Analysis of fev data

This program assesses the normality of variables in a study of pulmonary function in children. There is a data dictionary that provides more details about the data. The program was written by Steve Simon on 2024-09-02 and is placed in the public domain.

Libraries

The tidyverse library is the only one you need for this program.

```
        library(tidyverse)
```

List variable names

Since the variable names are not listed in the data file itself, you need to list them here.

```
        fev_names <- c(
            "age",
            "fev",
            "ht",
            "sex",
            "smoke")
```

Reading the data

Here is the code to read the data and show a glimpse.

```
        fev <- read_csv(
          file="../data/fev.csv",
          col_names=fev_names,
          col_types="nnncc")
        glimpse(fev)
```

```
Rows: 327
Columns: 5
$ age   <dbl> 9, 7, 9, 6, 8, 6, 8, 8, 5, 9, 5, 4, 9, 9, 8, 5, 8, 5, 9, 8, 9, 5…
$ fev   <dbl> 1.708, 1.720, 1.895, 1.919, 1.987, 1.602, 2.193, 2.258, 1.472, 2…
$ ht    <dbl> 57.0, 54.5, 57.0, 58.0, 58.5, 53.0, 58.5, 58.0, 50.0, 59.0, 49.0…
$ sex   <chr> "F,N\n8,1.724,67.5,F", "F,N\n9,1.558,53,M", "M,N\n8,2.336,61,F",…
$ smoke <chr> "N", "N", "N", "N", "N", "N", "N", "N", "N", "N", "N", "N", "N",…
```

Calculate mean and standard deviation for fev

To orient yourself to the data, calculate a few descriptive statistics.

```
fev |>
  summarize(
    fev_mean=mean(fev),
    fev_stdv=sd(fev))
```

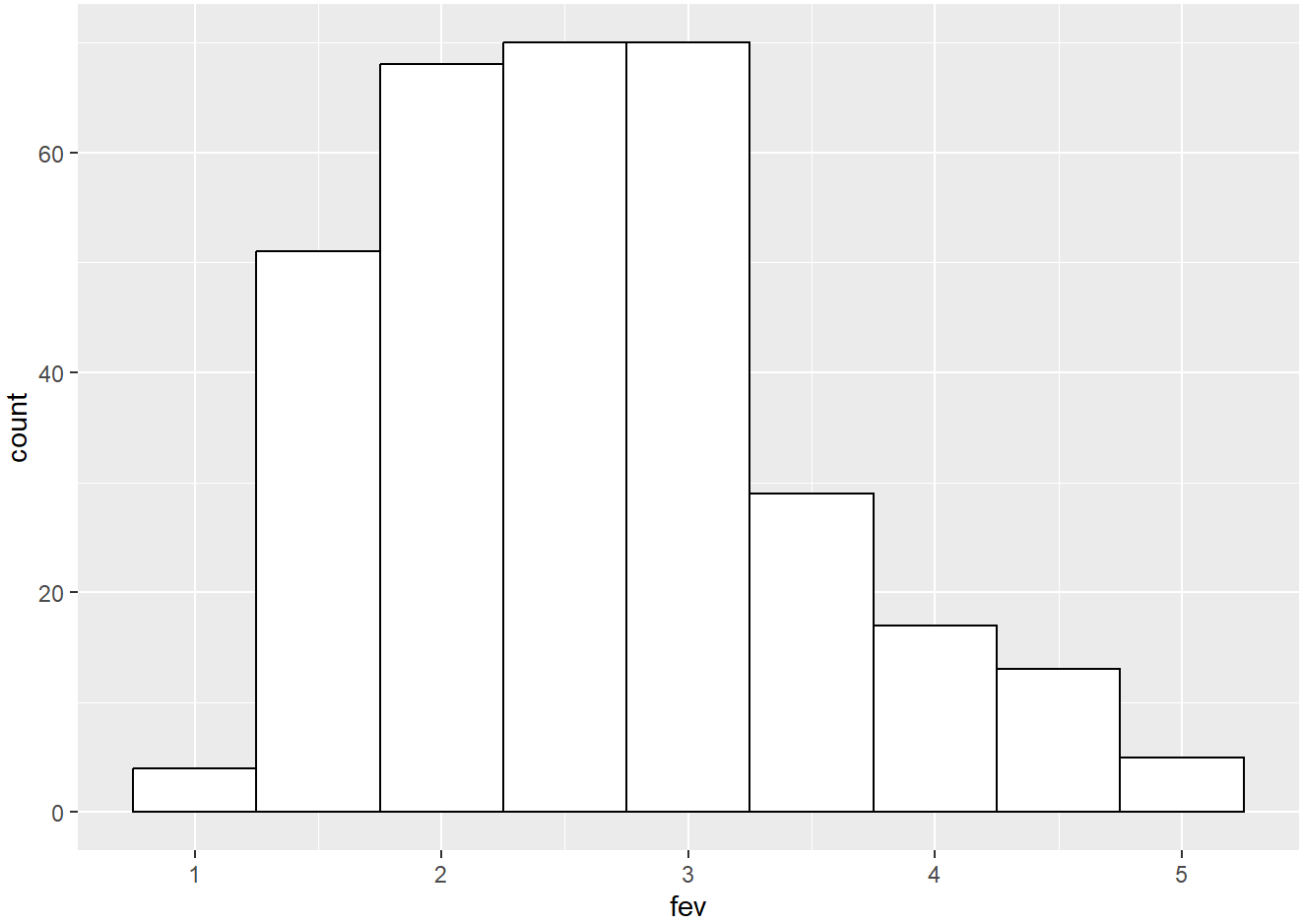
```
# A tibble: 1 × 2
  fev_mean fev_stdv
     <dbl>    <dbl>
1     2.62    0.844
```

