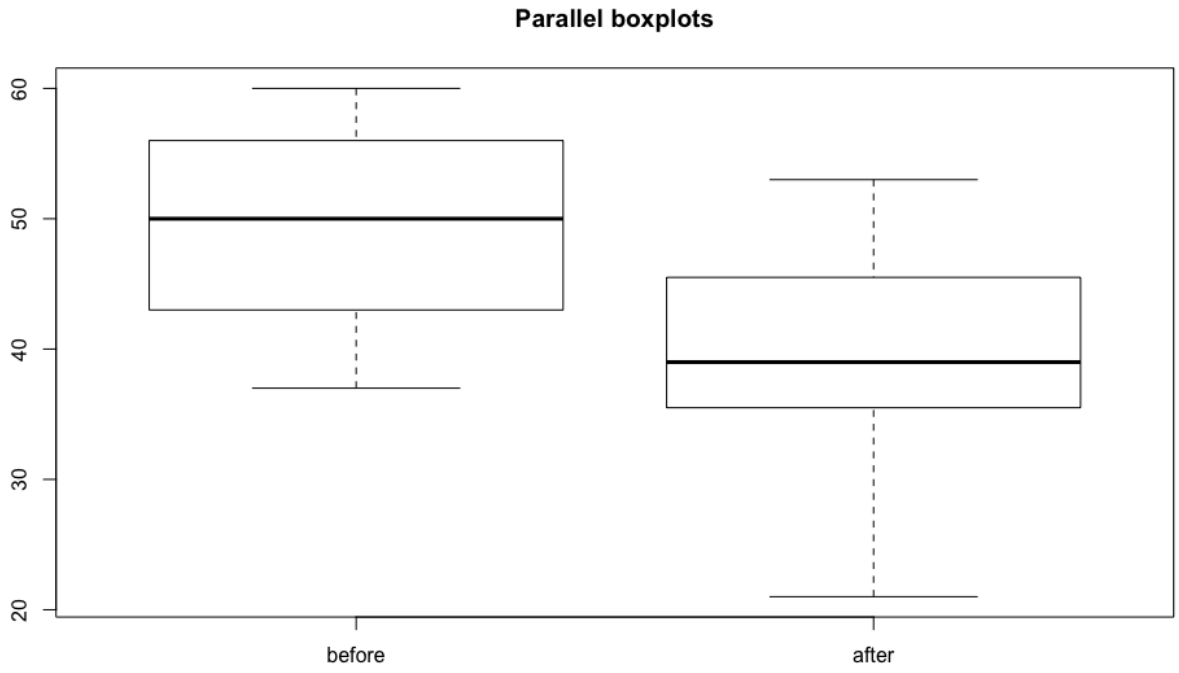
Statistical Methods for Data Science

HW 3 Solution

Exercise 8.1

(b) Before the change of firewall setting, $\min(X) = 37, \widehat{Q_1} = 43, \widehat{M} = 50, \widehat{Q_3} = 56, \max(X) = 60$.
The sample interquartile range is 56-43=13, all the data are within $1.5(\widehat{IQR})$from $\widehat{Q_1}$ or $\widehat{Q_3}$, and thus, we do not suspect any outliers.
After the change, $\min(X) = 21, \widehat{Q_1} = 35, \widehat{M} = 39, \widehat{Q_3} = 46, \max(X) = 53$.
In fact, according to Definition 8.7, any number between 35 and 36 is the first quantile, and any number between 45 and 46 is the third quantile.
Next, $\widehat{IQR} = 10$, and again, we see no outliers in this sample.



(c) Every element of the five-number summary has reduced following the change in firewall settings, suggesting that any number of intrusion attempts is now exceeded with a lower probability.

