Statistics 6 Review Exercise 2

Analysis of Cancer Patient Survival Times

Lucas Liona

Table of contents

Problem

A medical trial was conducted to investigate whether a new drug extended the life of a patient who had lung cancer. The survival times (in months) for 38 cancer patients who were treated with the drug are as follows:

5, 9, 10, 13, 14, 18, 18, 19, 20, 22, 25, 25, 25, 27, 28, 30, 30, 33, 36, 38, 39, 39, 40, 41, 41, 43, 44, 44, 45, 46, 46, 49, 50, 50, 54, 54, 59

The sample mean is approximately 31.1 months and the standard deviation is approximately 16.0 months.

Part (a): Normal Distribution Calculations

Assume that the survival time (in months) for patients on this drug is Normally distributed with a mean of 31.1 months and a standard deviation of 16.0 months.

(i) Probability of Surviving No More Than One Year

We need to calculate $P(X \leq 12)$, where $X \sim \text{Normal}(\mu = 31.1, \sigma = 16.0)$.

```r
# Parameters
mu <- 31.1
sigma <- 16.0

# Calculate P(X   12)
prob_less_than_12 <- pnorm(12, mean = mu, sd = sigma)
prob_less_than_12
```

```
[1] 0.1162879
```

The probability that a patient survives for no more than one year (12 months) is approximately 0.1163 or 11.63%.

(ii) Proportion Surviving Between One and Two Years

We need to calculate $P(12 < X \leq 24) = P(X \leq 24) - P(X \leq 12)$.

```r
# Calculate P(X   24)
prob_less_than_24 <- pnorm(24, mean = mu, sd = sigma)

# Calculate P(12 < X   24)
prob_between_12_and_24 <- prob_less_than_24 - prob_less_than_12
prob_between_12_and_24
```

```
[1] 0.2123238
```

The proportion of patients who survive between one year (12 months) and two years (24 months) is approximately 0.2123 or 21.23%.

(iii) 80th Percentile of Survival Times

We need to find the value a such that $P(X \leq a) = 0.8$.

```r
# Calculate the 80th percentile
percentile_80 <- qnorm(0.8, mean = mu, sd = sigma)
percentile_80
```

```
[1] 44.56594
```

The highest number of months that 80% of patients survive is approximately 17.63 months. This means that 80% of patients survive 17.63 months or less.

(iv) Central 80% of Survival Times

We need to find the 10th and 90th percentiles, which represent the central 80% of the distribution.

```
# Calculate the 10th percentile
percentile_10 <- qnorm(0.1, mean = mu, sd = sigma)

# Calculate the 90th percentile
percentile_90 <- qnorm(0.9, mean = mu, sd = sigma)

# Output the interval
central_80_percent <- c(percentile_10, percentile_90)
central_80_percent
```

```
[1] 10.59517 51.60483
```

The central 80% of survival times falls between 10.60 months and 51.60 months.

Part (b): Stem-and-Leaf Plot

Let's create a stem-and-leaf plot of the survival times.

```
# Input the data
survival_times <- c(5, 9, 10, 13, 14, 18, 18, 19, 20, 22, 25, 25, 25, 27, 28,
                    30, 30, 33, 36, 38, 39, 39, 40, 41, 41, 43, 44, 44, 45,
                    46, 46, 49, 50, 50, 54, 54, 59)

# Create a stem-and-leaf plot
stem(survival_times, scale = 1)
```

```

  The decimal point is 1 digit(s) to the right of the |
  0 | 59
  1 | 034889
  2 | 0255578
  3 | 0036899
  4 | 0113445669
  5 | 00449
```

The stem-and-leaf plot provides a visual representation of the distribution of survival times. Each stem (the digits to the left of the vertical line) represents the tens place, and each leaf (the digits to the right) represents the ones place of the survival times.