The mean fev is 2.6 liters and the standard deviation is 0.84 liters. I am not an expert on pulmonary function, but these values do not appear to be unusually large or small.

Histogram for fev, wide bars

The geom_histogram function draws a histogram. You should specify values for color (which is the outline of individual bars) and fill (which is what is inside the bars). Also be careful with your choice of binwidth. Don't rely on the default choice.

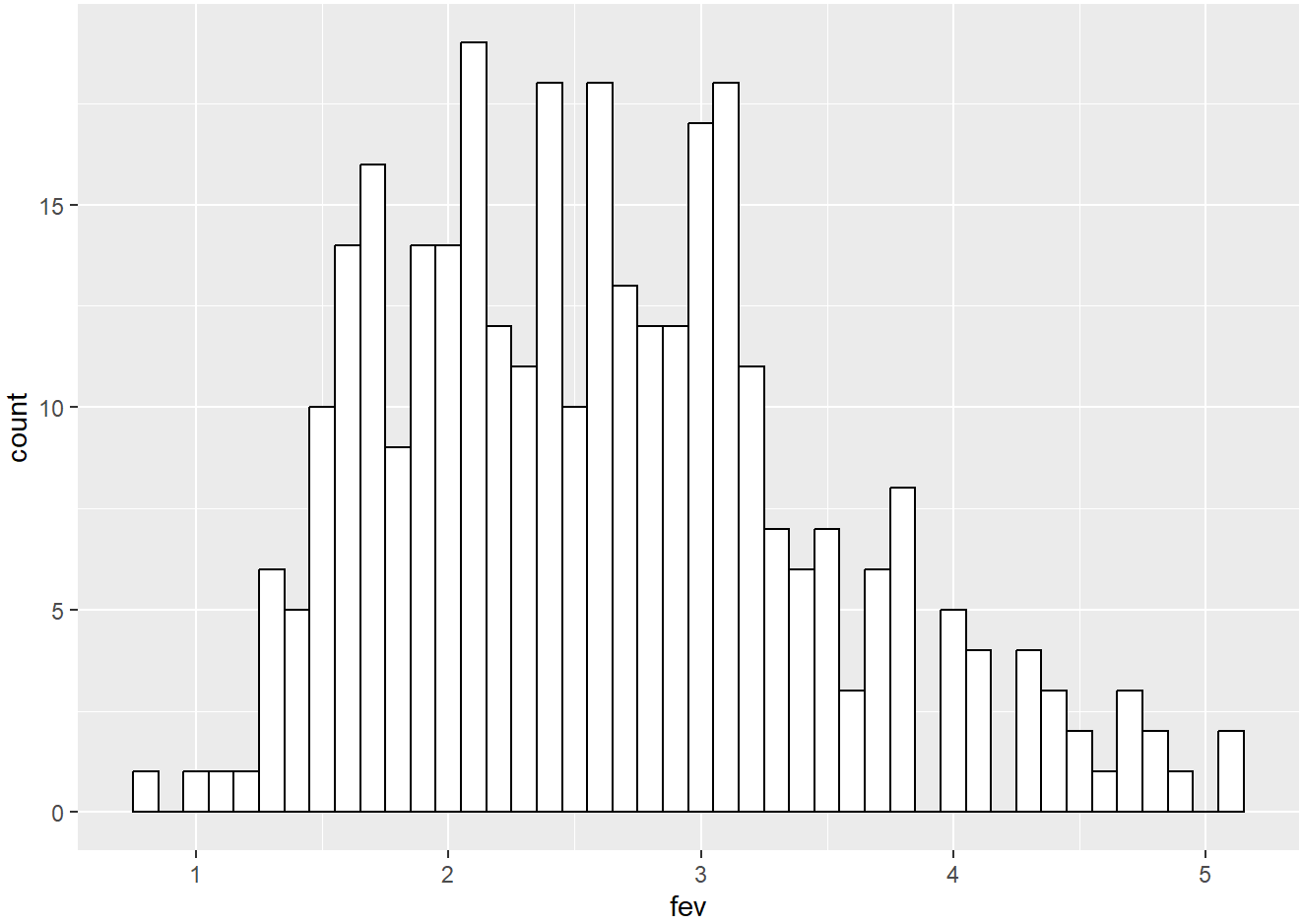
```
ggplot(data=fev, aes(x=fev)) +
  geom_histogram(
    binwidth=0.5,
    color="black",
    fill="white")
```