Exercise 8.2

(a) By Definition 8.3 and 8.8,

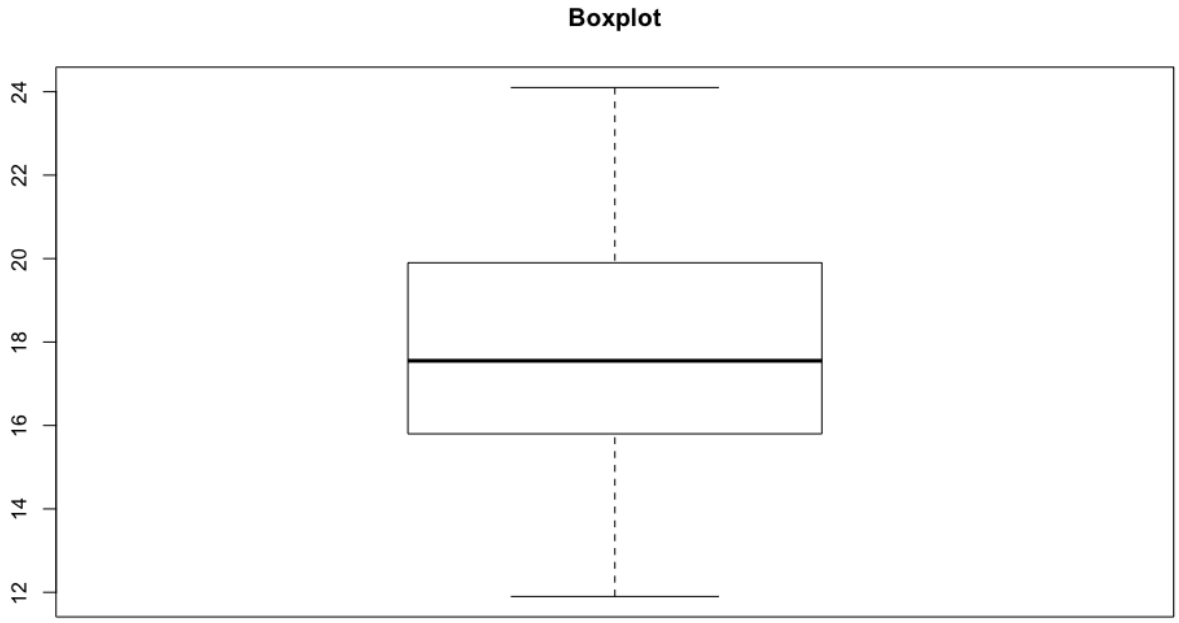
$$\overline{X} = 17.9540, \quad S^2 = 9.9682, \quad S = 3.1573$$

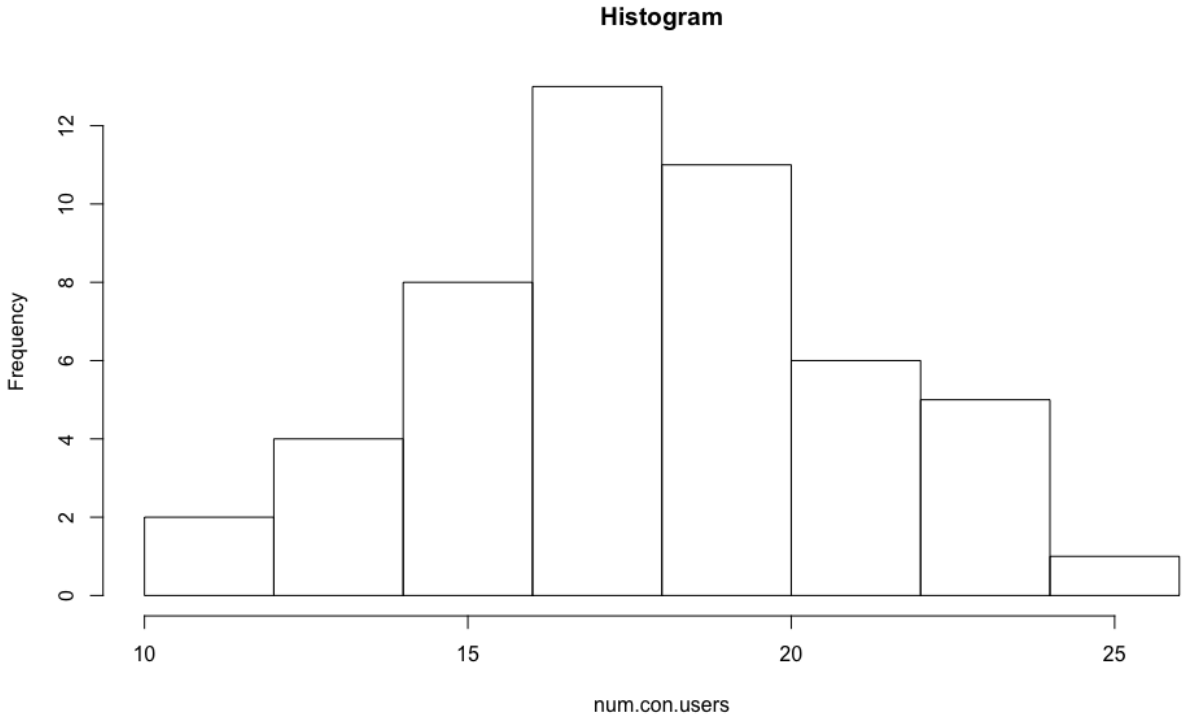
(b)

