Mathematical Statistics

Assignment 1

Student ID: 2025280050
Student Name: Muhammad Zeeshan

Northwestern Polytechnical University

Question 1.4.3

The data in Table 1.9 are presented to illustrate the role of renewable energy consumption in the U.S. energy supply in 2007 (source: `http://www.eia.doe.gov/fuelrenewable.html`). Renewable energy consists of biomass, geothermal energy, hydroelectric energy, solar energy, and wind energy.

(a) Construct a bar graph.

(b) Construct a Pareto chart.

(c) Construct a pie chart.

Data

Table 1: Renewable Energy Consumption (U.S., 2007)

Source	**Percentage**
Coal	22%
Natural gas	23%
Nuclear electric power	8%
Petroleum	40%
Renewable energy	7%

(a) Bar Graph

A bar graph displays each energy source and its percentage visually.

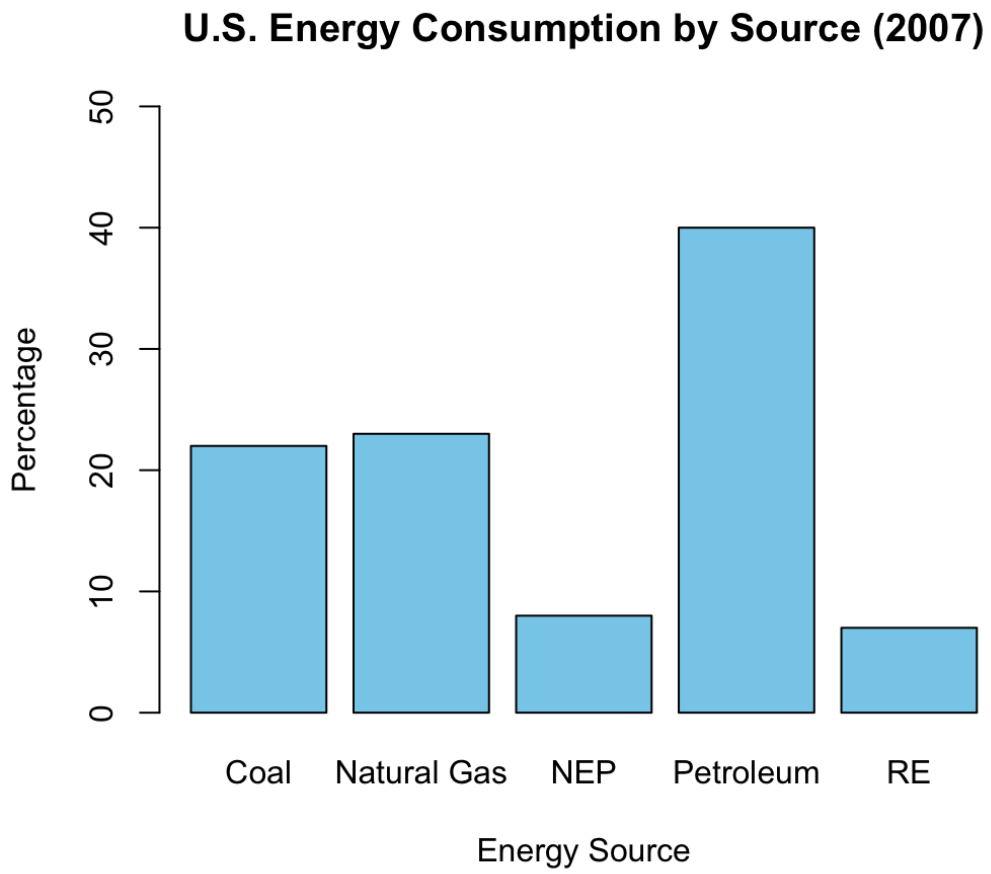

Figure 1: Bar graph of U.S. energy consumption by source (2007).

(b) Pareto Chart

The Pareto chart ranks the energy sources from largest to smallest contribution and shows the cumulative impact.

