MEDB 5501, Module03

2024-09-03

Topics to be covered

- What you will learn
  - The normal distribution
  - Normal probabilities and quantiles
  - Assessing normality
  - Using R to assess normality
  - Your homework

The bell shaped curve

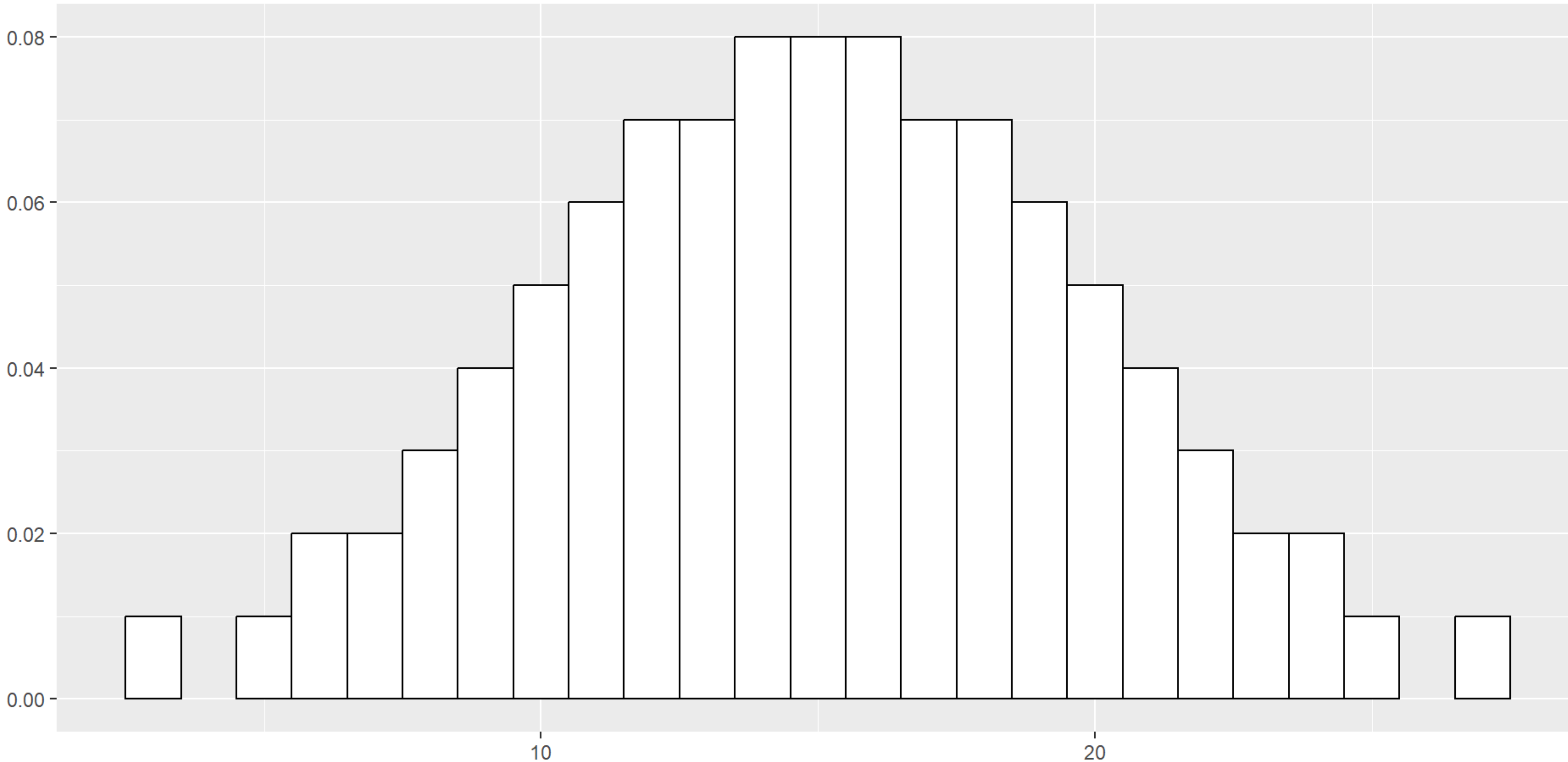
- Does your variation follow a bell shaped curve?
- Synonyms: normality, normal distribution
  - Values in the middle are most common
  - Frequencies taper off away from the center
  - Symmetry on either side

Speaker notes

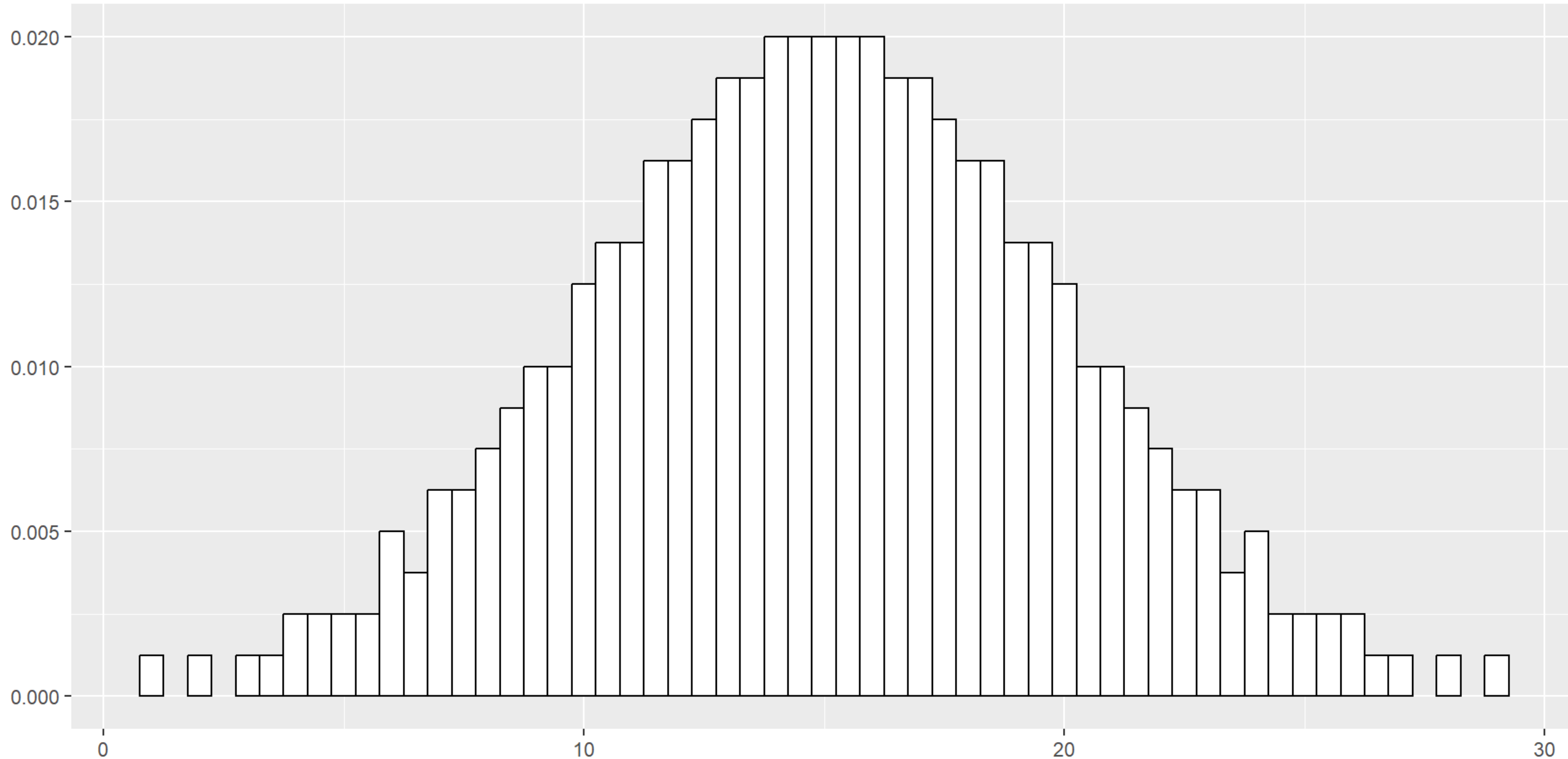
Speaker notes

Much variation in the real world follows a bell shaped curve, alternately called a normal distribution. You can assess whether you have a bell shaped curve using a histogram. Look for values in the middle being most common. The frequencies should taper off slowly as you moved away from the middle. The histogram should have symmetry. The left side of the histogram should be roughly equivalent to the right side of the histogram.