$$S(\overline{X}) = S/\sqrt{n} = 3.1573/\sqrt{50} = 0.4465.$$

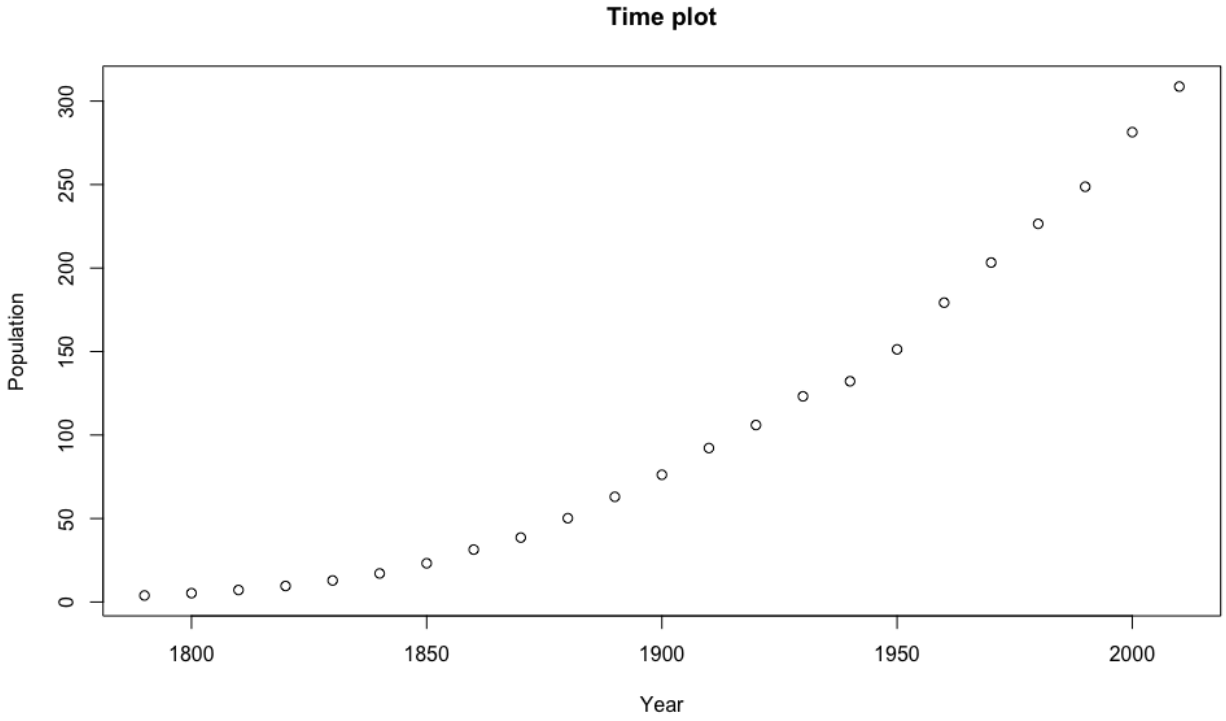
(c) Order the data from smallest to largest. The minimum is 11.9, the maximum is 24.1. For the sample of size 50, the median $\widehat{M}$ is any number between the $25th$ and $26th$ smallest, i.e., between 17.5 and 17.6. The first quantile $\widehat{Q_1}$ is the $13th$ smallest, i.e., 15.8 (exceeds $22\% \leq 25\%$ of the sample; is exceeded by $74\% \leq 75\%$ of the sample), and the third quantile $\widehat{Q_3}$ is the $37th$ smallest, i.e., 19.9.

The five-point summary is (11.9, 15.8, 17.55, 19.9, 24.1).
Below are boxplot and histogram.