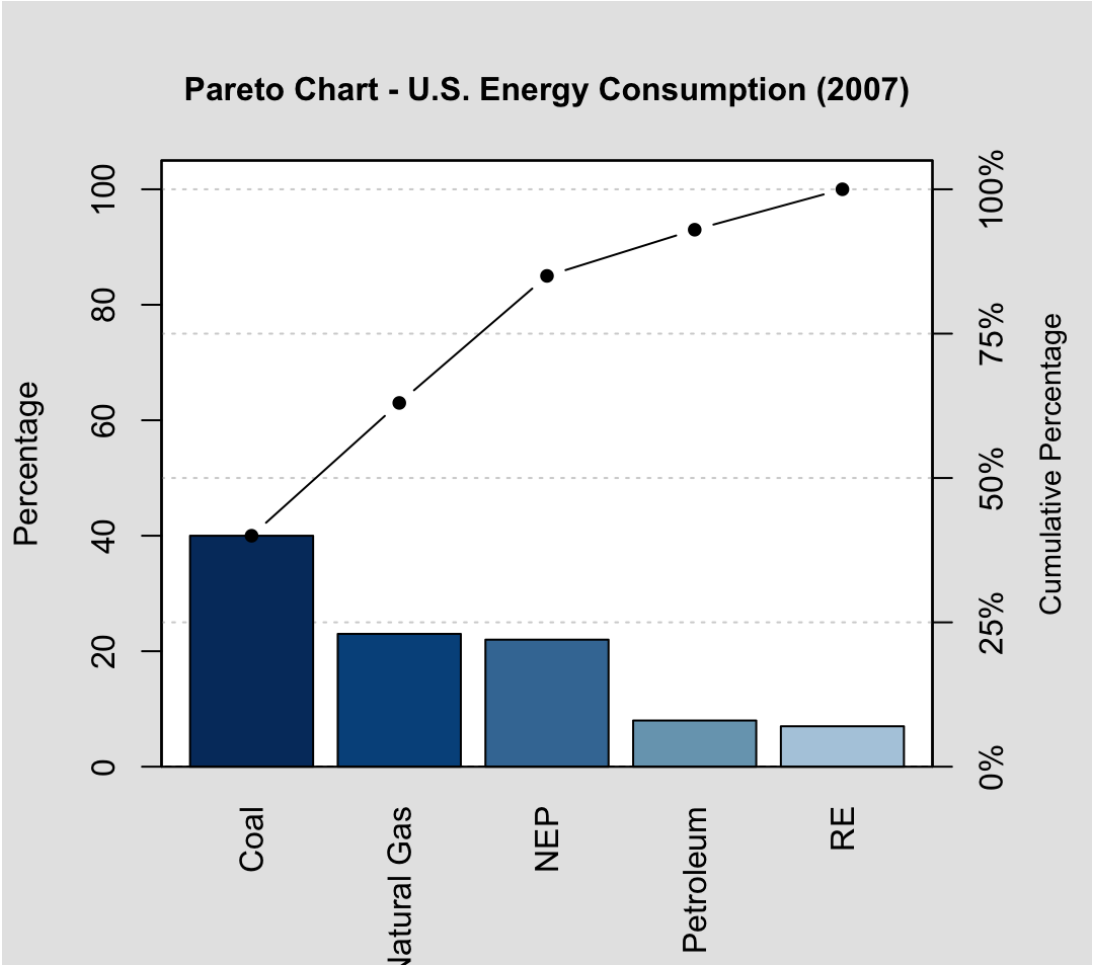

Figure 2: Pareto chart of energy consumption (2007).

(c) Pie Chart

The pie chart provides a proportional view of energy consumption distribution.

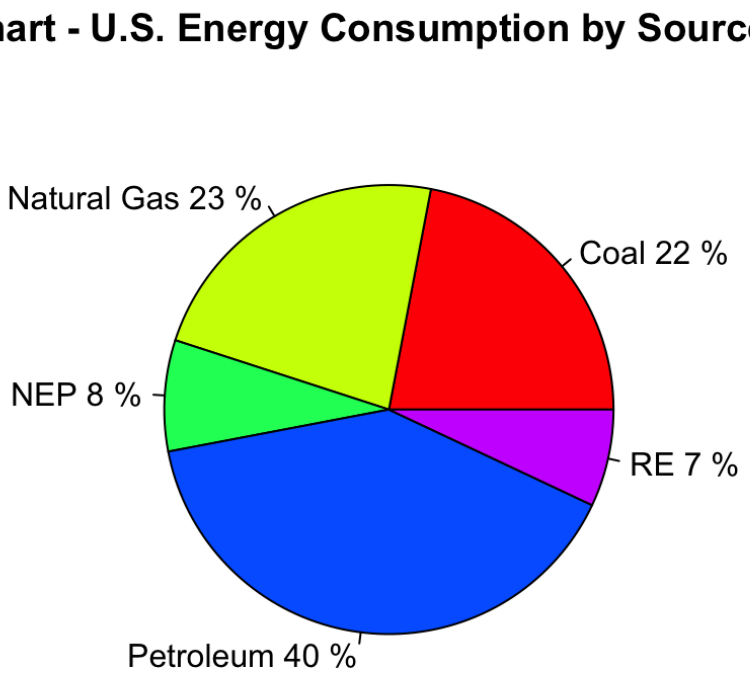

Figure 3: Pie chart of energy consumption (2007).

