Housefly wing lengths

Dilip Nikhil Francies

28th September

The data

What does the distribution of housefly wing lengths (in mm) look like? This was apparently an important question in 1955, and the data set we look at is a famous one, from Sokal & Hunter, reproduced here:

https://seattlecentral.edu/qelp/sets/057/057.html

The data set can be scanned in from the web:

```
wing.length <- scan("https://seattlecentral.edu/qelp/sets/057/s057.txt")
```

This reads the data in as a vector. To use `ggplot()`, we need the data in a data frame. The following code creates a data frame called `wings.df` that contains a variable called `Length`:

```
wings.df <- data.frame(Length = wing.length)
```

We also need to load the `ggplot2` package to use the functions in it:

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

Summary statistics

Find the sample mean:

```
mean(wings.df$Length)
```

```
## [1] 45.5
```

The sample mean fly wing length is 45.5 millimeters.

Find the plug-in standard deviation and the sample standard deviation:

```
# Plug-in
sqrt(mean(wings.df$Length^2) - mean(wings.df$Length)^2)
```

```
## [1] 3.9
```

```
# Sample SD
sd(wings.df$Length)
```

```
## [1] 3.919647
```

There's hardly any difference. To one decimal place, the standard deviation of fly wing lengths is 3.9 millimeters.

We can also describe the data using the five-number summary:

```
summary(wings.df)
```

```
##      Length
##  Min.   :36.0
##  1st Qu.:43.0
##  Median :45.5
##  Mean   :45.5
##  3rd Qu.:48.0
##  Max.   :55.0
```

```r
IQR(wings.df$Length)
```