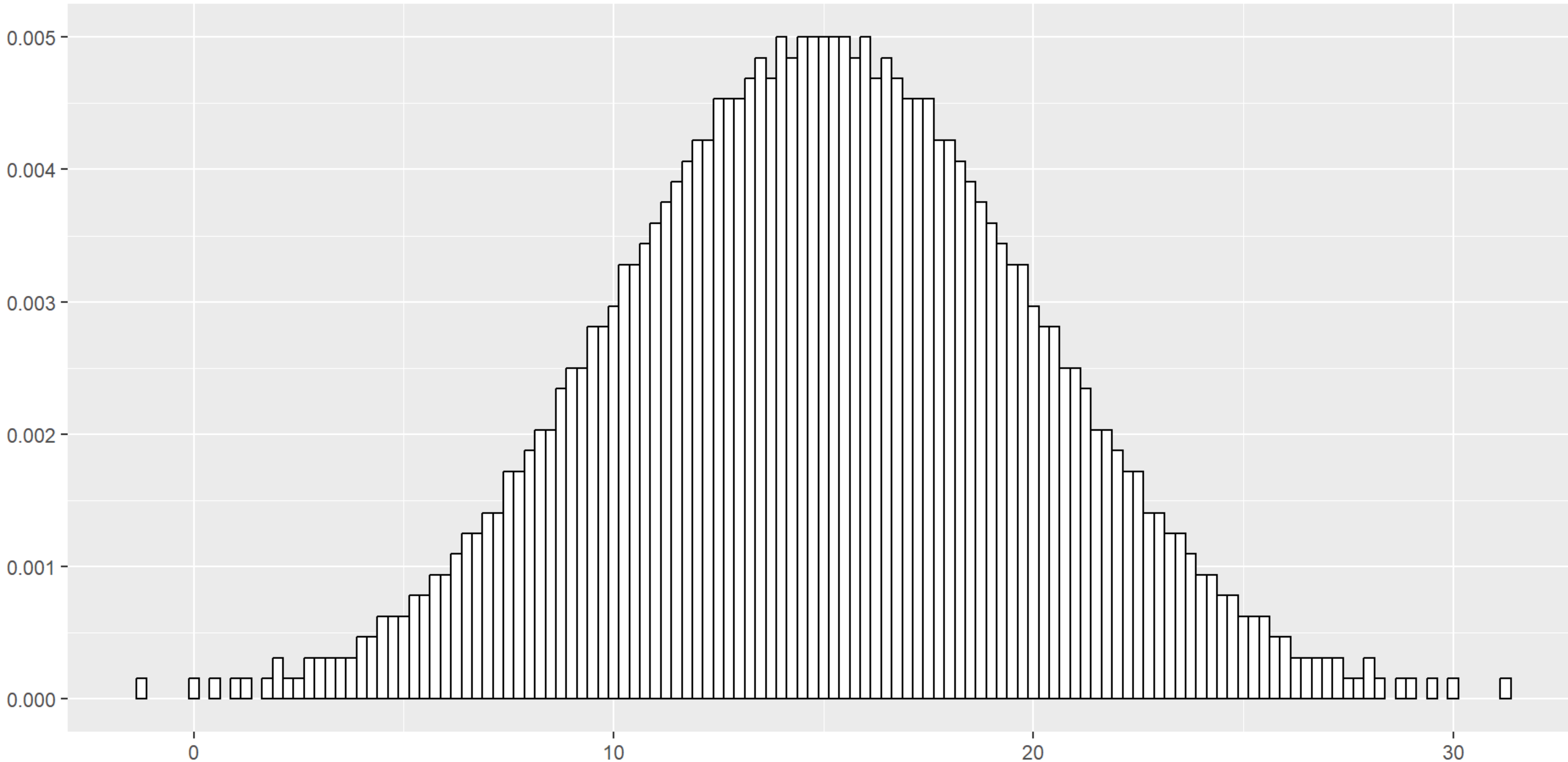
Bell-shaped curve is a limit



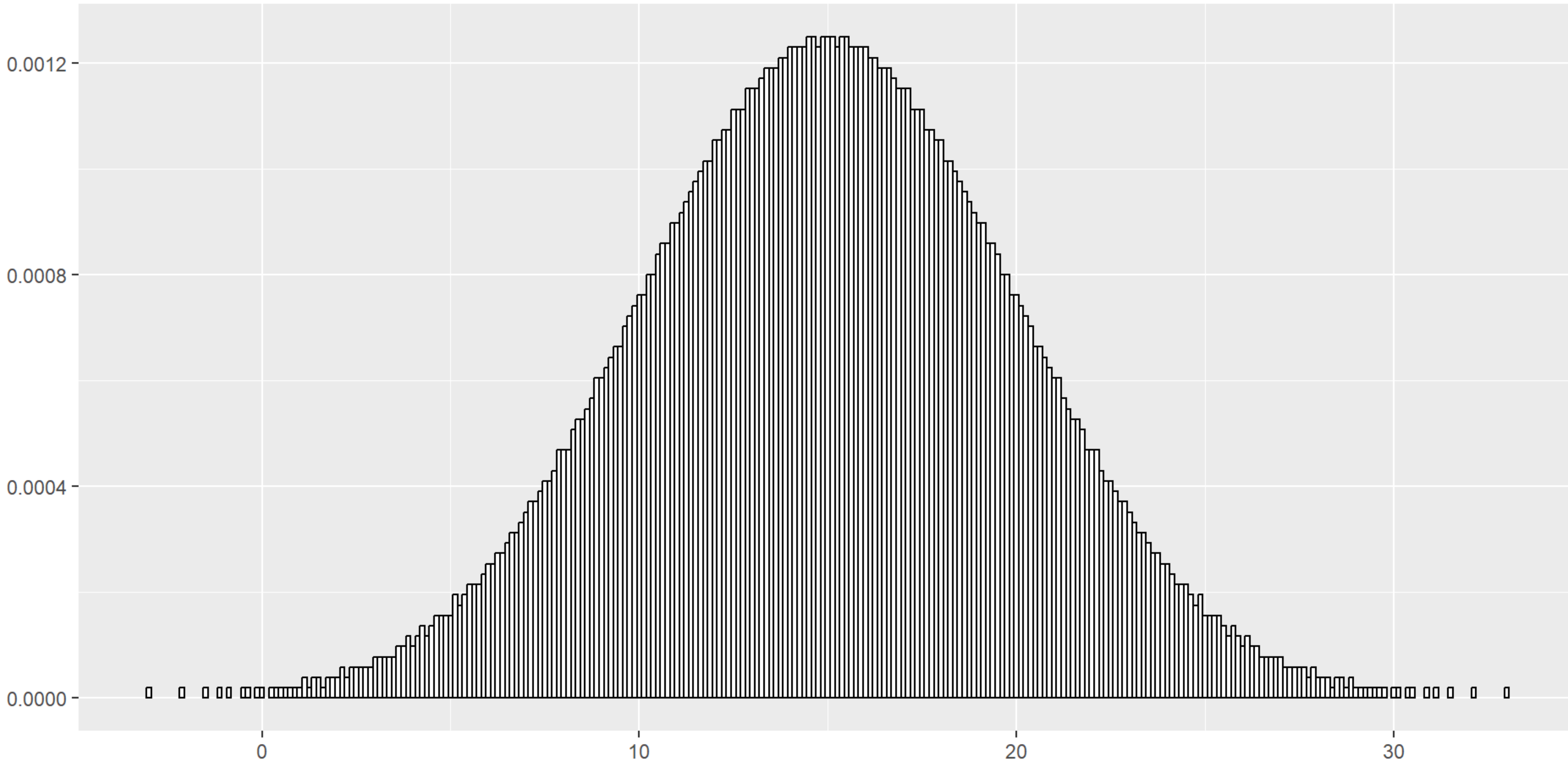
Bell-shaped curve is a limit



Bell-shaped curve is a limit



Bell-shaped curve is a limit



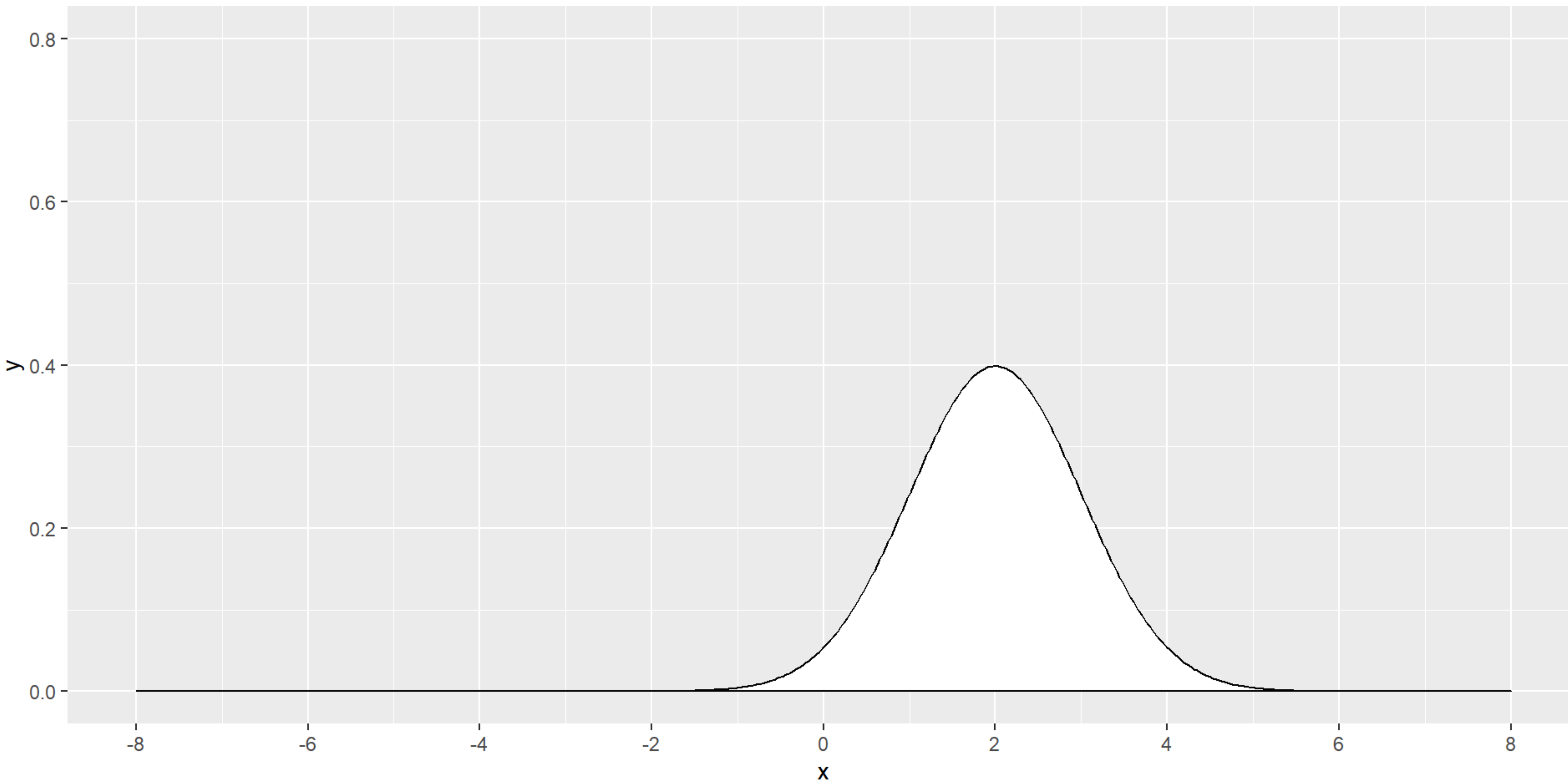
Overview

The normal distribution is an important component of most statistical analyses. The formula for the normal distribution is quite complex,

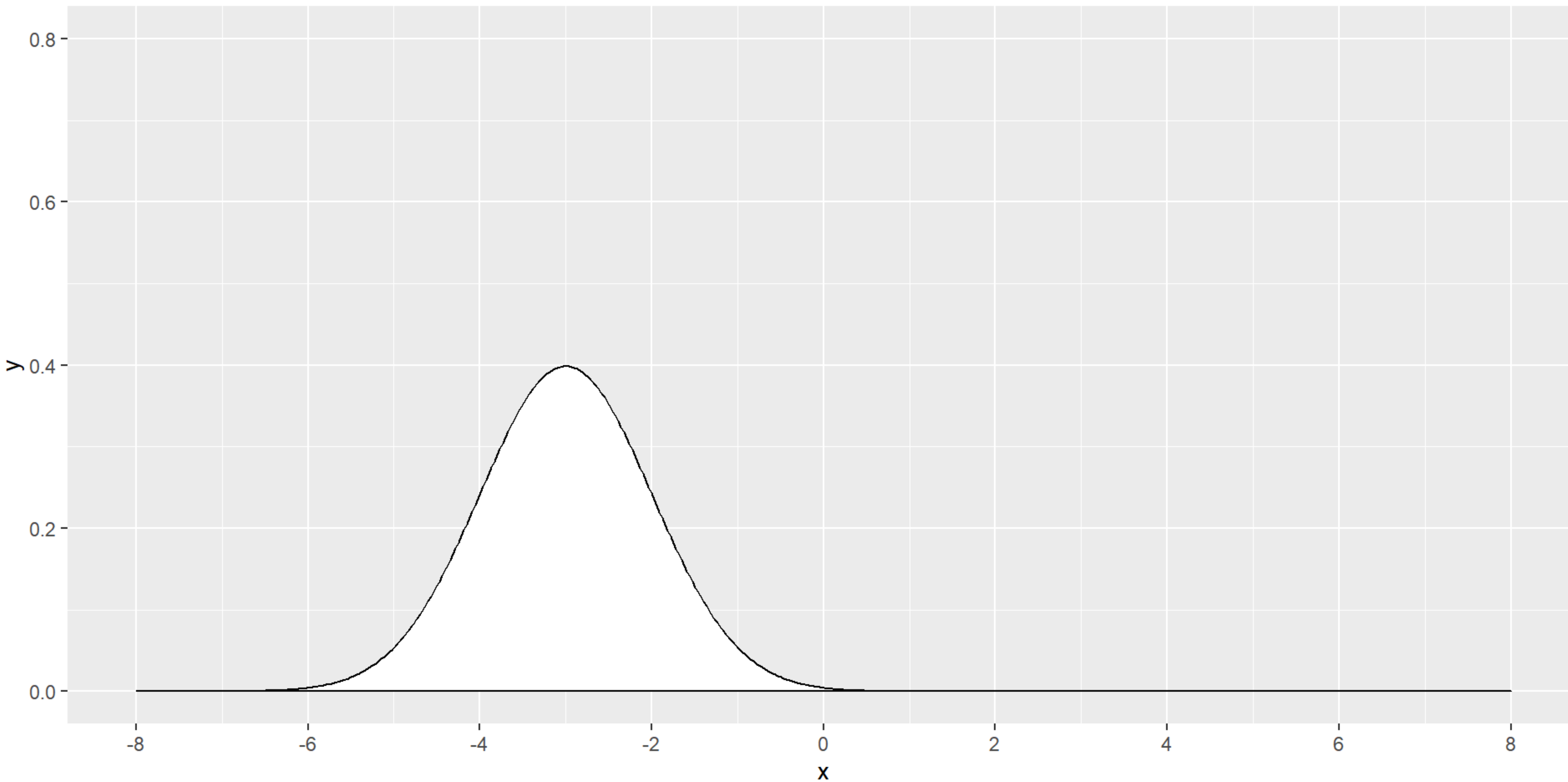
$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

but the shape is readily recognizable.