Conclusion

From the analysis, petroleum represents the largest share (40%) of U.S. energy consumption in 2007, followed by natural gas (23%) and coal (22%). Renewable energy accounts for 7%, highlighting its minor but important role in the total energy supply.

Question 1.4.14

The following table gives radon concentrations in pCi/liter (picocurie per liter) obtained from 40 houses in a certain area.

2.9	0.6	13.5	17.1	2.8	3.8	16.0	2.1	6.4	17.2
7.9	0.5	13.7	11.5	2.9	3.6	6.1	8.8	2.2	9.4
15.9	8.8	9.8	11.5	12.3	3.7	8.9	13.0	7.9	11.7
6.2	6.9	12.8	13.7	2.7	3.5	8.3	15.9	5.1	6.0

(a) Construct a stem-and-leaf display.

(b) Construct a frequency histogram and interpret.

(c) Construct a pie chart and interpret.

Solution

(a) Stem-and-leaf display

Using the integer part as the stem and the first decimal digit as the leaf, the stem-and-leaf display is:

```
 0 | 5 6
 2 | 1 2 7 8 9 9
 3 | 5 6 7 8
 5 | 1
 6 | 0 1 2 4 9
 7 | 9 9
 8 | 3 8 8 9
 9 | 4 8
11 | 5 5 7
12 | 3 8
13 | 0 5 7 7
15 | 9 9
16 | 0
17 | 1 2
```

Interpretation: the data range from 0.5 to 17.2, with many observations in the 2.x and 12–13.x ranges.

(b) Frequency histogram

I used bin width 2 (pCi/l) with breaks at 0,2,4,...,18. The frequency counts per bin are:

Bin (pCi/l)	Count
$[0, 2)$	2
$[2, 4)$	10
$[4, 6)$	1
$[6, 8)$	7
$[8, 10)$	6
$[10, 12)$	3
$[12, 14)$	6
$[14, 16)$	2
$[16, 18)$	3
Total	40

Interpretation: the modal bin is $[2, 4)$ (10 observations); most values lie between 2 and 14 pCi/L, with a handful in the higher bins (up to 17.2), suggesting a slight right tail.