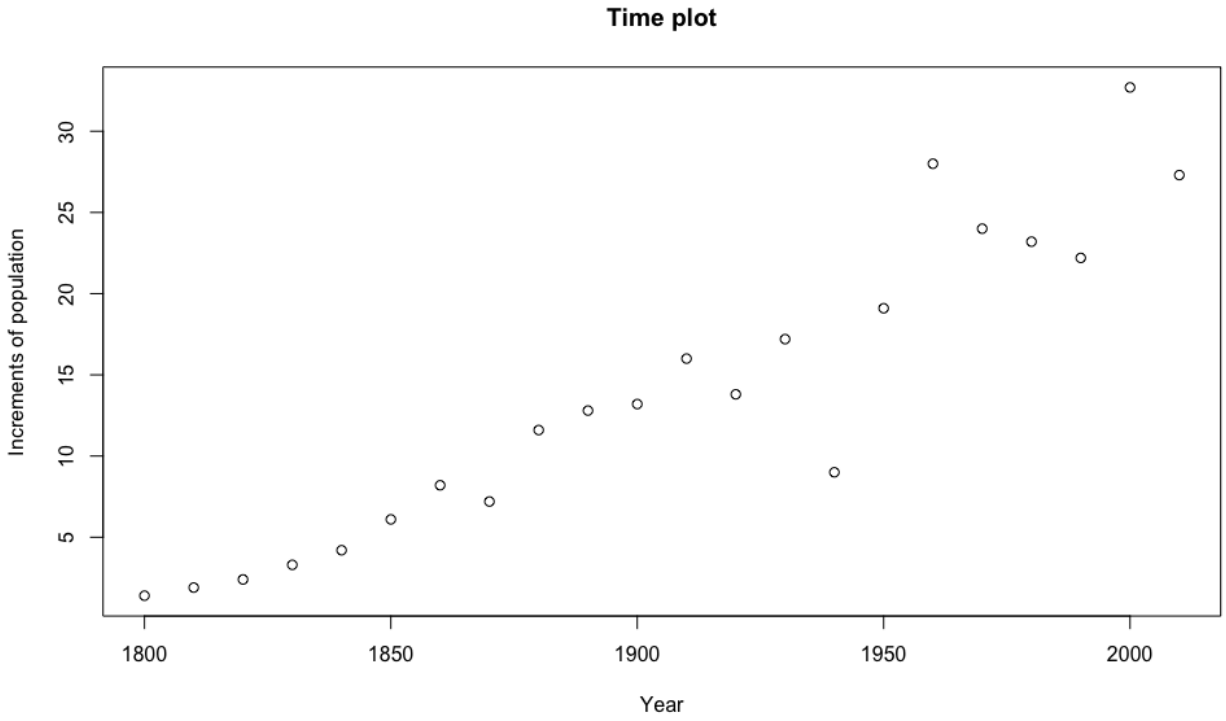


Exercise 8.5 The time plot is on the next page. There is a steady nonlinear growth of the population although during the $20th$ century it may have become linear. The U.S. population increases every decade.



Exercise 8.6

(a) The sample mean is 13.85 (mln), the sample median is 13.00 (mln), and then variance is 87.60 (mln^2). On the average, the U.S. population increased by 13.85 (mln) people per decade. Due to an obvious time trend seen in (b), the data do not follow the same distribution, therefore, the sample mean and the sample variance are not unbiased estimates of the population mean and variance.