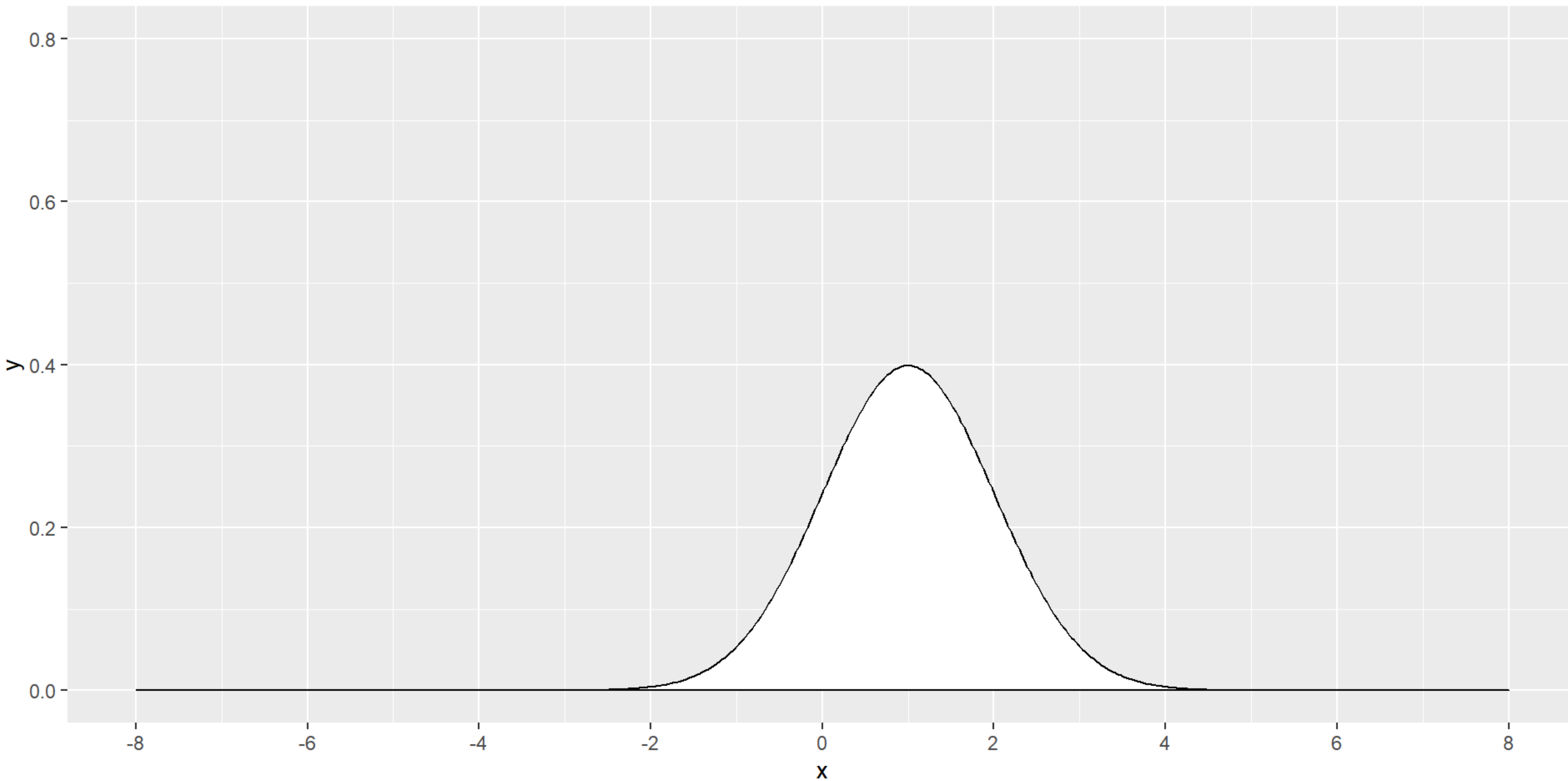
Normal distribution with mu=2, sigma=1



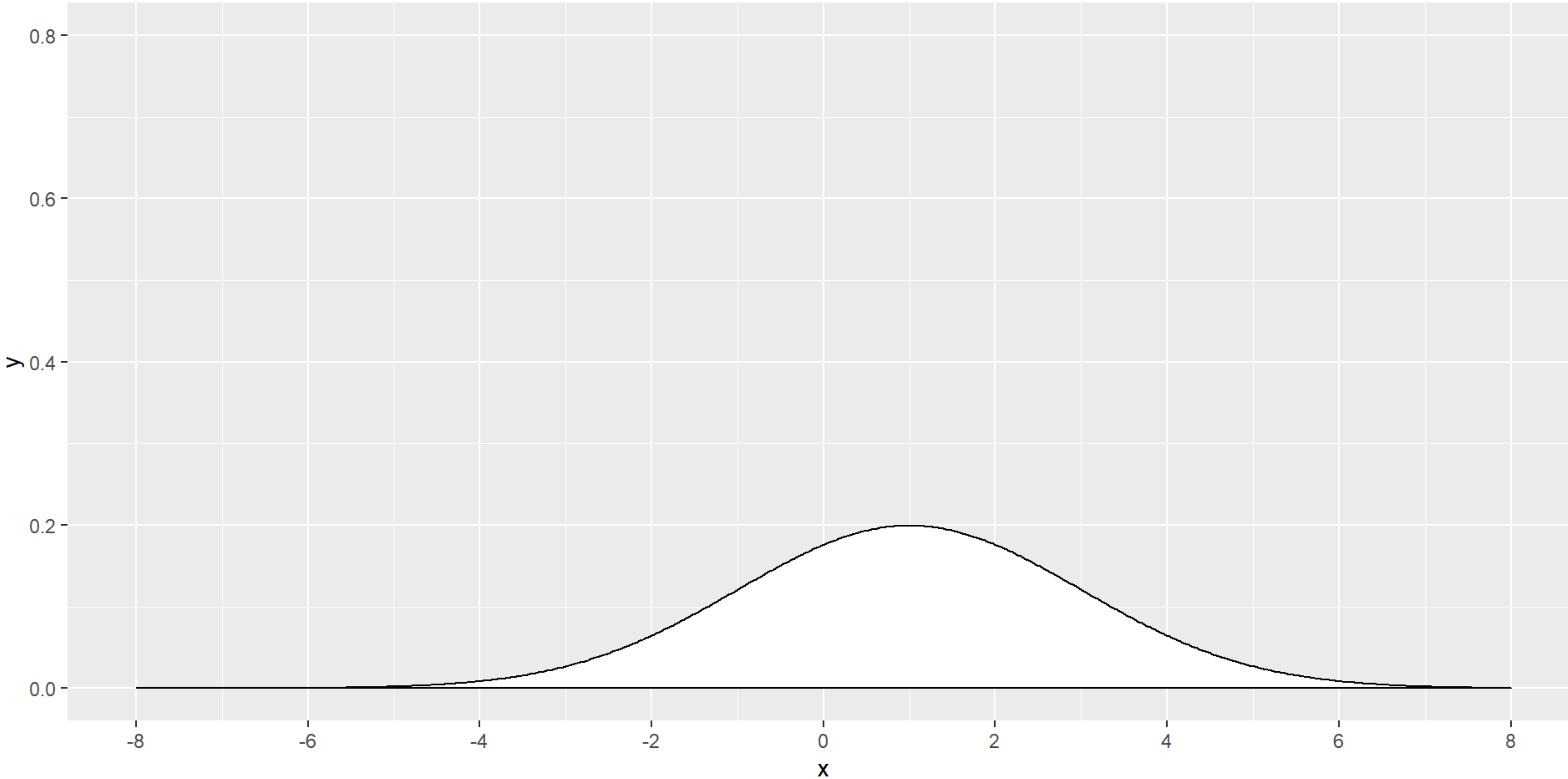
Normal distribution with mu=-3, sigma=1



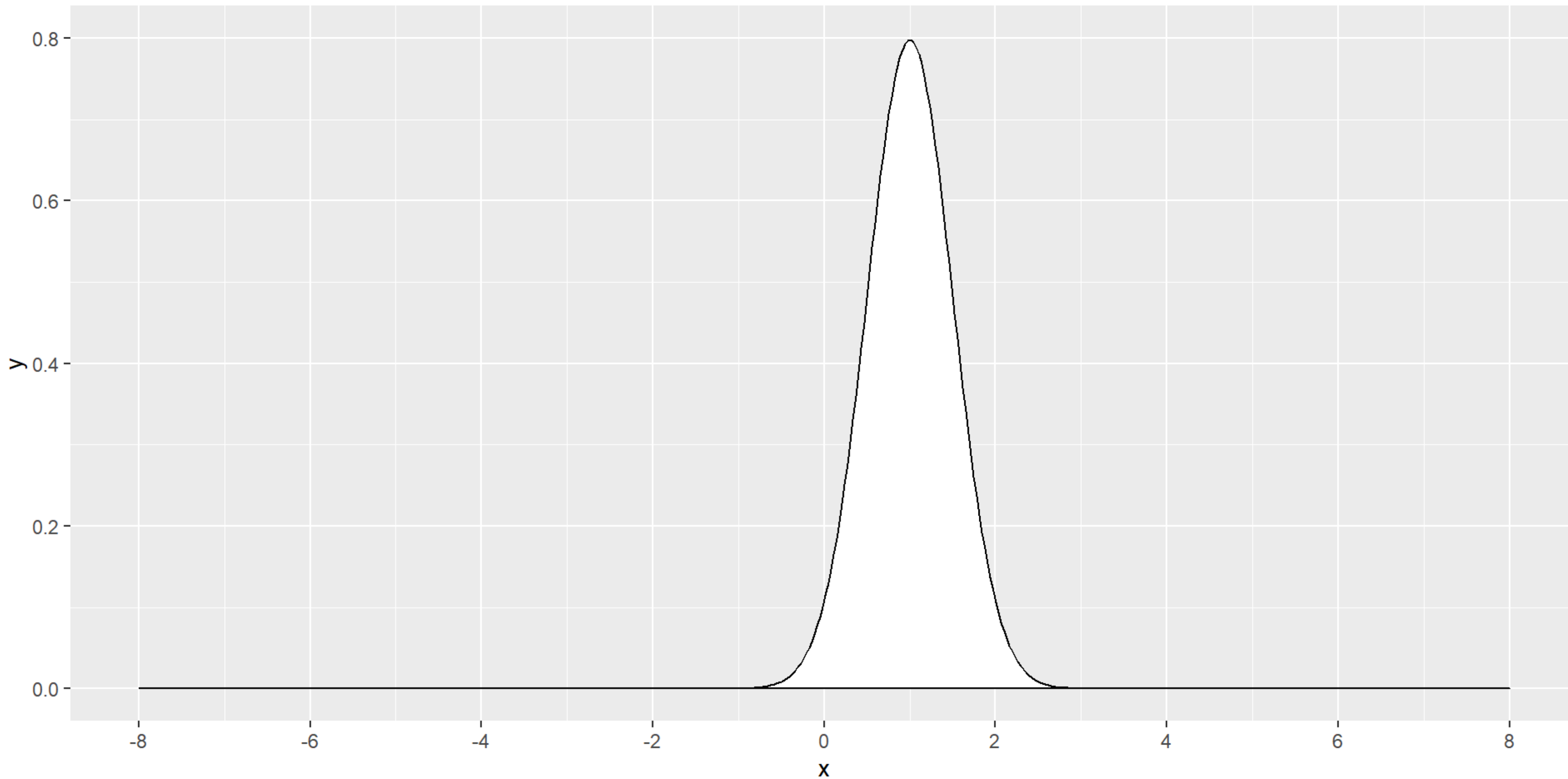
Normal distribution with mu=1, sigma=1



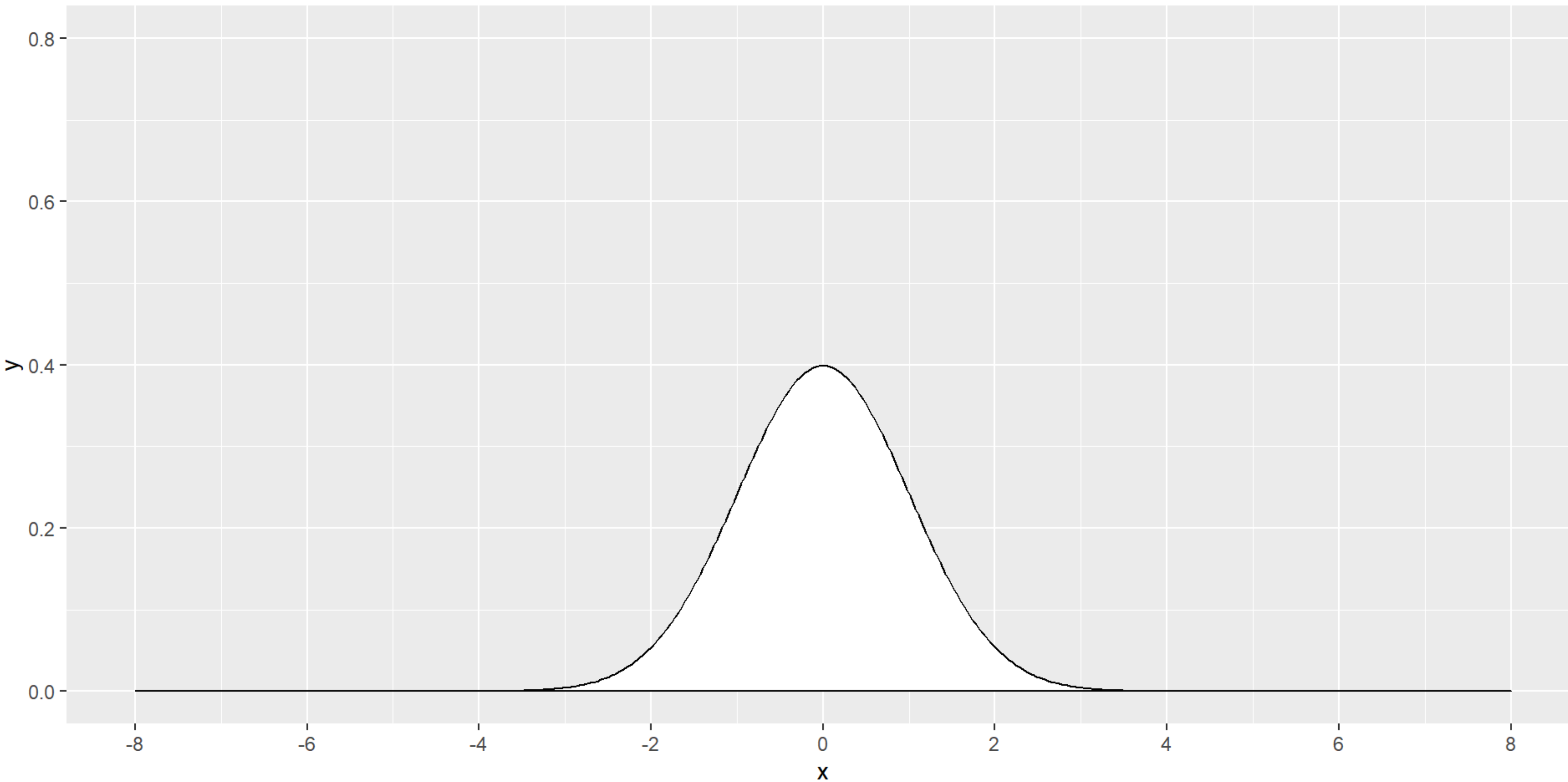
Normal distribution with mu=1, sigma=2



Normal distribution with mu=1, sigma=0.5



Standard normal distribution (mu=0, sigma=1)



Why concern yourself with the bell shaped curve?

- You can characterize individual observations
- You can characterize summary measures

Break #1

- What you have learned
  - The normal distribution
- What's coming next
  - Normal probabilities and quantiles

simon-5501-03-normal-calculations.qmd, 1

```
---
title: "Normal probabilities and quantiles"
format:
  revealjs:
    slide-number: true
    embed-resources: true
editor: source
---

This program displays the standard normal curve along with various
probabilities and quantiles. It was written by Steve Simon on 2024-09-01 and
is placed in the public domain.
```

Speaker notes

Here is a program template that illustrates how to calculate probabilities and quantiles for the standard normal distribution

The first few lines are the documentation header.

simon-5501-03-normal-calculations.qmd, 2

````
## Load the tidyverse library

```{r setup}
#| message: false
#| warning: false
library(tidyverse)
```
````

Speaker notes

You only need the tidyverse libraries for this program.

simon-5501-03-normal-calculations.qmd, 3

````
## Using R to draw the standard normal curve

use seq to calculate 100 evenly spaced values between -4 and +4 and dnorm to
compute the bell curve at each point. Use geom_polygon to paint the area
surrounded by the bell curve.

```{r standard-normal}
x <- seq(-4, 4, length=100)
y <- dnorm(x)
data.frame(x, y) |>
  ggplot(aes(x, y)) +
    geom_polygon(fill="white", color="black") -> normal_curve
normal_curve
```
````

Speaker notes

This code draws the bell shaped curve for a standard normal distribution. The graph is saved as normal_curve to allow for future modifications.

simon-5501-03-normal-calculations.qmd, 4

```
## P[Z < 1.5]

Use geom_vline and geom_label to draw a vertical reference line and add text
to the normal curve. The pnorm function computes the standard normal
probability.

```{r prob-1}
a <- 1.5
normal_curve +
  geom_vline(xintercept=a) +
  geom_label(x=a, y=0.4, label=a) +
  geom_label(x=a-0.5, y=0, label="Area = ?")
pnorm(1.5)
```
```

Speaker notes

This code computes the probability that a standard normal variable is less than a particular number and displays the area associated with this probability.

simon-5501-03-normal-calculations.qmd, 5

```
## P[Z < -0.5]

```{r prob-2}
a <- -0.5
normal_curve +
  geom_vline(xintercept=a) +
  geom_label(x=a, y=0.4, label=a) +
  geom_label(x=a-0.5, y=0, label="Area = ?")
pnorm(-0.5)
```
```

Speaker notes

Here’s a similar calculation.

simon-5501-03-normal-calculations.qmd, 6

````
## P[Z > 1]

When you are calculating the probability on the right (probability greater