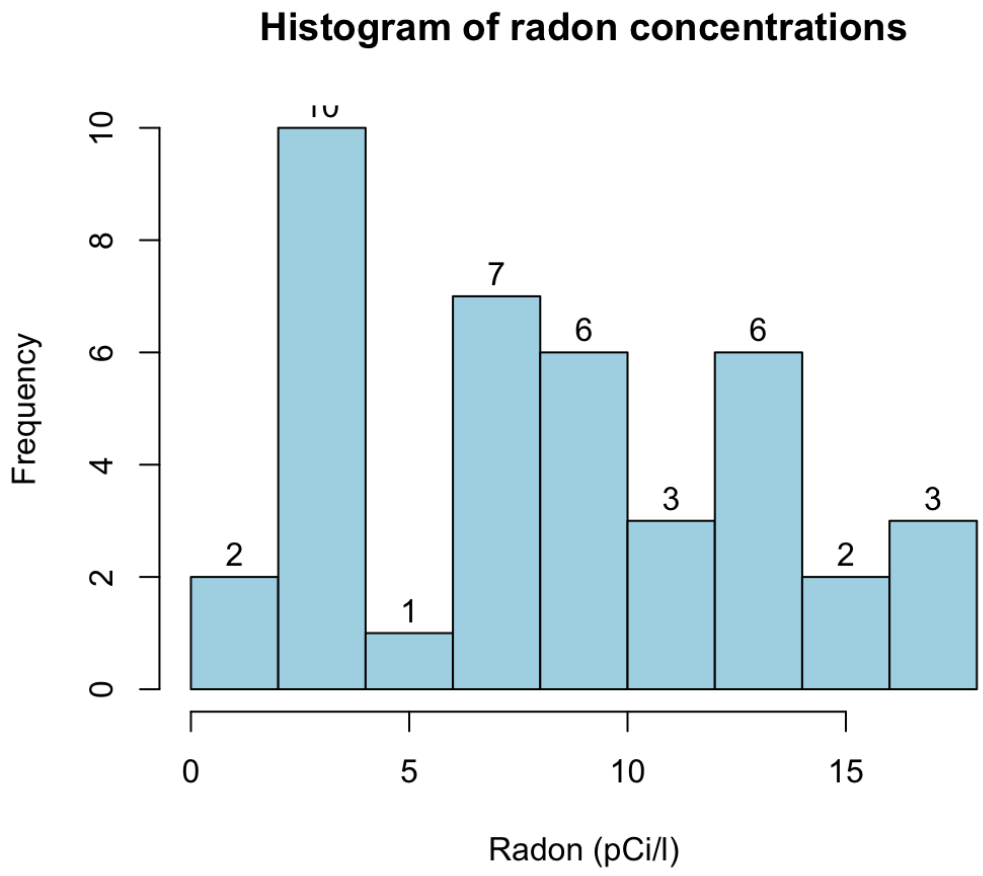

Figure 4: Histogram of radon concentrations.

(c) Pie chart

I grouped the data into five ranges: $[0, 4), [4, 8), [8, 12), [12, 16), [16, 18)$. Counts and percentages:

Range	Count	Percent
0–< 4	12	30.0%
4–< 8	8	20.0%
8–< 12	9	22.5%
12–< 16	8	20.0%
16–< 18	3	7.5%

Interpretation: the largest proportion (30%) is in the lowest range (0–4 pCi/L). The middle ranges (4–16) together contain most observations; only a small fraction (7.5%) are in 16–18 pCi/L.

(To include the actual pie chart image, save the R plot as a PNG/PDF and include with `\includegraphics{...}`)

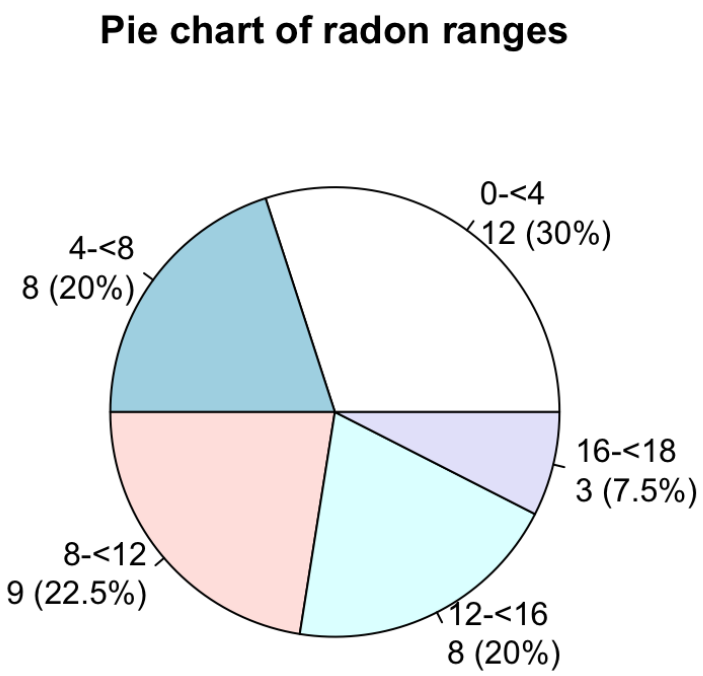

Figure 5: Pie chart of radon concentrations.

Question 1.5.5

Maximal static inspiratory pressure (PImax) is an index of respiratory muscle strength. The following data show the measure of PImax (cm H_2O) for 15 cystic fibrosis patients.

105	135	80	105	115	45	95	115
100	40	85	115	90	95	70	

(a) Find the lower and upper quartiles, median, and interquartile range. Check for any outliers and interpret.

(b) Construct a box plot and interpret.

(c) Are there any outliers?