See below for interpretation

Histogram for fev, narrow bars

```
ggplot(data=fev, aes(x=fev)) +
  geom_histogram(
    binwidth=0.1,
    color="black",
    fill="white")
```

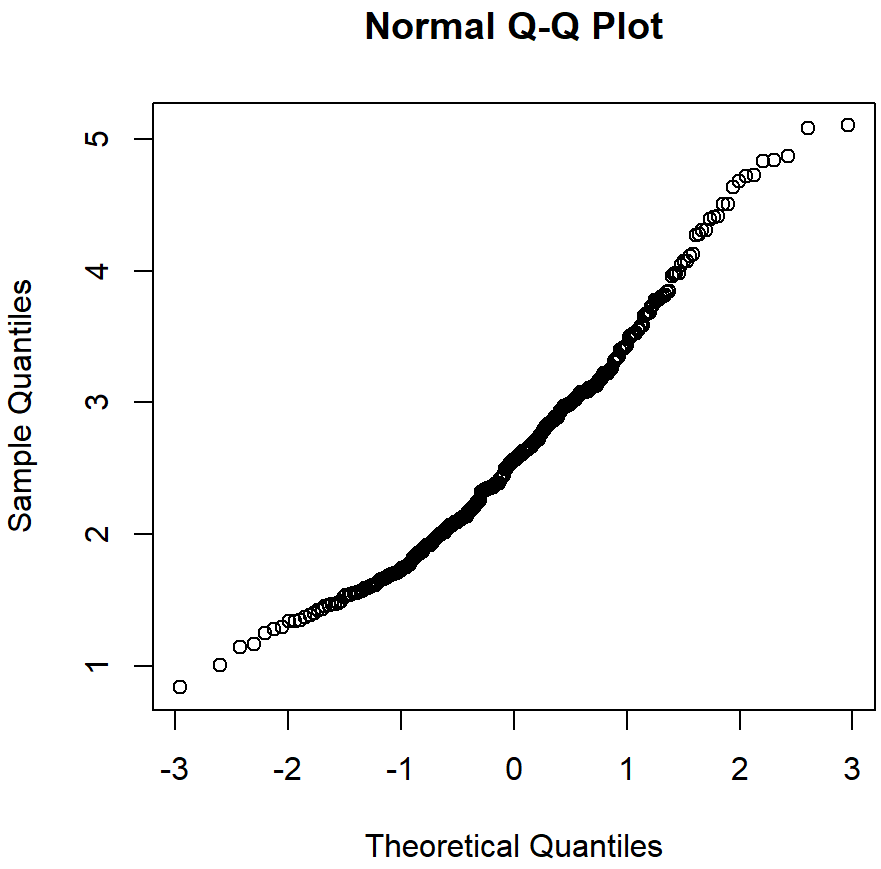


Although some may interpret these histograms as showing a slight skewness, I would interpret them as being approximately normal.

Normal probability plot for fev

The qqnorm function produces a normal probability plot. The default option for most plots is landscape orientation (the