than some number), use 1-pnorm.

```{r prob-3}
a <- 1
normal_curve +
  geom_vline(xintercept=a) +
  geom_label(x=a, y=0.4, label=a) +
  geom_label(x=a+0.5, y=0, label="Area = ?")
1- pnorm(1)
```
````

Speaker notes

For greater than probabilities (probabilities corresponding to area to the right), subtract the pnorm result from 1.

simon-5501-03-normal-calculations.qmd, 7

````
## P[Z > -2]

```{r prob-4}
a <- -2
normal_curve +
  geom_vline(xintercept=a) +
  geom_label(x=a, y=0.4, label=a) +
  geom_label(x=a+0.5, y=0, label="Area = ?")
1- pnorm(-2)
```
````

Speaker notes

Here’s another greater than probability calculation.

simon-5501-03-normal-calculations.qmd, 8

```
## P[-2.5 < Z < 2.5]

When you are calculating the probability between two values, compute pnorm of
the larger value minus pnorm of the smaller value.

```{r prob-5a}
a <- 2.5
normal_curve +
  geom_vline(xintercept=-a) +
  geom_vline(xintercept= a) +
  geom_label(x=-a, y=0.4, label=-a) +
  geom_label(x= a, y=0.4, label= a) +
  geom_label(x=0, y=0, label="Area = ?")
pnorm(2.5)- pnorm(-2.5)
```
```

Speaker notes

For probabilities between two values, calculate the difference.

simon-5501-03-normal-calculations.qmd, 9

````
## P[-0.5 < Z < 0.5]

```{r prob-6}
a <- 0.5
normal_curve +
  geom_vline(xintercept=-a) +
  geom_vline(xintercept= a) +
  geom_label(x=-a, y=0.4, label=-a) +
  geom_label(x= a, y=0.4, label= a) +
  geom_label(x=0, y=0, label="Area = ?")
pnorm(0.5) - pnorm(-0.5)
```
````

Speaker notes

Here’s a similar probability calculation.

simon-5501-03-normal-calculations.qmd, 10

```
## 25th percentile of a standard normal

Use qnorm to calculate quantiles of the standard normal distribution.

```{r quantile-1}
p <- 0.25
a <- qnorm(p)
normal_curve +
  geom_vline(xintercept=a) +
  geom_label(x=a, y=0.4, label="Quantile = ?") +
  geom_label(x=a-0.5, y=0, label=p)
qnorm(0.25)
```
```

Speaker notes

The normal quantile is the value associated with a specified probability. Use the qnorm function for quantile calculations.

simon-5501-03-normal-calculations.qmd, 11

````
## 90th percentile of a standard normal

```{r quantile-2}
p <- 0.9
a <- qnorm(p)
normal_curve +
  geom_vline(xintercept=a) +
  geom_label(x=a, y=0.4, label="Quantile = ?") +
  geom_label(x=a-0.5, y=0, label=p)
qnorm(0.9)
```
````

Speaker notes

Here is another quantile calculation.

Break #2

- What you have learned
  - Normal probabilities and quantiles
- What's coming next
  - Assessing normality

Assessing normality

- No variable follows a perfect normal distribution
  - But many are close
- How to assess (approximate) normality
  - Histogram
  - Boxplot
  - Normal probability plot
- Avoid formal tests of normality

Speaker notes

While no variables are going to fit the bell shaped curve exactly, many variables come close. Your subjective interpretation of graphs is the best way to assess normality, recognizing that you would be satisfied with a reasonable approximation to normality. Three graphs used commonly for this are the histogram, the boxplot and the normal probability plot.