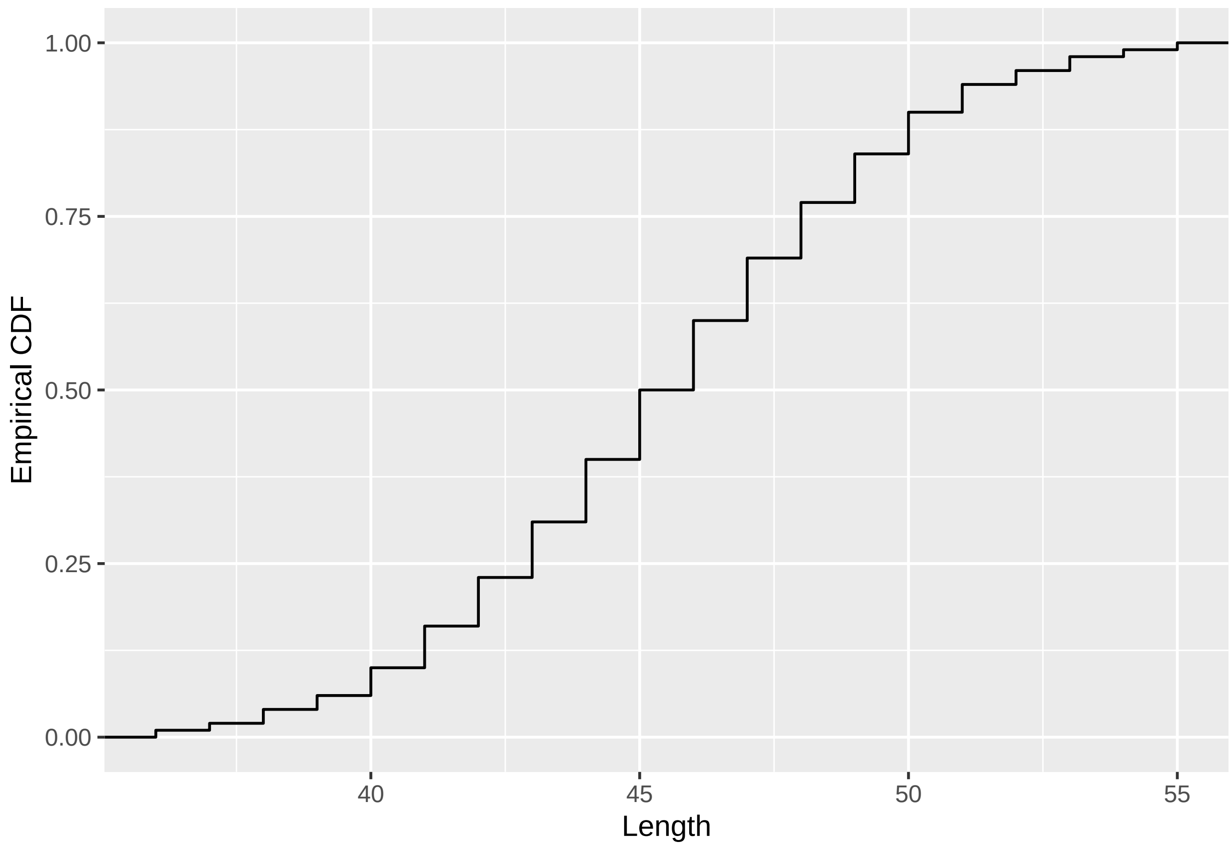
```
## [1] 5
```

The median fly wing length is 45.5 millimeters. The interquartile range is (q3-q1) = 5 millimeters.

Plot the data

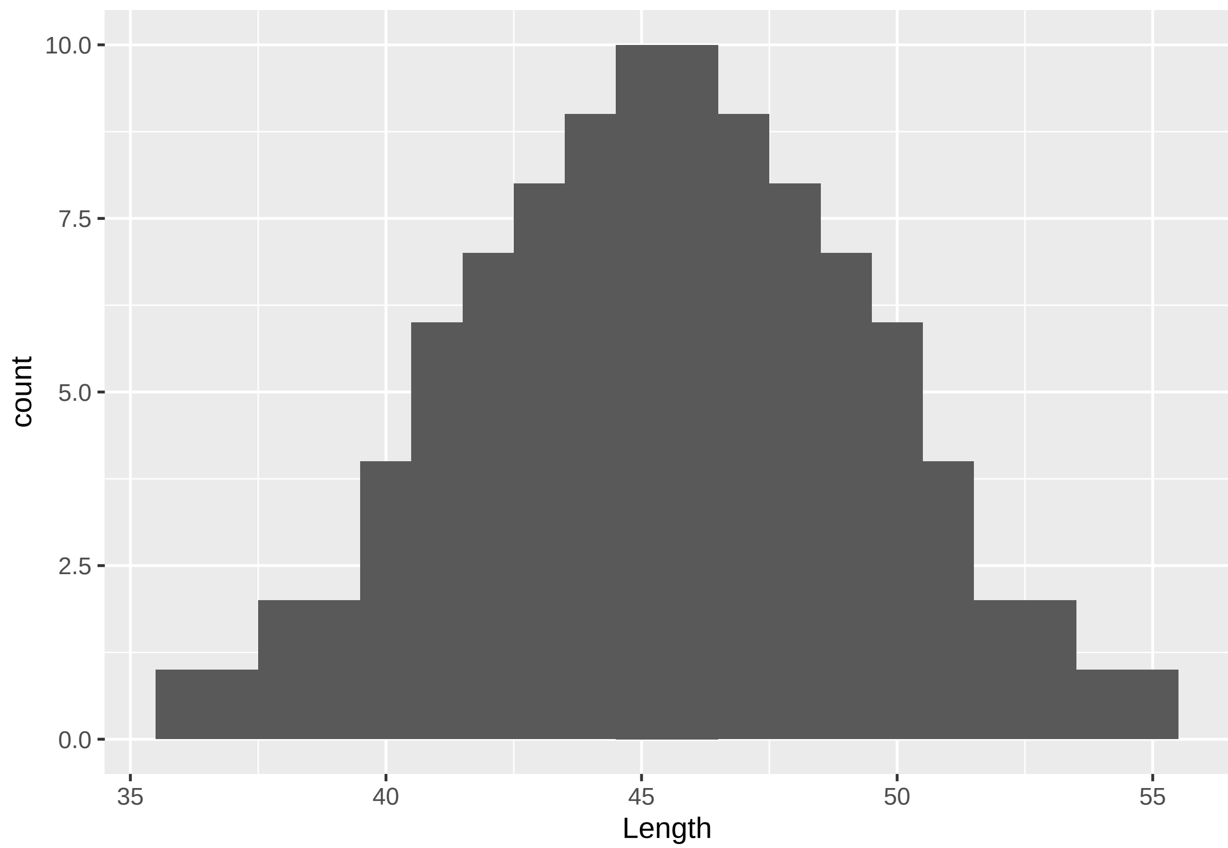
Plot the empirical CDF:

```r
ggplot(wings.df, aes(x = Length)) + stat_ecdf() + ylab("Empirical CDF")
```