width is larger than the height). The q-q plot, however, looks best if figure width and height are equal.

```
qqnorm(fev$fev)
```



The normal probability plot is reasonably close to a straight line, indicating that the data comes reasonably close to following a normal distribution.

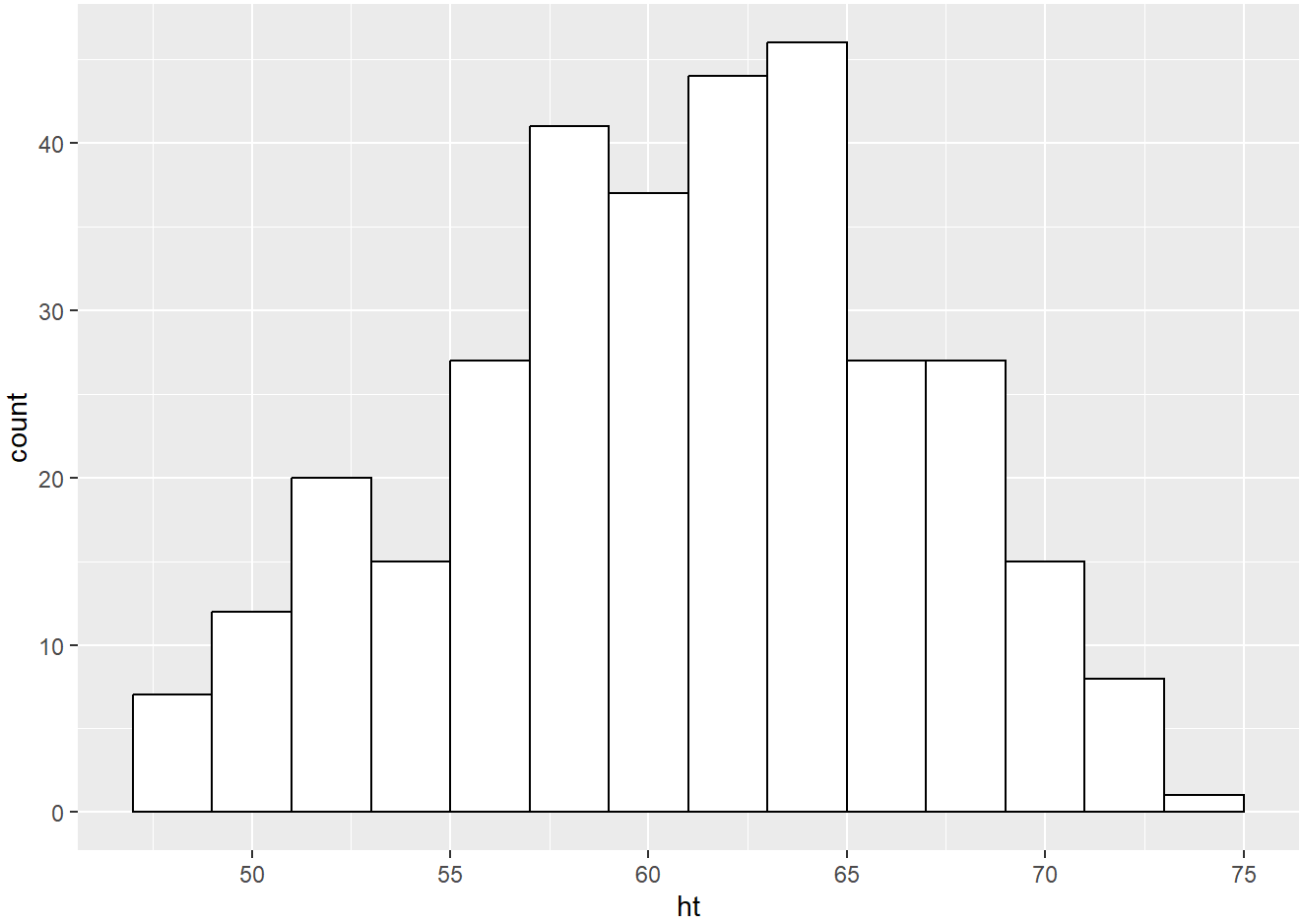
Question 3

```
fev |>
  summarize(
    fev_mean=mean(ht),
    fev_stdv=sd(ht))
```

```
# A tibble: 1 × 2
  fev_mean fev_stdv
     <dbl>    <dbl>
1     61.1     5.72
```