Histogram

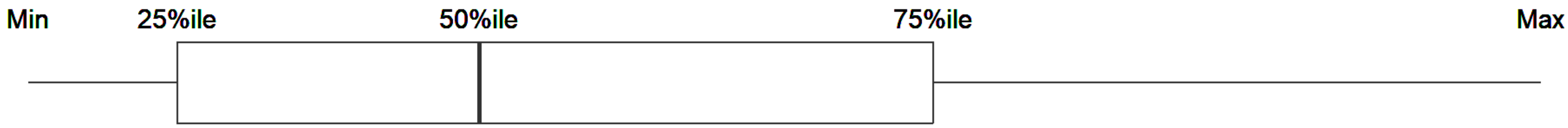
- Peak in the middle
- Roughly symmetric
- Falls off exponentially
- Warning!! Bar width can influence your interpretation
  - Try two or more bar widths

Speaker notes

Look for a peak in the middle of the histogram. If you see two separate peaks, the data is not normally distributed. The histogram should be roughly symmetric, but don’t expect perfect symmetry. The bars should fall off more or less exponentially from either side of the peak.

Be careful, because the number of bars that you draw can influence your interpretation. Something that looks normal with a small number of wide bars might look not so normal with a large number of narrow bars.

Sample Boxplot



Speaker notes

What to look for in the boxplot

- Median halfway between 25th and 75th percentile.
- Whiskers are same size
- Whiskers not too short, not too long

Constructing a normal probability plot, 1

- Calculate rank
- Divide by (n+1)
- Compute corresponding normal percentiles
- Compare these on a graph to the original data
- Roughly straight line implies normality

Constructing a normal probability plot, 2

Use the rank function to assign 1 to the smallest value, 2 to the next smallest value, etc. up to n for the largest value.

```
   x r
1  7 4
2  3 2
3 23 9
4  2 1
5  5 3
6 13 6
7 11 5
8 17 7
9 19 8
```

Constructing a normal probability plot, 3

Divide the rank by (n+1) to get evenly spaced percentages.

```
   x r pctile
1  7 4    0.4
2  3 2    0.2
3 23 9    0.9
4  2 1    0.1
5  5 3    0.3
6 13 6    0.6
7 11 5    0.5
8 17 7    0.7
9 19 8    0.8
```

Constructing a normal probability plot, 4

Calculate percentiles from a normal distribution using the evenly spaced percentages.

```
   x r pctile          z
1  7 4    0.4 -0.2533471
2  3 2    0.2 -0.8416212
3 23 9    0.9  1.2815516
4  2 1    0.1 -1.2815516
5  5 3    0.3 -0.5244005
6 13 6    0.6  0.2533471
7 11 5    0.5  0.0000000
8 17 7    0.7  0.5244005
9 19 8    0.8  0.8416212
```

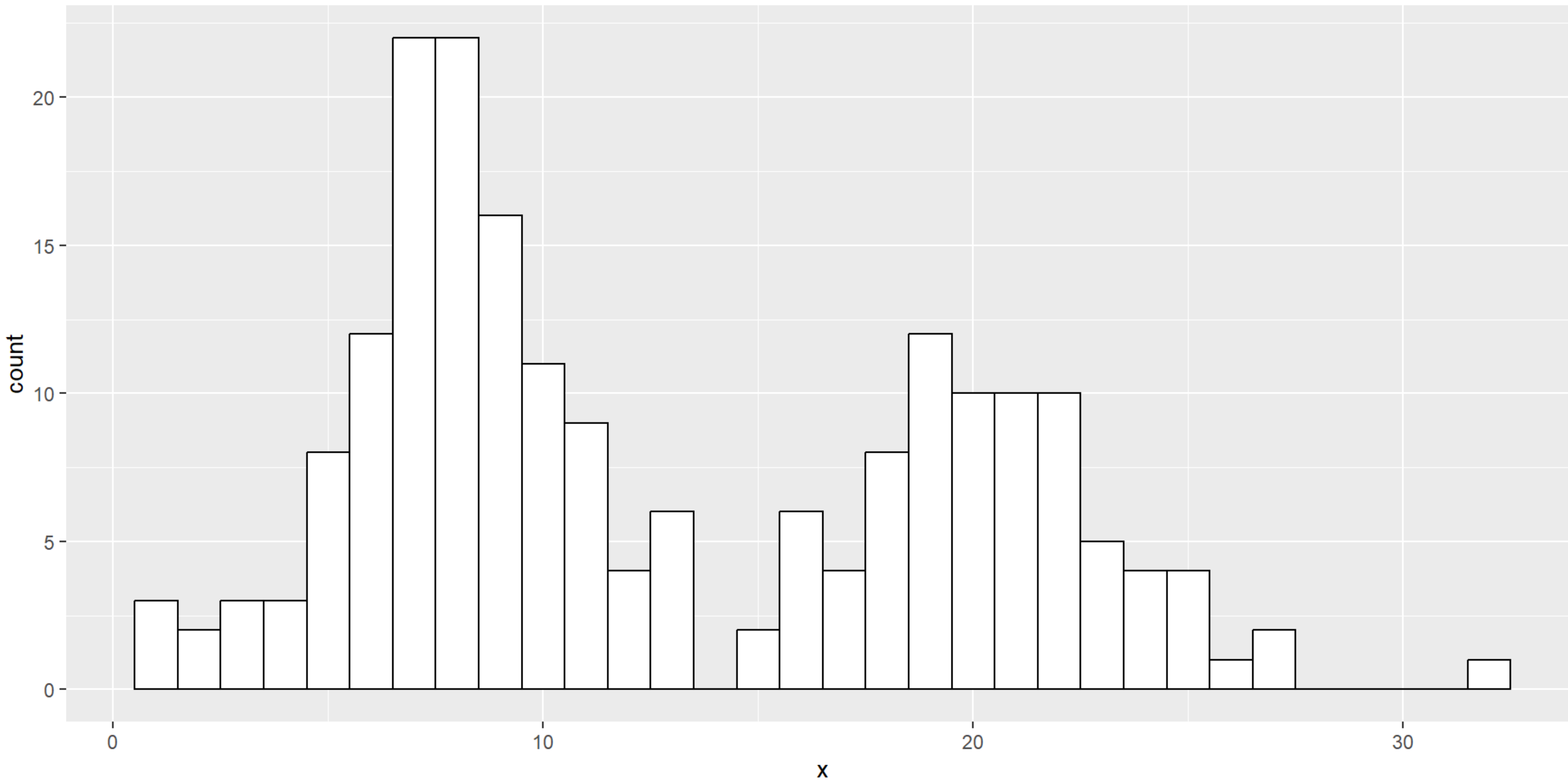
Constructing a normal probability plot, 5

Plot the percentiles from the normal distribution to the original data. A reasonably straight line is evidence of normality.

Constructing a normal probability plot, 6

The qqnorm function will do all these steps for you automatically.

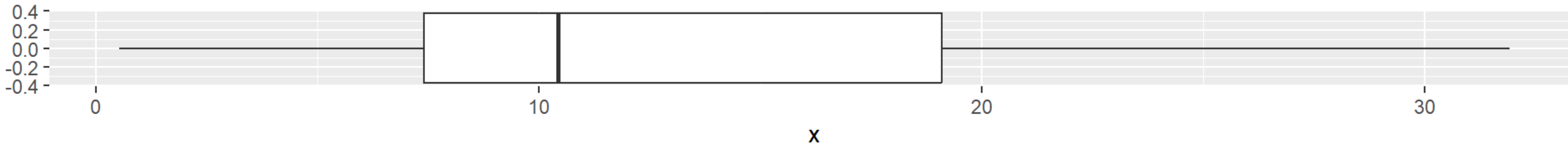
Bimodal data, histogram



Speaker notes

Here’s a histogram that shows a bimodal distribution. The frequencies are not highest in the center of the data. This is not a bell shaped curve.

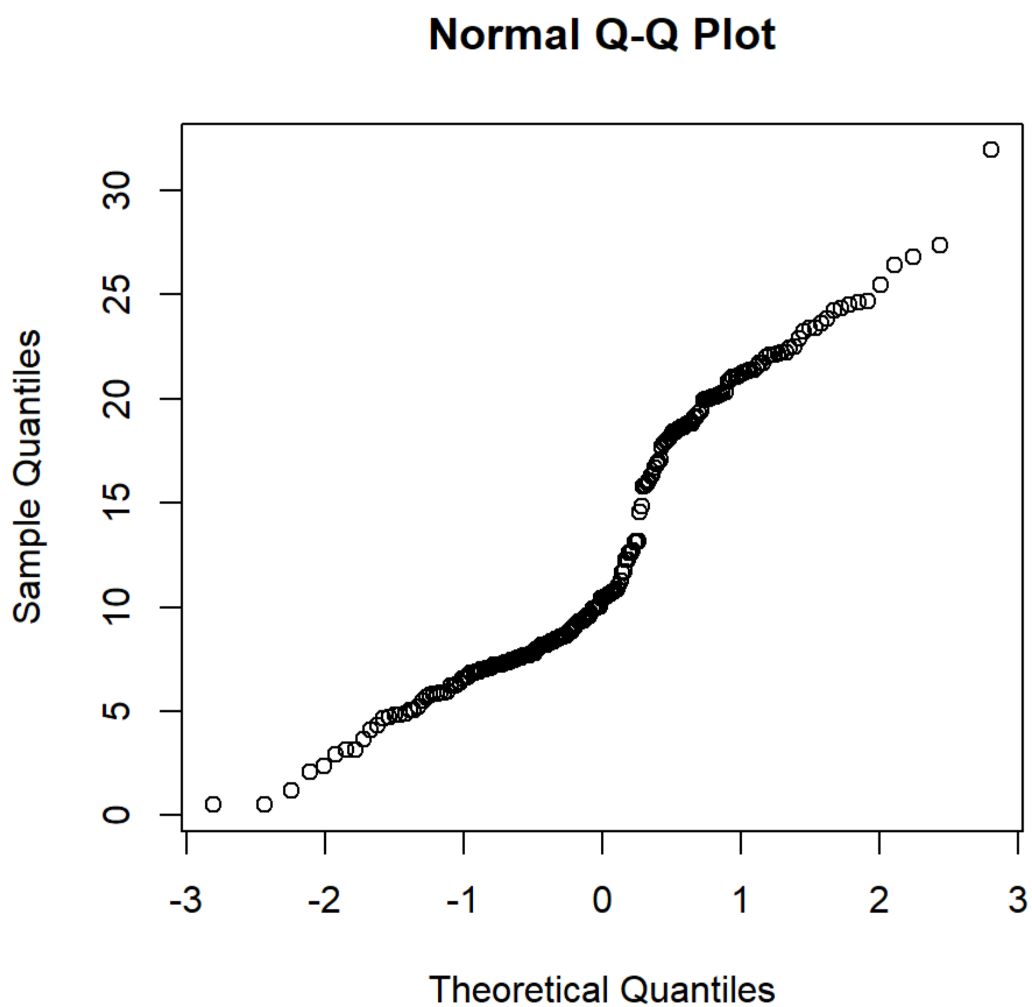
Bimodal data, boxplot



Speaker notes

The boxplot is not as useful as the histogram for detecting bimodal distributions.