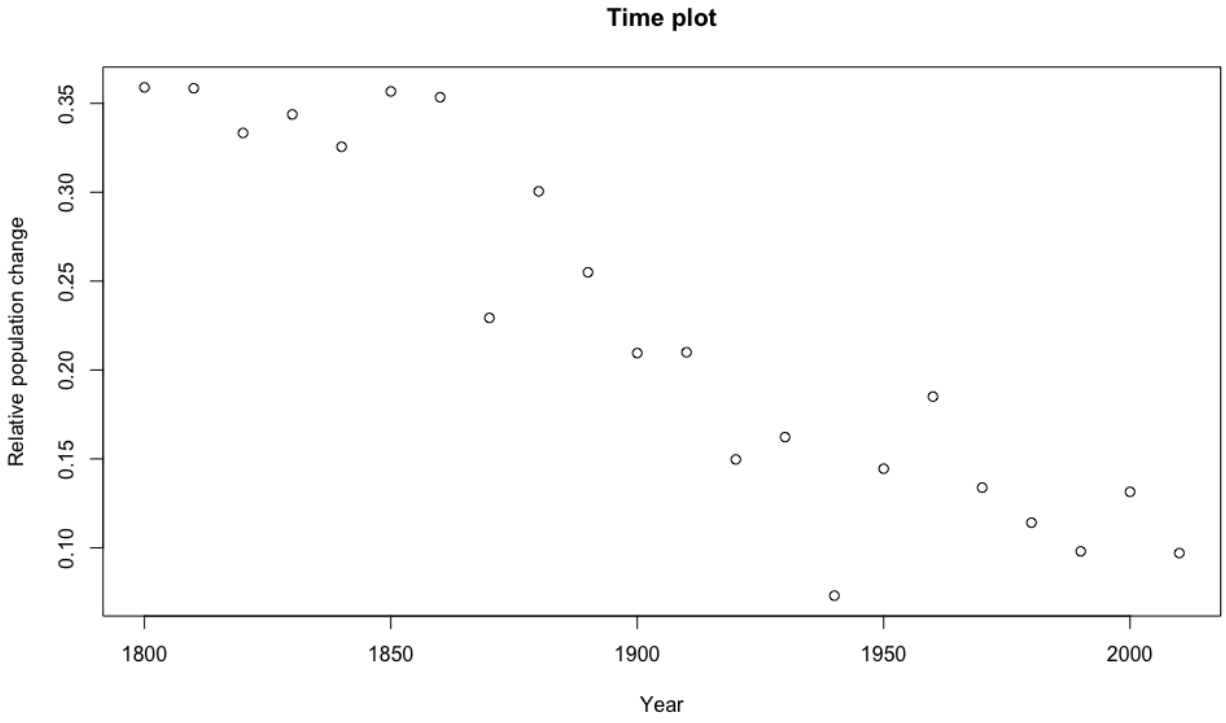


(b) The time plot is in the front page. It shows an overall increasing trend except for the period between 1960 and 1990 when the population growth was slowing.

Exercise 8.7

(a) The sample mean is 0.22, the sample median is 0.21, and the variance is 0.01. That is, the U.S. population has increased by 22%, on the average every 10 years.

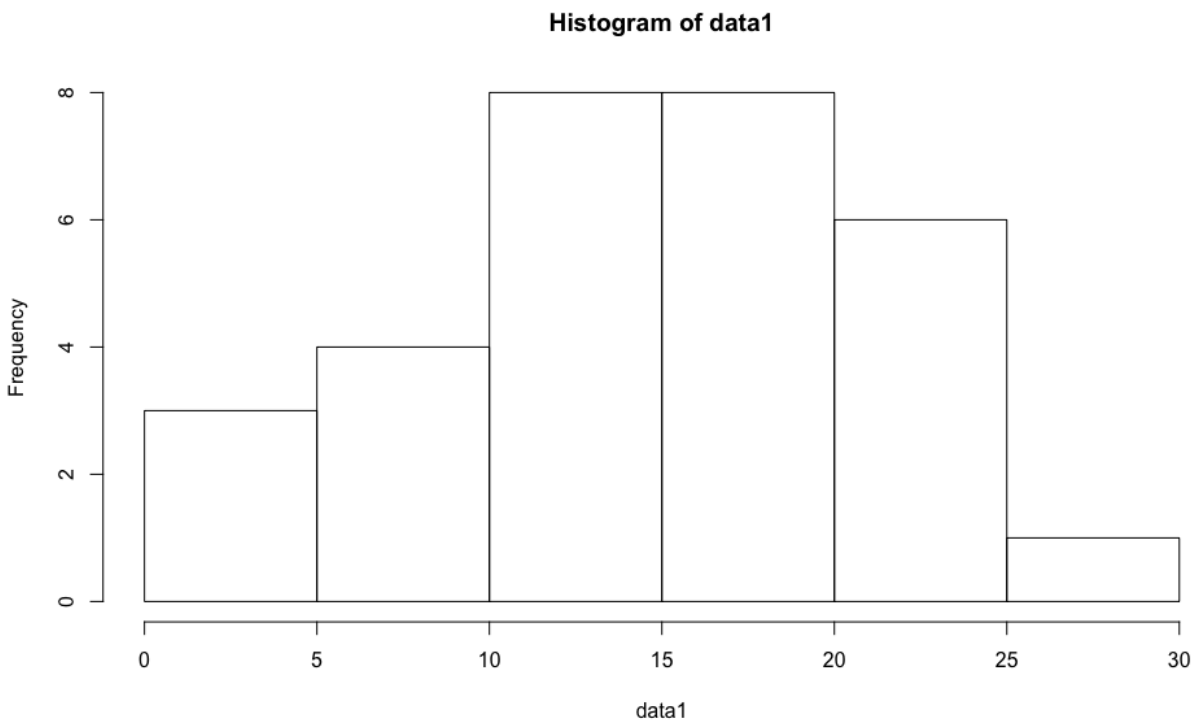
(b) The time plot is the following. We see that in general, the proportional population change decreases, and the trend is almost linear. In 1880-1860, the population increased by about 35% each decade whereas its relative growth never exceeded 20% since 1920.