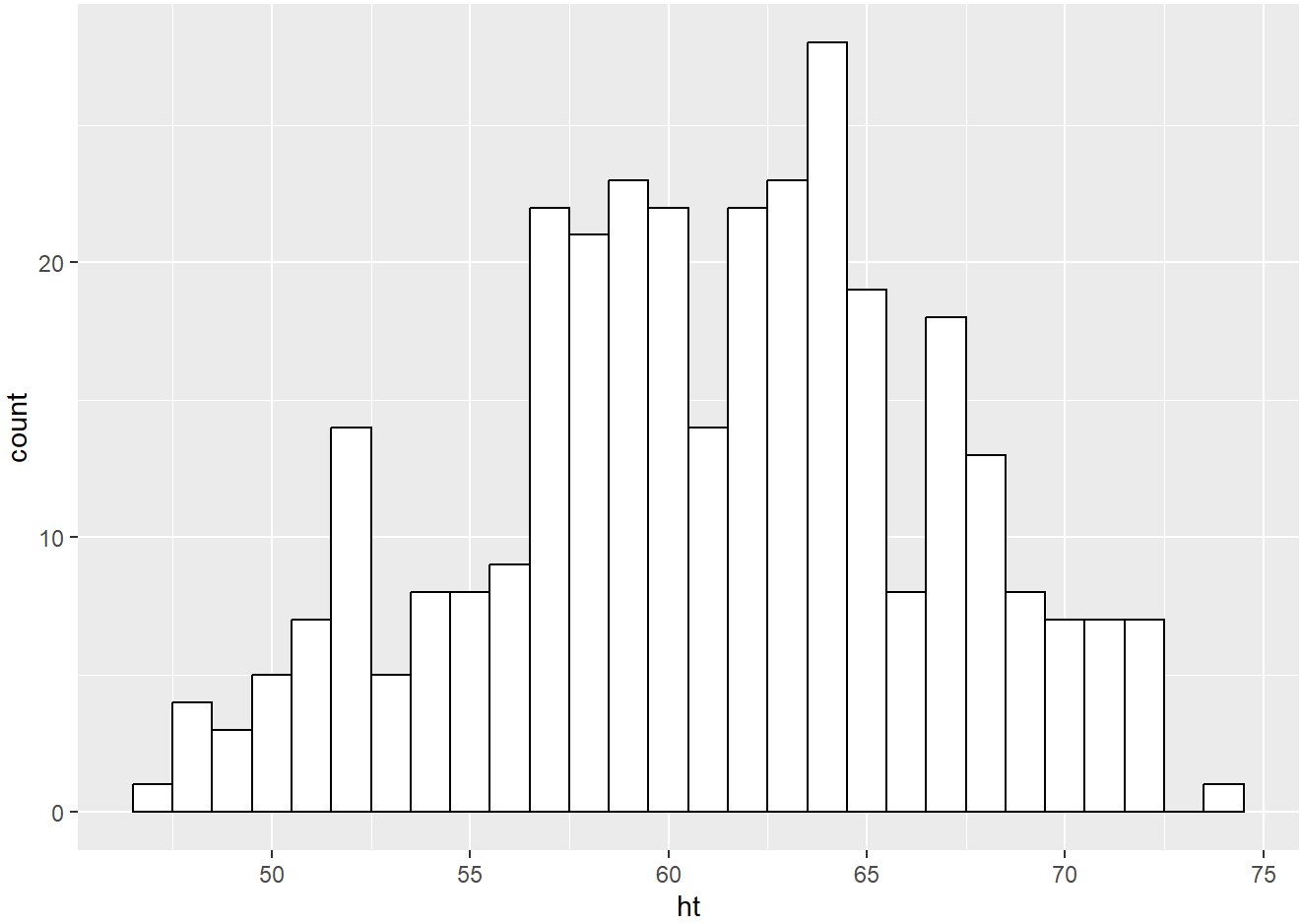
Histogram for ht, wide bars

```
ggplot(data=fev, aes(x=ht)) +
  geom_histogram(
    binwidth=2,
    color="black",
    fill="white")
```