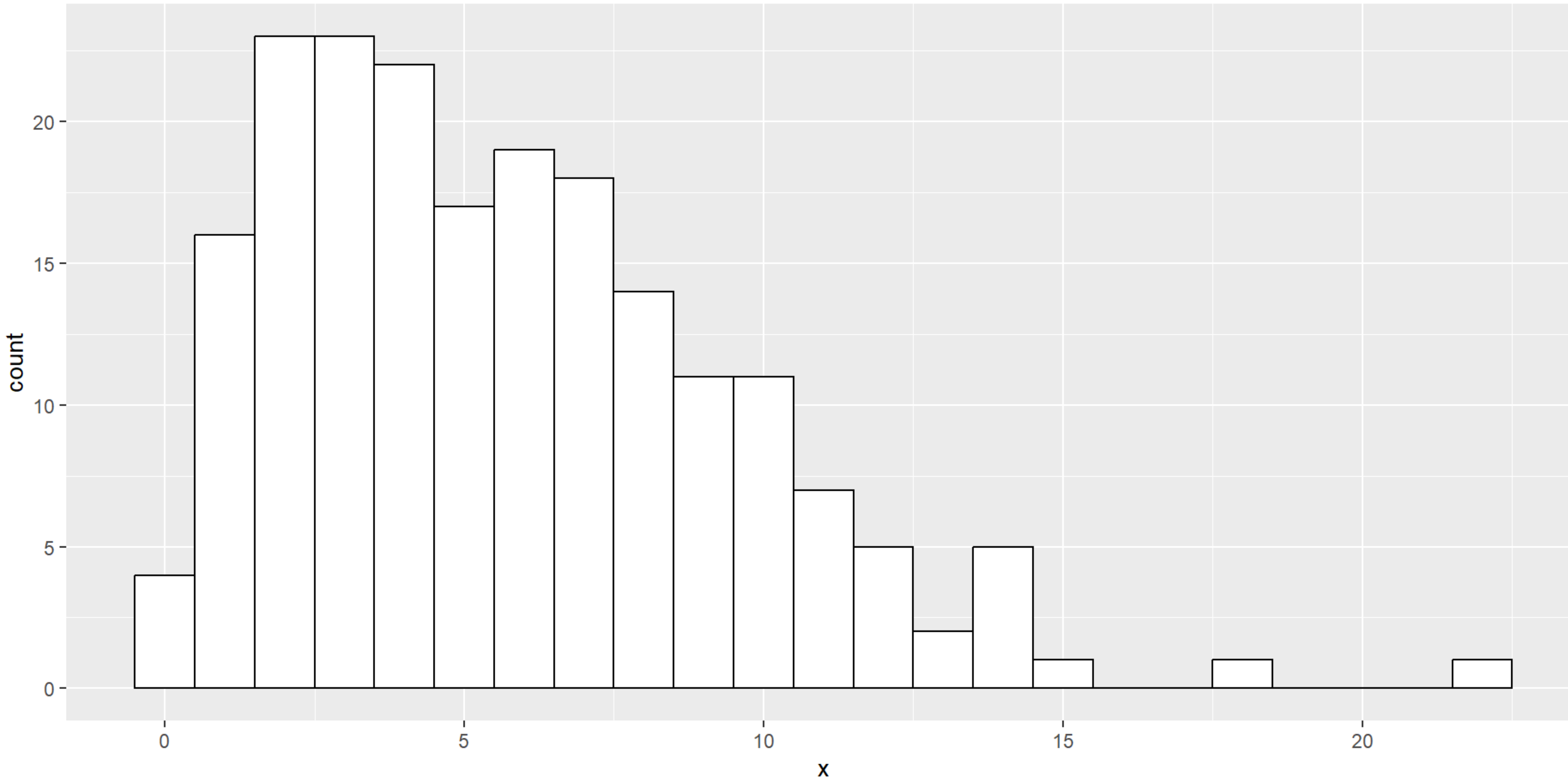
Bimodal data, qq plot



Speaker notes

On the qq plot, a bimodal pattern is often represented as two lines with a sharp jump between them.

Skewed data, histogram

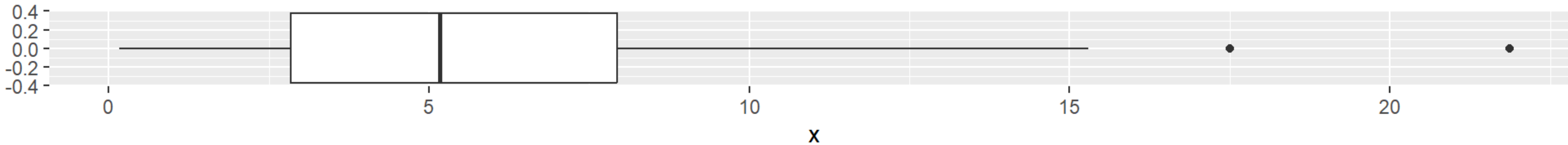


Speaker notes

Speaker notes

Here’s a histogram that shows a skewed or asymmetric distribution. This is not a bell shaped curve.

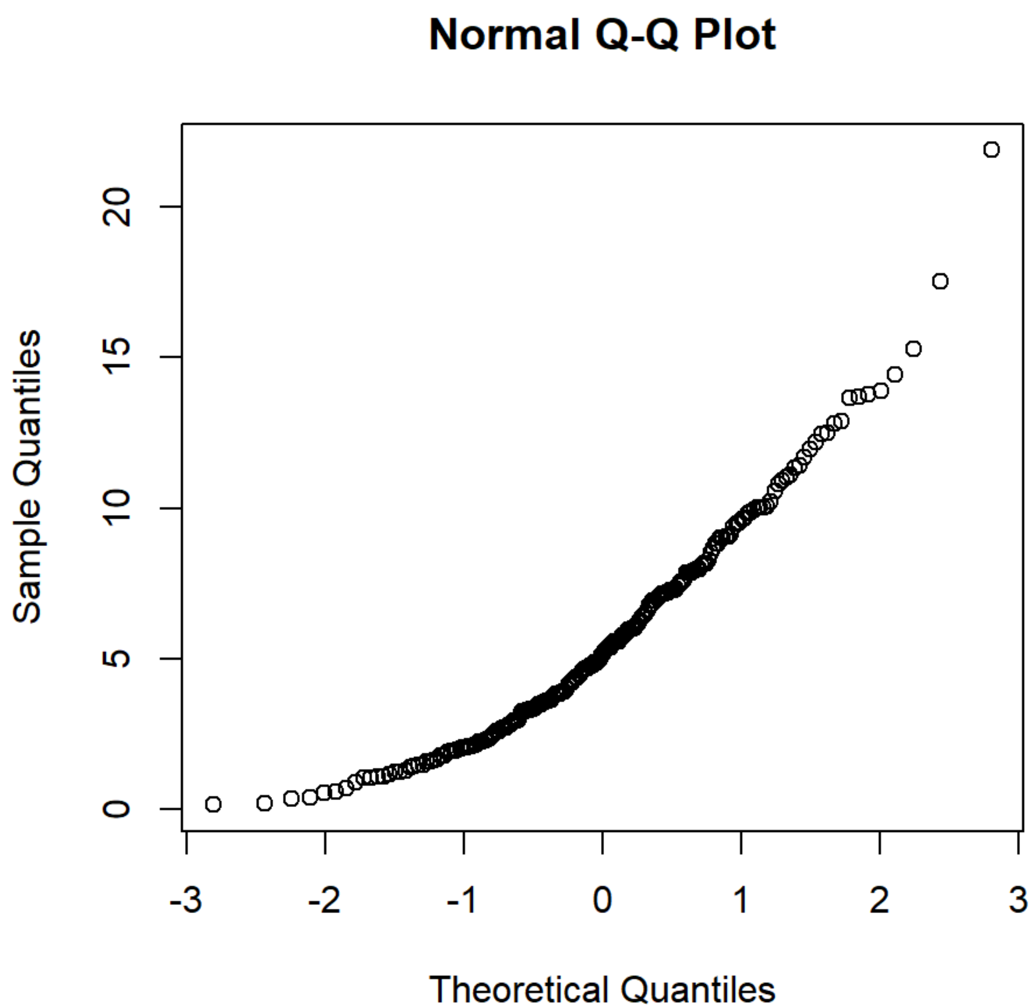
Skewed distribution, boxplot



Speaker notes

An asymmetry in the box and/or the whiskers is an indication of a skewed distribution.

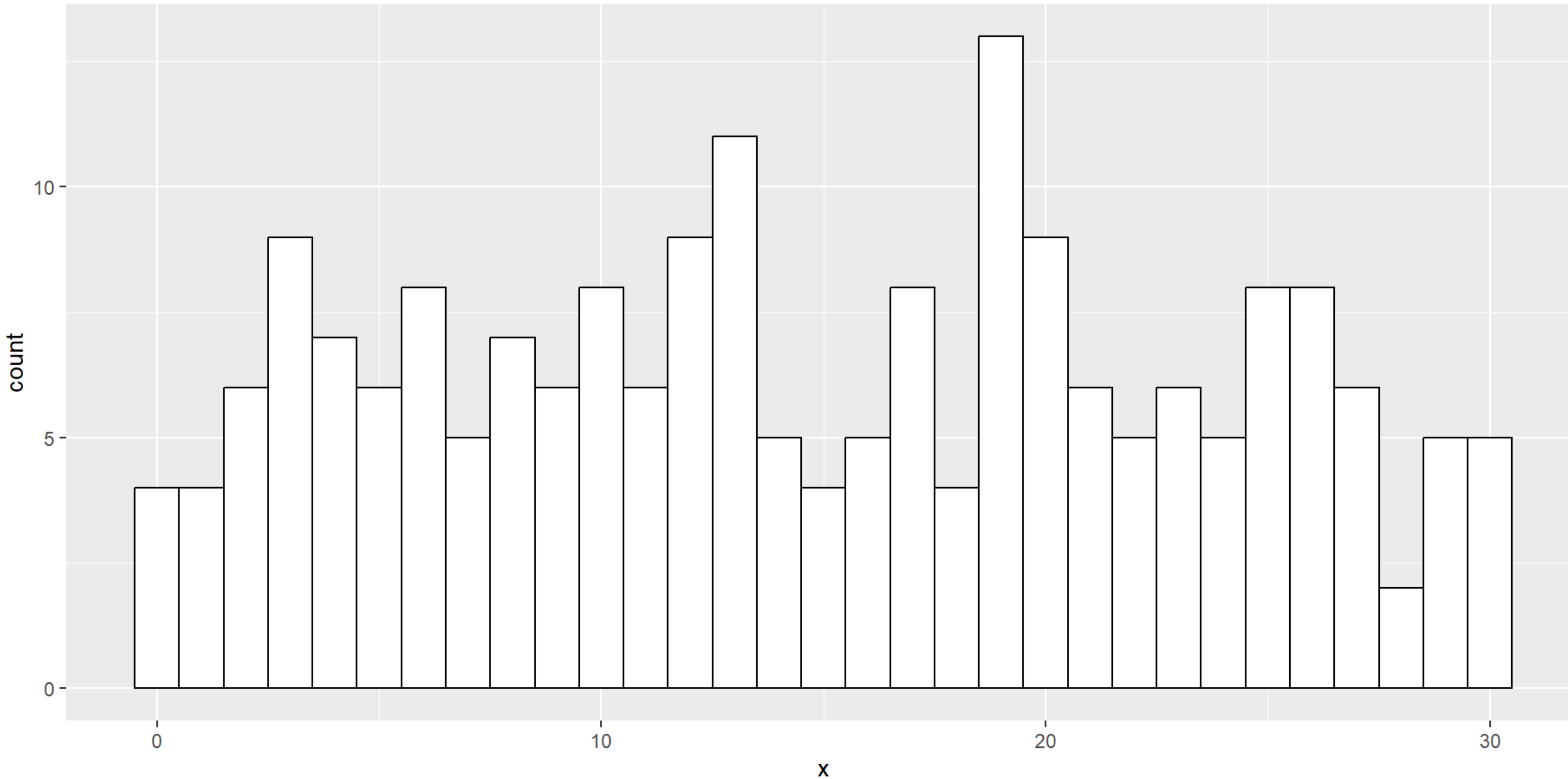
Skewed distribution, qq plot



Speaker notes

A curved pattern for the normal probability plot indicates skewness.