For a more readable version, we can create a manual stem-and-leaf plot:

```r
# Create vectors for each stem
stem0 <- survival_times[survival_times < 10]
stem1 <- survival_times[survival_times >= 10 & survival_times < 20]
stem2 <- survival_times[survival_times >= 20 & survival_times < 30]
stem3 <- survival_times[survival_times >= 30 & survival_times < 40]
stem4 <- survival_times[survival_times >= 40 & survival_times < 50]
stem5 <- survival_times[survival_times >= 50 & survival_times < 60]

# Extract the ones place for each value
leaf0 <- stem0 %% 10
leaf1 <- stem1 %% 10
leaf2 <- stem2 %% 10
leaf3 <- stem3 %% 10
leaf4 <- stem4 %% 10
leaf5 <- stem5 %% 10

# Sort the leaves
leaf0 <- sort(leaf0)
leaf1 <- sort(leaf1)
leaf2 <- sort(leaf2)
leaf3 <- sort(leaf3)
leaf4 <- sort(leaf4)
leaf5 <- sort(leaf5)

# Display the stem-and-leaf plot
cat("Stem | Leaf\n")
```

```
Stem | Leaf
```

```r
cat("----------------\n")
```

```
----------------
```

```r
cat("   0 |", paste(leaf0, collapse = " "), "\n")
```

```
   0 | 5 9
```

```r
cat("   1 |", paste(leaf1, collapse = " "), "\n")
```

```
   1 | 0 3 4 8 8 9
```

```r
cat("   2 |", paste(leaf2, collapse = " "), "\n")
```

```
   2 | 0 2 5 5 5 7 8
```

```r
cat("   3 |", paste(leaf3, collapse = " "), "\n")
```

```
   3 | 0 0 3 6 8 9 9
```

```r
cat("   4 |", paste(leaf4, collapse = " "), "\n")
```

```
   4 | 0 1 1 3 4 4 5 6 6 9
```

```r
cat("   5 |", paste(leaf5, collapse = " "), "\n")
```

```
   5 | 0 0 4 4 9
```

```r
cat("----------------\n")
```

```
----------------
```

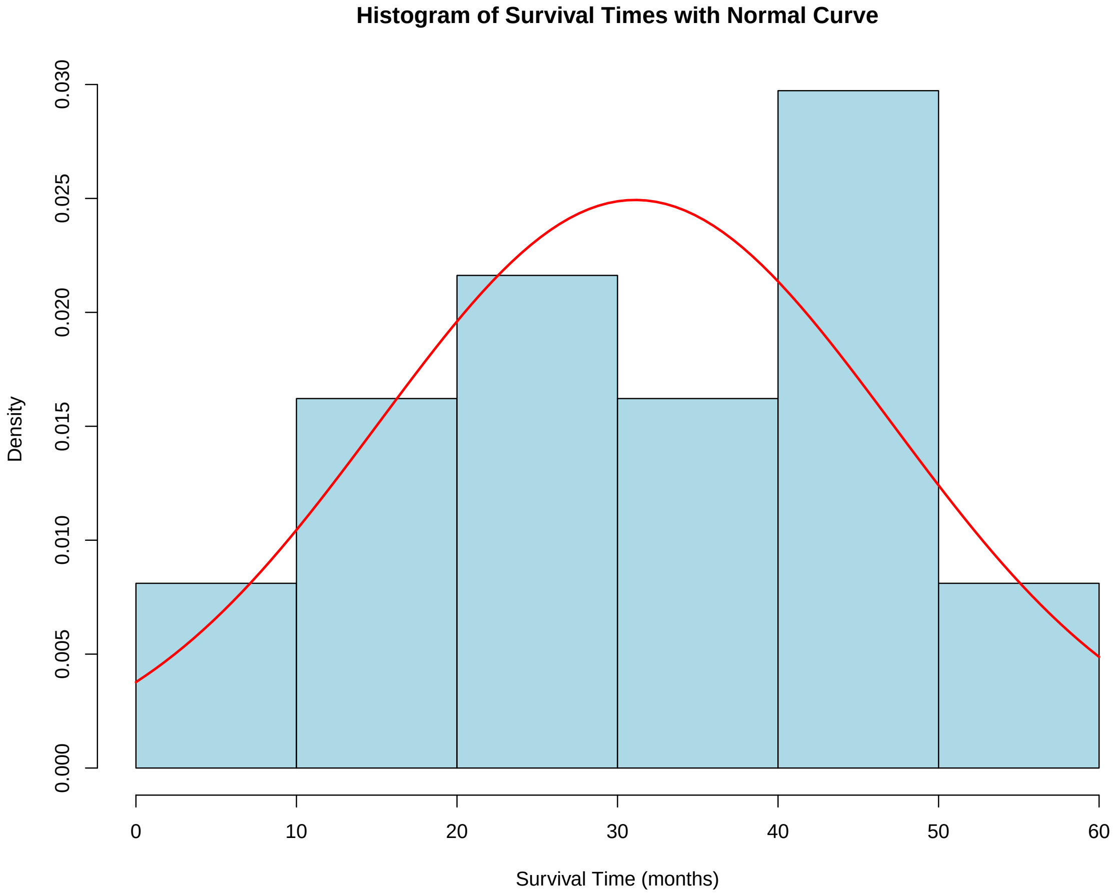
```r
cat("Units: 1|0 = 10\n")
```

```
Units: 1|0 = 10
```