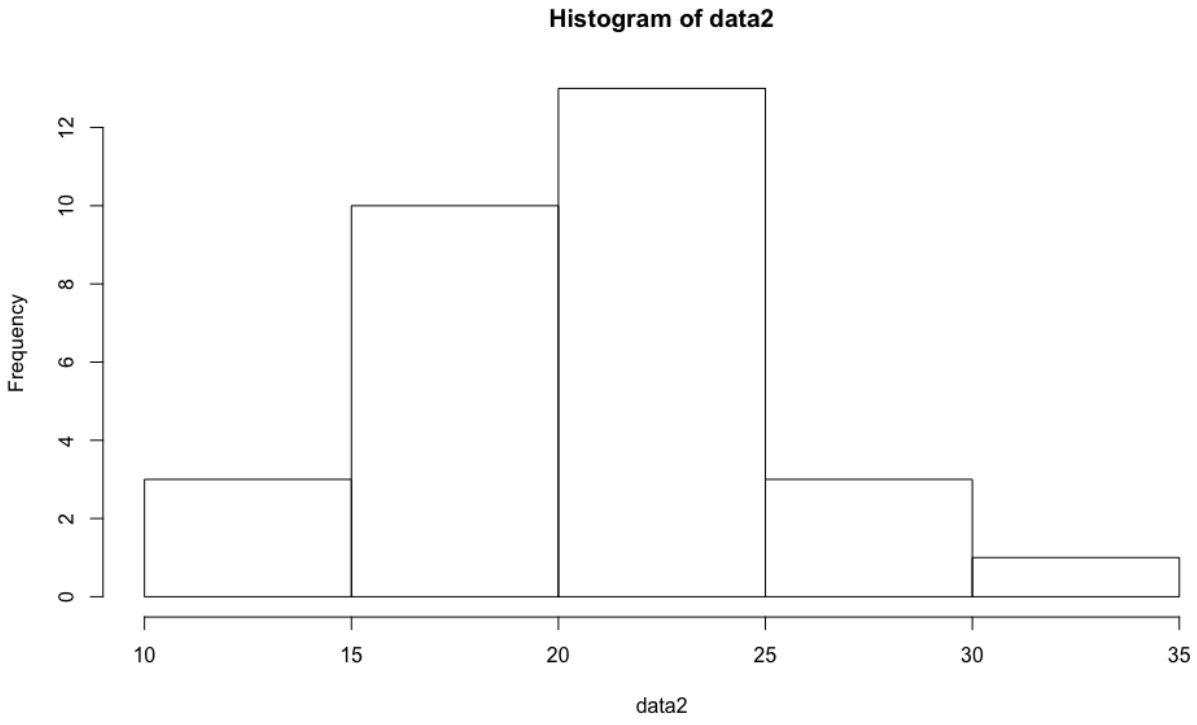


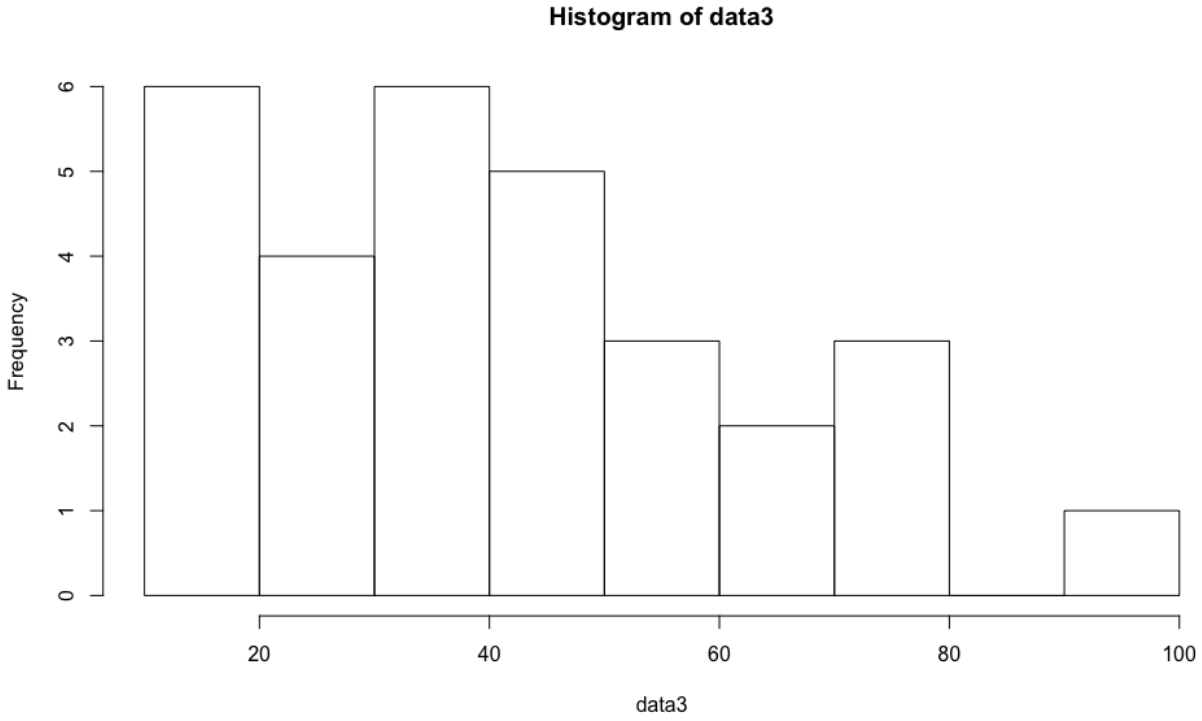
(c) An increasing trend of the absolute increments in Ex 8.6 corresponds to a decreasing trend of the relative increments in Ex 8.7. We should expect a rather strong negative correlation: large increments correspond to small relative increments, and vice versa.
Indeed, the correlation coefficient equals -0.93. A steady reduction of relative increments is not surprising. Had the population increased by a constant percent every decade, it would have grown exponentially fast. However, the human population does not grown at an exponential rate.
On the other hand, had the population increased by a constant number every decade, it would have grown linearly. The increments increase, therefore the population grows faster than a linear function. Thus, increasing absolute increments and decreasing relative increments show that the U.S. population growth is faster than linear but slower than exponential.

Exercise 8.8

(a) Histograms of the given data sets are shown on the next page. The first distribution is slightly left-skewed, the second is symmetric, and the third is right-skewed.


Histogram of data1
Frequency
0
2
4
6
8
0
5
10
15
20
25
30
data1


Histogram of data2
Frequency
0
2
4
6
8
10
12
10
15
20
25
30
35
data2



(b) Set 1: $\overline{X} = 14.9667$, $\widehat{M} = 15.5$. As expected for a left-skewed distribution, $\overline{X} < \widehat{M}$.
Set 2: $\overline{X} = 20.8333$, $\widehat{M} = 21.0$. As expected for a symmetric distribution, $\overline{X} \approx \widehat{M}$.
Set 3: $\overline{X} = 41.3$, $\widehat{M} = 39.5$. As expected for a right-skewed distribution, $\overline{X} > \widehat{M}$.

Additional Ex:

R Codes:

```
x=rnorm(5000,5,2)
quantile(x,0.975)
```

R Results:

```
97.5%
8.983331
```

The result is close to 8.92. It's not surprised, because 1.96 is the approximate value to the 97.5 percentile point of the standard normal distribution.

```
#####################################################################
# R code for HW 3 Exercises from Chapter 8                          #
#####################################################################

########
#8.1(b)#
########

before <- c(56, 47, 49, 37, 38, 60, 50, 43, 43, 59, 50, 56, 54, 58)
after <- c(53, 21, 32, 49, 45, 38, 44, 33, 32, 43, 53, 46, 36, 48, 39, 35, 37,
36, 39, 45)

# > (quantile(before, type = 1))
# 0%  25%  50%  75% 100%
# 37   43   50   56   60
# >

# > (quantile(after, type = 1))
# 0%  25%  50%  75% 100%
# 21   35   39   45   53
# >

boxplot(before, after, names = c("before", "after"))


########
#8.2(a)#
########

num.con.users <- c(17.2, 22.1, 18.5, 17.2, 18.6, 14.8, 21.7, 15.8, 16.3, 22.8,
24.1, 13.3, 16.2, 17.5, 19, 23.9, 14.8, 22.2, 21.7, 20.7, 13.5, 15.8, 13.1,
16.1, 21.9, 23.9, 19.3, 12, 19.9, 19.4, 15.4, 16.7, 19.5, 16.2, 16.9, 17.1,
20.2, 13.4, 19.8, 17.7, 19.7, 18.7, 17.6, 15.9, 15.2, 17.1, 15, 18.8, 21.6,
11.9)

# > (mean(num.con.users))
# [1] 17.954
# > (var(num.con.users))
# [1] 9.968249
# > (sd(num.con.users))
# [1] 3.157253
# >

########
#8.2(b)#
########

# > (sd(num.con.users)/sqrt(length(num.con.users)))
# [1] 0.4465031
```

```
# >

########
#8.2(c)#
########

# > (quantile(num.con.users))
# 0%    25%    50%    75%   100%
# 11.900 15.825 17.550 19.875 24.100
# >

boxplot(num.con.users)

########
#8.2(d)#
########

#compute the interquantile range

q1 <- quantile(num.con.users)[2]
q3 <- quantile(num.con.users)[4]

# > (iqr <- q3 - q1)
# 75%
# 4.05
# >

# Alternatively, use the IQR function:
# > (IQR(num.con.users))
# [1] 4.05
# >

#find the outliers if any
lower <- q1 - 1.5 * iqr
upper <- q3 + 1.5 * iqr

# > (num.con.users[num.con.users < lower | num.con.users > upper])
# numeric(0)
# >

########
#8.2(e)#
########

hist(num.con.users)

########
#8.5 #
########
```

```
year <- seq(1790, 2010, 10)
pop <- c(3.9, 5.3, 7.2, 9.6, 12.9, 17.1, 23.2, 31.4, 38.6, 50.2, 63, 76.2, 92.2,
106, 123.2, 132.2, 151.3, 179.3, 203.3, 226.5, 248.7, 281.4, 308.7)

plot(pop ~ year, xlab = "Year", ylab = "Population")

########
#8.6(a)#
########

increments <- pop[2:length(pop)] - pop[1:(length(pop) - 1)]

# > (mean(increments))
# [1] 13.85455
# > (median(increments))
# [1] 13
# > (var(increments))
# [1] 87.60355
# >

########
#8.6(b)#
########

plot(increments ~ year[-1], xlab = "Year", ylab = "Increments of population")

########
#8.7(a)#
########

increments.relative <- (pop[2:length(pop)] - pop[1:(length(pop) - 1)])/pop[1:(length(pop) -
1)]

# > (mean(increments.relative))
# [1] 0.2238006
# > (median(increments.relative))
# [1] 0.2097488
# > (var(increments.relative))
# [1] 0.01025038
# >

########
#8.7(b)#
########

plot(increments.relative ~ year[-1], xlab = "Year", ylab = "Relative population change")

########
#8.7(c)#
########
```

```
# > (cor(year[-1], increments.relative))
# [1] -0.9278865
# >

########
#8.8 #
########

#Data set 1

data1 <- c(19, 24, 12, 19, 18, 24, 8, 5, 9, 20, 13, 11, 1, 12, 11, 10, 22, 21,
7, 16, 15, 15, 26, 16, 1, 13, 21, 21, 20, 19)

hist(data1)

# > (mean(data1))
# [1] 14.96667
# > (median(data1))
# [1] 15.5
#>

#Data set 2

data2 <- c(17, 24, 21, 22, 26, 22, 19, 21, 23, 11, 19, 14, 23, 25, 26, 15, 17,
26, 21, 18, 19, 21, 24, 18, 16, 20, 21, 20, 23, 33)

hist(data2)

# > mean(data2)
# [1] 20.83333
# > median(data2)
# [1] 21
# >

#Data set 3

data3 <- c(56, 52, 13, 34, 33, 18, 44, 41, 48, 75, 24, 19, 35, 27, 46, 62, 71,
24, 66, 94, 40, 18, 15, 39, 53, 23, 41, 78, 15, 35)

hist(data3)

# > mean(data3)
# [1] 41.3
# > median(data3)
# [1] 39.5
# >
```