Solution

(a) **Find the lower and upper quartiles, median, and interquartile range. Check for any outliers and interpret.**

First, arrange the data in ascending order:

$$40,\ 45,\ 70,\ 80,\ 85,\ 90,\ 95,\ 95,\ 100,\ 105,\ 105,\ 115,\ 115,\ 115,\ 135$$

The number of observations is $n = 15$.

$$Q_2 = \text{Median} = 8^{\text{th}} \text{ observation} = 95$$

The lower half (below the median) is:

$$40,\ 45,\ 70,\ 80,\ 85,\ 90,\ 95$$

So,

$$Q_1 = 4^{\text{th}} \text{ observation} = 80$$

The upper half (above the median) is:

$$100,\ 105,\ 105,\ 115,\ 115,\ 115,\ 135$$

Hence,

$$Q_3 = 4^{\text{th}} \text{ observation in upper half} = 115$$

Therefore, the interquartile range is:

$$IQR = Q_3 - Q_1 = 115 - 80 = 35$$

To check for outliers, we use:

$$\text{Lower Bound} = Q_1 - 1.5(IQR) = 80 - 1.5(35) = 27.5$$

$$\text{Upper Bound} = Q_3 + 1.5(IQR) = 115 + 1.5(35) = 167.5$$

Since all data values are between 40 and 135, there are **no outliers**.

Summary:

$$Q_1 = 80, \quad Q_2 = 95, \quad Q_3 = 115, \quad IQR = 35$$

(b) **Construct a box plot and interpret.**

The box plot will extend from the minimum value (40) to the maximum (135), with the box spanning from $Q_1 = 80$ to $Q_3 = 115$ and a median line at 95.

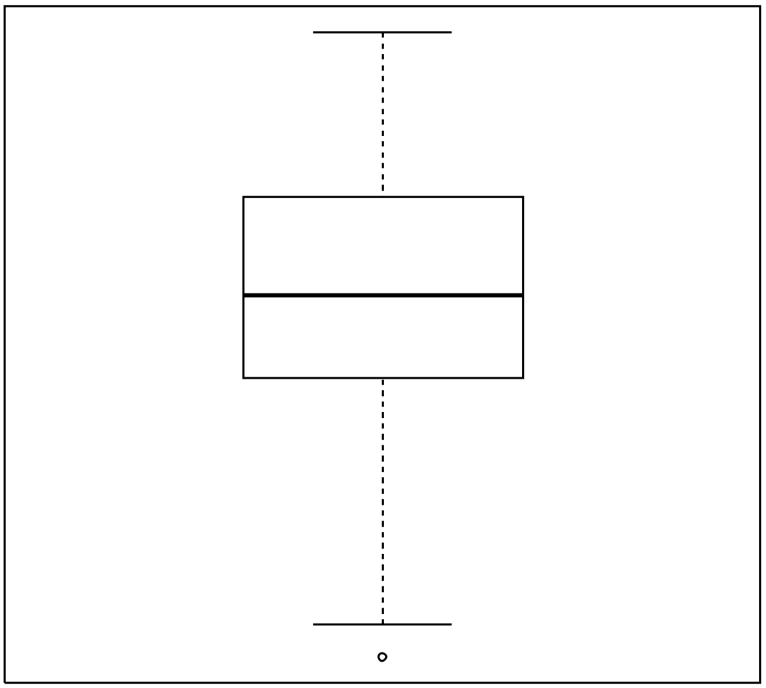

Figure 6: Box plot of PImax values for cystic fibrosis patients.

Interpretation: The distribution of PImax values is slightly right-skewed since the upper whisker (115–135) is longer than the lower whisker (40–80). Most patients have PImax values between 80 and 115 cm H_2O.

(c) **Are there any outliers?**

No, there are no outliers since all observations fall within the acceptable range of 27.5 to 167.5 cm H_2O.

Question 1.5.7

(a) For any grouped data with l classes with group frequencies f_i and class midpoints m_i, show that

$$\sum_{i=1}^{l} f_i(m_i - \bar{x}) = 0.$$

(b) Verify this result for the data given in *Exercise 1.5.6*.

Given: Classes 0–4, 5–9, 10–14, 15–19, 20–24 with frequencies

$$f_i = 5,\ 14,\ 15,\ 10,\ 6.$$

Class midpoints:

$$m_i = 2,\ 7,\ 12,\ 17,\ 22.$$

Solution

(a)

$$\sum_{i=1}^{l} f_i(m_i - \bar{x}) = \sum_{i=1}^{l} f_i m_i - \sum_{i=1}^{l} f_i \bar{x} = \sum_{i=1}^{l} f_i m_i - \bar{x} \sum_{i=1}^{l} f_i = n\bar{x} - n\bar{x} = 0,$$

since $\bar{x} = \dfrac{\sum_i f_i m_i}{\sum_i f_i}$ and $\sum_i f_i = n$.