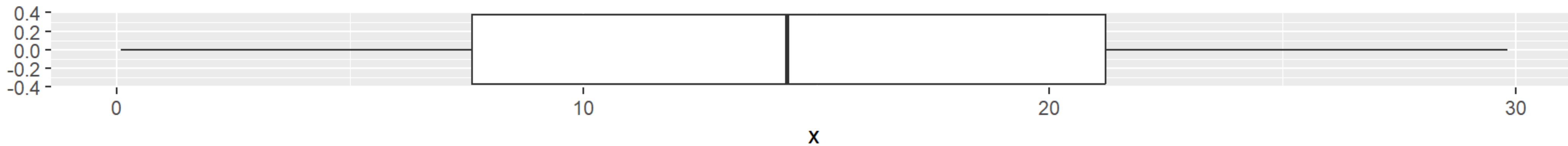
Light-tailed data, histogram



Speaker notes

Here’s a histogram that shows a symmetric distribution, but the frequencies do not taper off as you move away from the center. This is not a bell shaped curve.

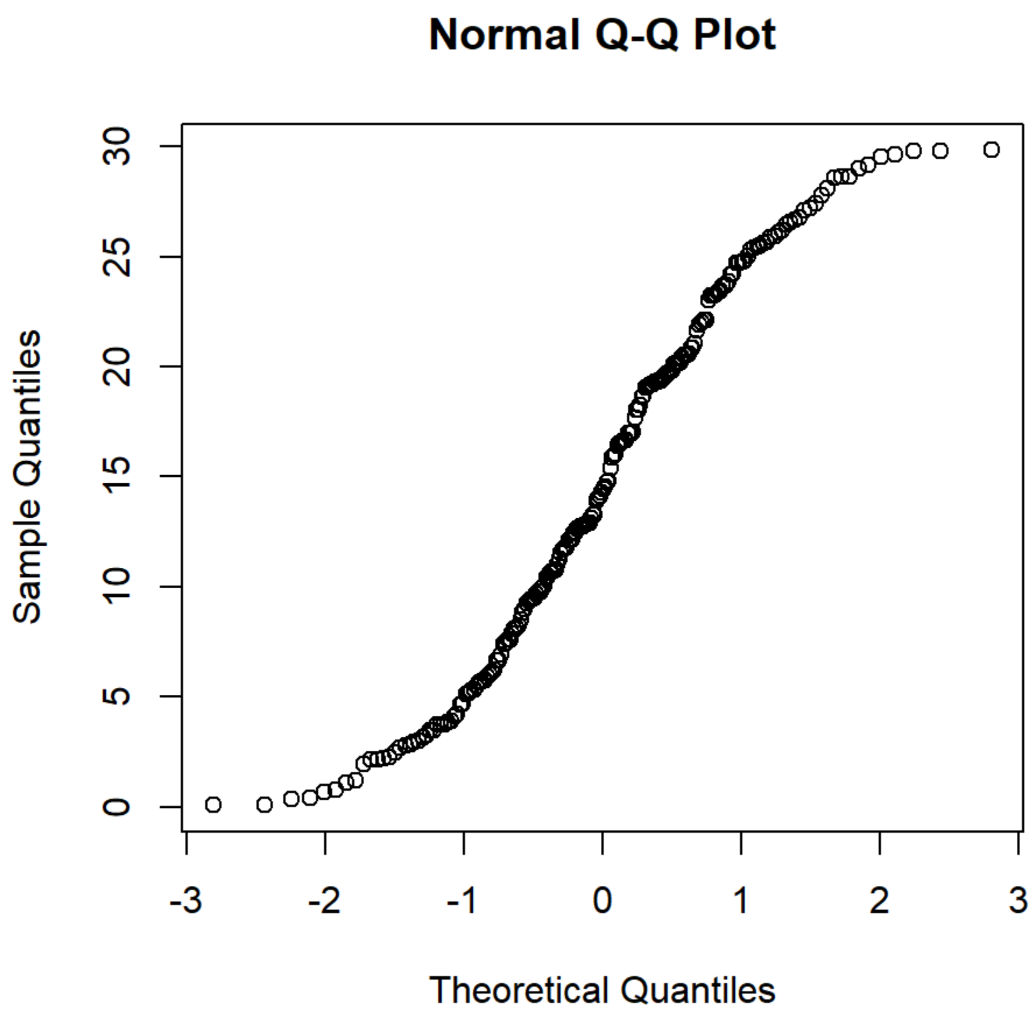
Light-tailed distribution, boxplot



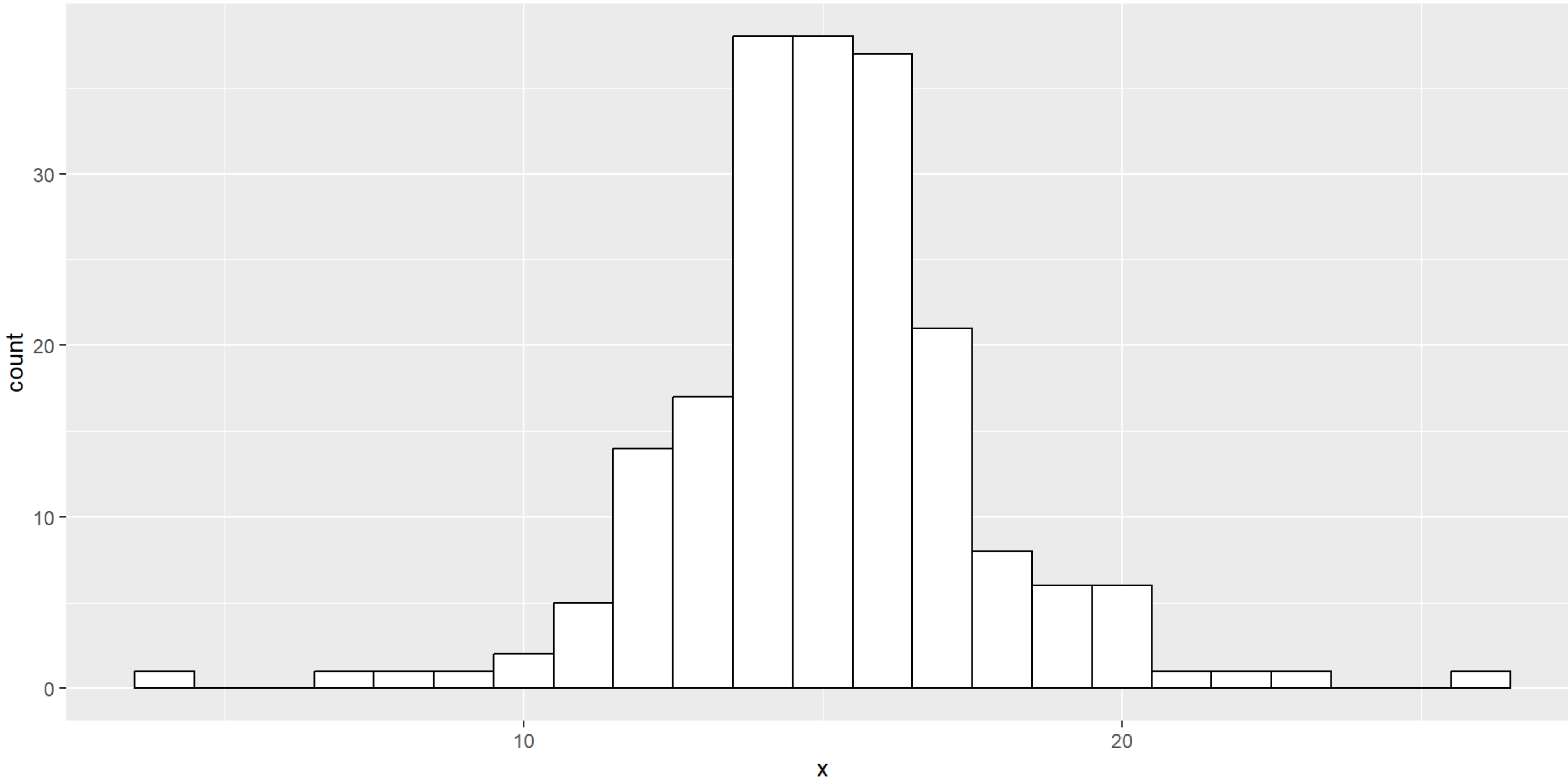
Speaker notes

A boxplot with very short whiskers is evidence of a light tailed distibution.

Light-tailed distribution, qq plot



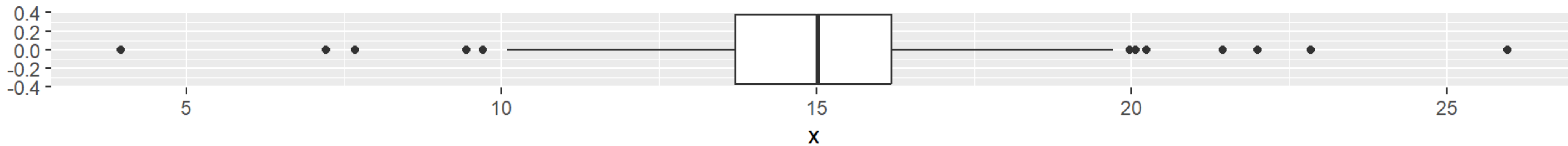
Heavy-tailed distribution, histogram



Speaker notes

Here’s a histogram that shows a symmetric distibution, but the frequencies taper off at first, but then flatten out. This is called a heavy tailed distribution and it tends to produce outliers, extreme values, on both sides. This is not a bell shaped curve.