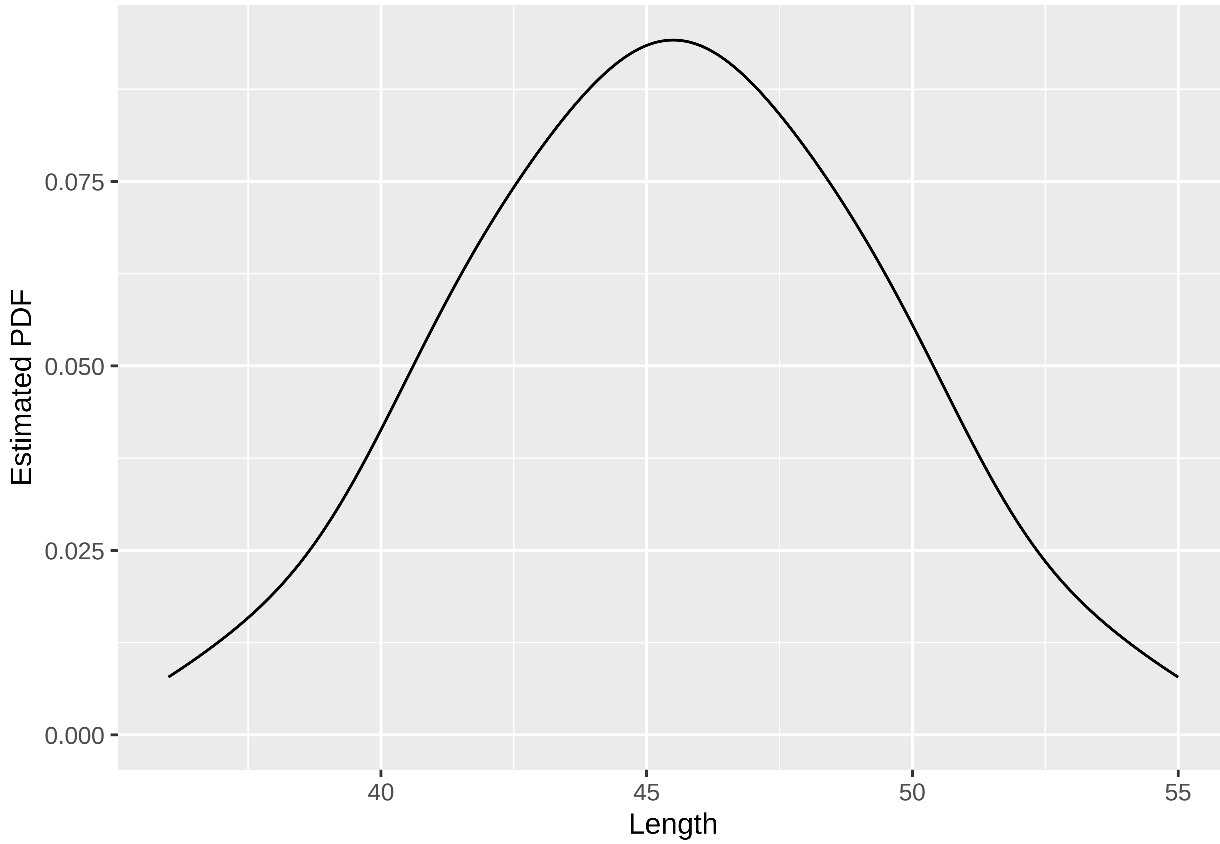
Draw a histogram:

```r
ggplot(wings.df,aes(x = Length)) + geom_histogram(breaks = seq(35.5, 55.5, 1))
```