Histogram for ht, narrow bars

```
ggplot(data=fev, aes(x=ht)) +
  geom_histogram(
    binwidth=1,
    color="black",
    fill="white")
```

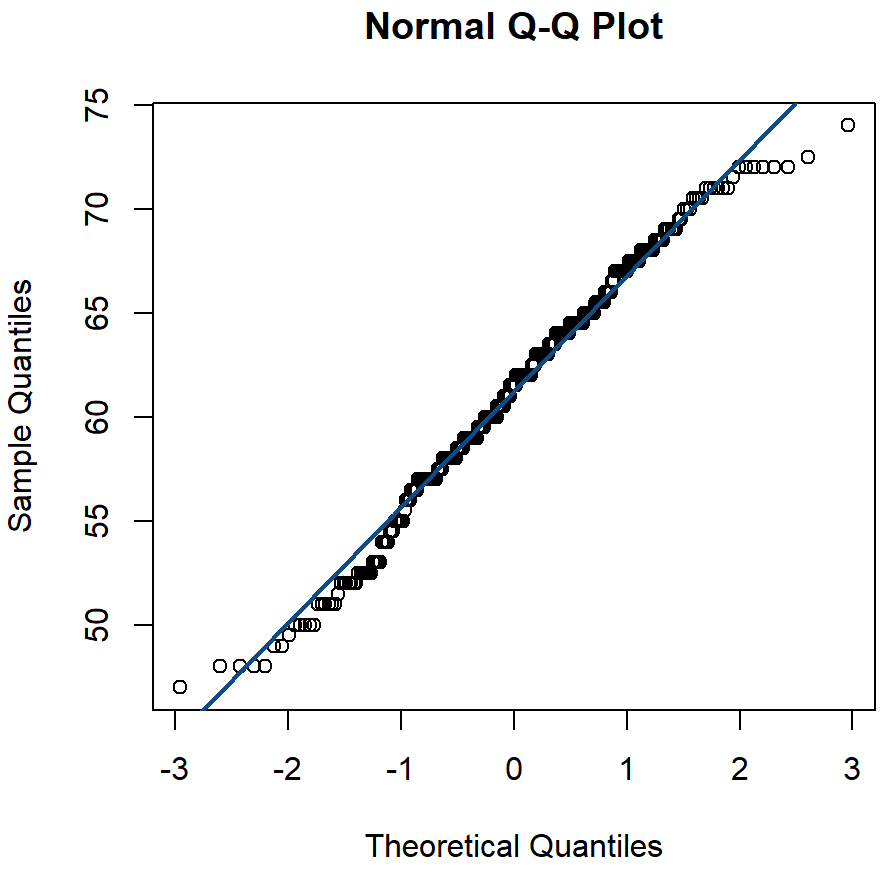


Both histograms are similar.

Normal probability plot for ht

Question 4

```
qqnorm(fev$ht)
qqline(fev$ht, col = "dodgerblue4", lwd = 2)
```



Most of the data points fall along a straight line, which suggests that the data is approximately normally distributed. However, some points fall below the line, indicating that the data has fewer extreme values than expected.