(b)

Using the data from Exercise 1.5.6: $m_i = 2, 7, 12, 17, 22$, $f_i = 5, 14, 15, 10, 6$, and $\bar{x} = 11.8$.

$$\begin{aligned} 5(2 - 11.8) &= -49.0, \\ 14(7 - 11.8) &= -67.2, \\ 15(12 - 11.8) &= 3.0, \\ 10(17 - 11.8) &= 52.0, \\ 6(22 - 11.8) &= 61.2, \end{aligned}$$

and their sum equals 0, verifying the identity numerically.

Question 1.5.10

The radon concentration (in pCi/liter) data obtained from 40 houses in a certain area are given below.

2.9	0.6	13.5	17.1	2.8	3.8	16.0	2.1	6.4	17.2
7.9	0.5	13.7	11.5	2.9	3.6	6.1	8.8	2.2	9.4
15.9	8.8	9.8	11.5	12.3	3.7	8.9	13.0	7.9	11.7
6.2	6.9	12.8	13.7	2.7	3.5	8.3	15.9	5.1	6.0

(a) Find the mean, variance, and range for these data.

(b) Find lower and upper quartiles, median, and interquartile range. Check for any outliers.

(c) Construct a box plot.

(d) Construct a histogram and interpret.

(e) Locate on your histogram $\bar{x}$, $\bar{x} \pm s$, $\bar{x} \pm 2s$, and $\bar{x} \pm 3s$. Count the data points in each of the intervals $\bar{x} \pm s$, $\bar{x} \pm 2s$, and $\bar{x} \pm 3s$. How do these counts compare with the empirical rule?

Solution

(a) Mean, variance, standard deviation, range

Number of observations: $n = 40$. Sum of data $\sum x = 333.6$.

$$\bar{x} = \frac{\sum x}{n} = \frac{333.6}{40} = 8.34.$$

Sum of squared deviations: $\sum (x_i - \bar{x})^2 = 944.376$. Sample variance (using $n - 1$):

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} = \frac{944.3759999999999}{39} \approx 24.21476923.$$

Sample standard deviation:

$$s = \sqrt{s^2} \approx 4.92085046.$$

Range:

$$\max(x) - \min(x) = 17.2 - 0.5 = 16.7.$$

(b) Quartiles, median, IQR, outlier check

Using interpolation (R type = 7) we obtain

$$Q_1 = 3.675, \qquad Q_2 = 8.10, \qquad Q_3 = 12.425.$$

$$\text{IQR} = Q_3 - Q_1 = 8.75.$$

1.5·IQR fences:

$$\text{Lower fence} = Q_1 - 1.5 \cdot \text{IQR} = -9.45, \qquad \text{Upper fence} = Q_3 + 1.5 \cdot \text{IQR} = 25.55.$$

All observations lie in $[0.5, 17.2] \subset [-9.45, 25.55]$, so there are no outliers by the 1.5·IQR rule.

(c) Boxplot

Construct the box from $Q_1 = 3.675$ to $Q_3 = 12.425$ with median at 8.10. Whiskers extend to the minimum 0.5 and maximum 17.2 (no outliers). The sample mean $\bar{x} = 8.34$ may be plotted as a point for comparison. Interpretation: center ≈ 8.1, IQR $= 8.75$, slight right skew; no outliers.

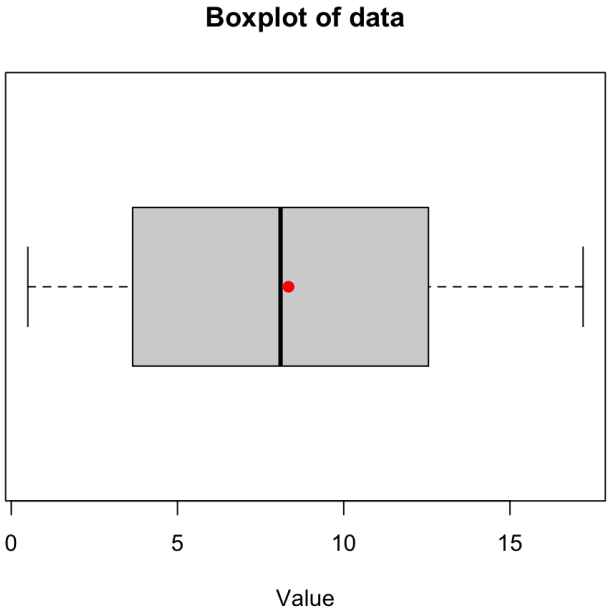

Figure 7: Box plot of radon concentration data.