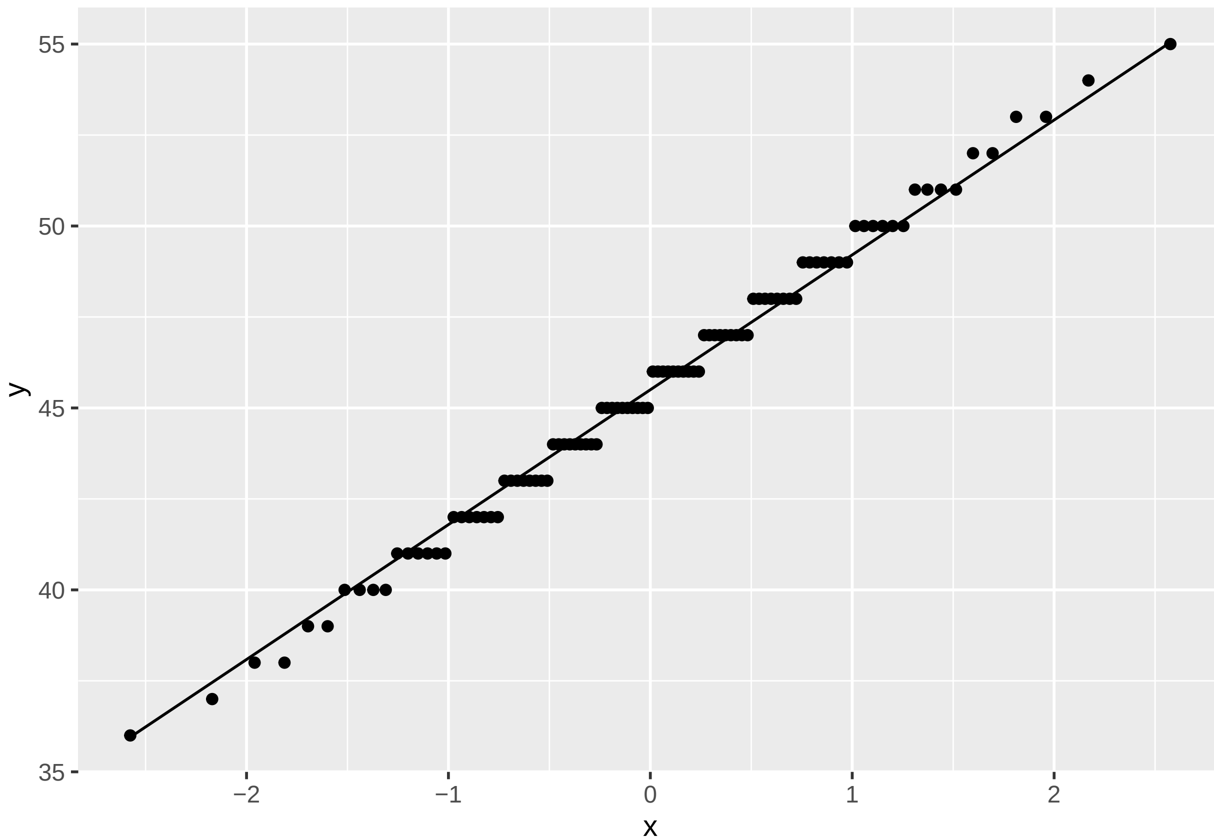
Draw a density estimate:

```
ggplot(wings.df, aes(x=Length)) + geom_density() + ylab("Estimated PDF")
```



Draw a normal quantile-quantile plot:

```
ggplot(wings.df, aes(sample = Length)) + stat_qq() + stat_qq_line()
```



Does the data look like it follows a normal distribution, apart from small discrepancies like rounding? Yes, the data does look like it follows a normal distribution. Even from the QQ line which is straight, and the fact that the points align perfectly well makes it look like its a normal distribution.

There's one thing about the data that might make one suspicious that the data is made up. The suspicious thing is: Perfection. In reality, the data is not as perfect as the data that is shown above. There will be noise, and outliers which in real practical data, which is absent in thic case.