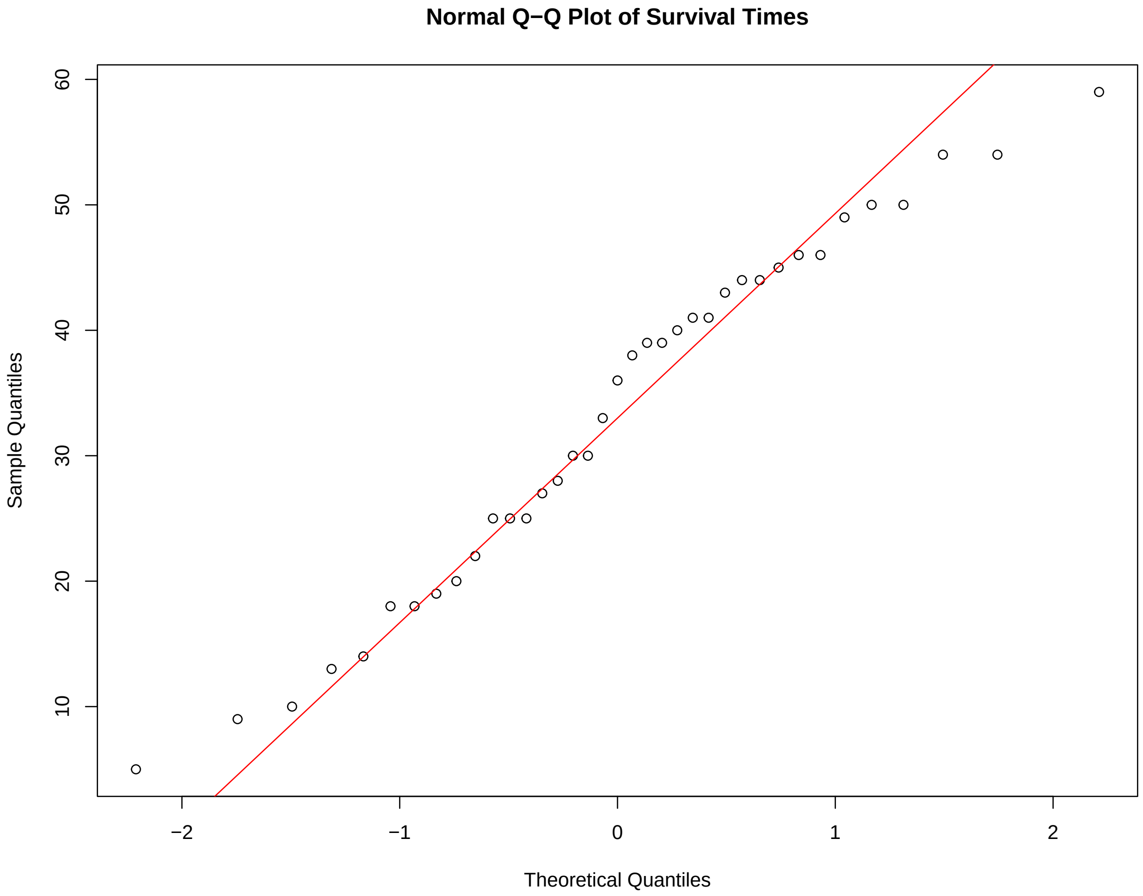
Part (c): Assessing Normality

To visually assess whether the data follows a normal distribution, let's create additional plots.

```r
# Histogram with normal curve overlay
hist(survival_times, freq = FALSE, main = "Histogram of Survival Times with Normal Curve",
     xlab = "Survival Time (months)", col = "lightblue")
curve(dnorm(x, mean = mu, sd = sigma), add = TRUE, col = "red", lwd = 2)
```



```r
# QQ-plot
qqnorm(survival_times, main = "Normal Q-Q Plot of Survival Times")
qqline(survival_times, col = "red")
```



Looking at the stem-and-leaf plot, histogram, and Q-Q plot, we can assess whether the data appears to follow a normal distribution:

1. The stem-and-leaf plot shows that the data is somewhat bimodal, with peaks around 18-25 months and 44-50 months, rather than a single central peak as would be expected in a normal distribution.
2. The histogram reinforces this observation, showing a distribution that doesn't have the classic bell shape of a normal distribution.
3. The Q-Q plot shows deviations from the diagonal line, particularly at the tails, suggesting departures from normality.

These observations suggest that the assumption of normality may not be valid for this data. The distribution appears to be bimodal rather than bell-shaped, which could indicate that there are actually two distinct groups of patients with different response patterns to the drug.

This finding could be clinically significant and might warrant further investigation into potential factors that could explain these different survival patterns.