(d) Histogram and interpretation

Plot a histogram (e.g. using Sturges' rule) and add vertical lines for the mean $\bar{x} = 8.34$ and for $\bar{x} \pm ks$ $(k = 1, 2, 3)$ to visualize spread.

Interpretation: the bulk of observations lies roughly between 3 and 13. The distribution shows a slight right skew (mean slightly larger than median). No extreme outliers are present.

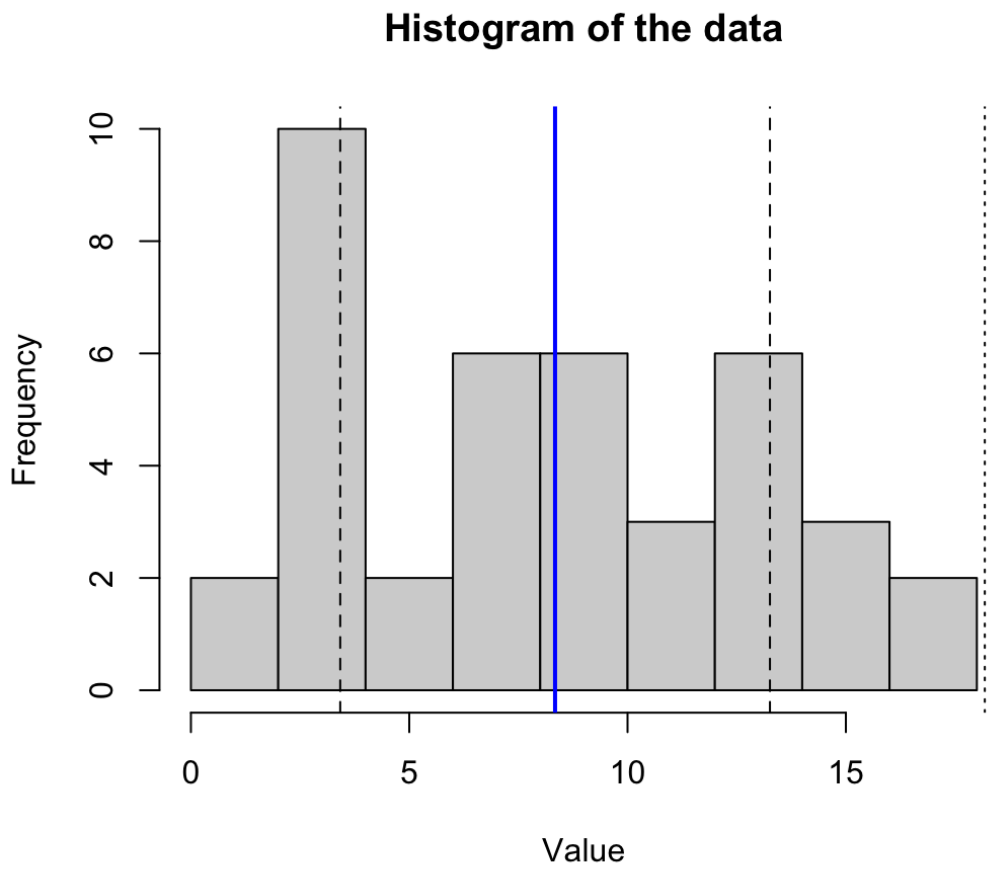

Figure 8: Histogram of radon concentration data with mean and standard deviation markers.

(e) Empirical-rule intervals and counts

$$\bar{x} = 8.34, \qquad s \approx 4.92085046.$$

$$\bar{x} \pm s \approx [3.41915,\ 13.26085], \quad \bar{x} \pm 2s \approx [-1.5017,\ 18.1817], \quad \bar{x} \pm 3s \approx [-6.4226,\ 23.1026].$$

Counts:

within $\bar{x}\pm s$: $24/40 = 60.0\%$, within $\bar{x}\pm 2s$: $40/40 = 100\%$, within $\bar{x}\pm 3s$: $40/40 = 100\%$.

Comparison with normal empirical rule (68%, 95%, 99.7%): the sample has slightly fewer points within 1s and more within 2s and 3s than a normal distribution would predict.