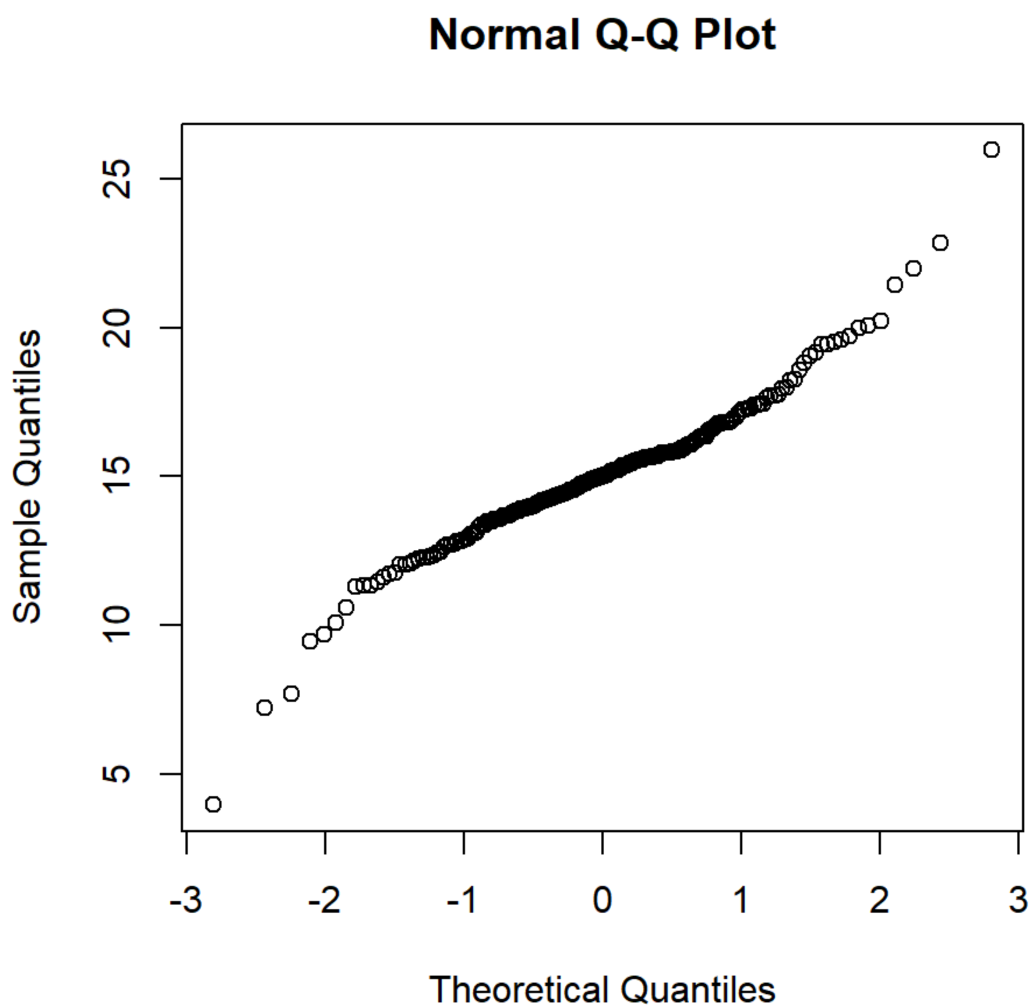
Heavy-tailed distribution, boxplot



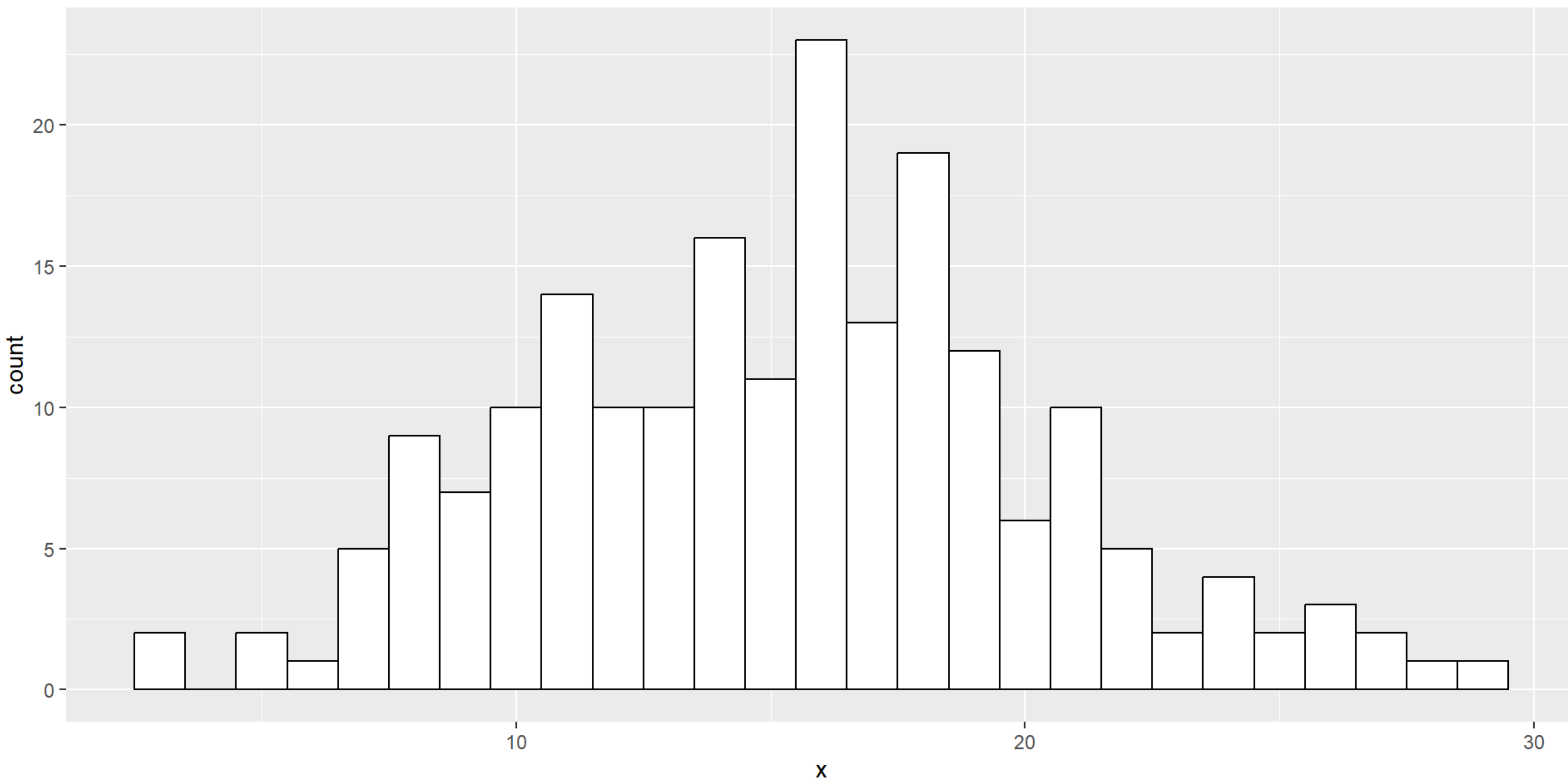
Speaker notes

The boxplot is not as useful as the histogram for detecting bimodal distributions.

Heavy tailed data, qq plot



A normal distribution, histogram



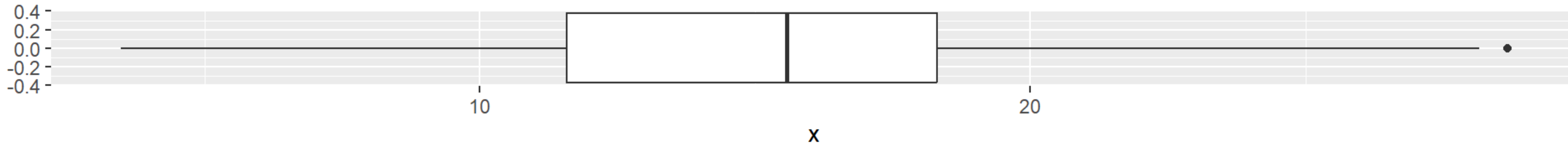
Speaker notes

Speaker notes

Here’s a histogram that shows a symmetric distribution, with the most frequent values in the center and frequencies that taper off on either side.

This is a bell shaped curve.

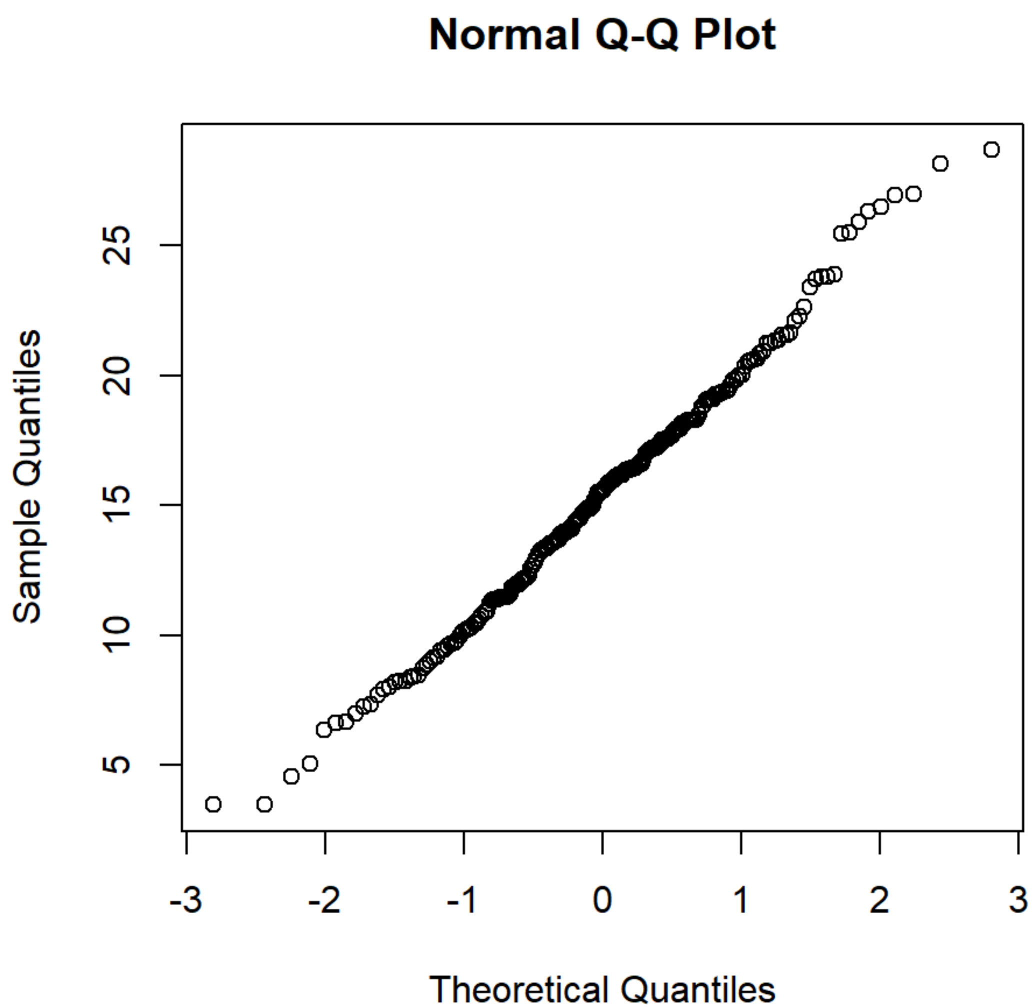
A normal distribution, boxplot



Speaker notes

The boxplot has a roughly symmetric box and roughly symmetric whiskers. The whiskers are a bit longer than the box itself, but not a lot longer.

A normal distribution, qq plot



Speaker notes

Speaker notes

A roughly straight line indicates a normal distribution.

Break #3

- What you have learned
  - Assessing normality
- What's coming next
  - Using R to assess normality

Data dictionary for fev, 1

```
---
data_dictionary: fev (.csv, sas7bdat, .sav, .txt)

copyright: >
  The author of the jse article holds the copyright, but does not list
  conditions under which it can be used. Individual use for educational
  purposes is probably permitted under the Fair Use provisions of
  U.S. Copyright laws.

description: >
  Forced Expiratory Volume (FEV) in children. The data was collected
  in Boston in the 1970s.

additional_description:
  https://jse.amstat.org/v13n2/datasets.kahn.html
```

Speaker notes

Here is a dataset you will need for your programming assignment. It is a study of pulmonary function in children.

Data dictionary for fev, 2

```
download_url:
  https://www.amstat.org/publications/jse/datasets/fev.dat.txt
format:
  csv: comma delimited
  sas7bdat: proprietary (SAS)
  sav: proprietary (SPSS)
  txt: fixed width

varnames:
  not included

missing_value_code:
  not needed

size:
  rows: 654
  columns: 5
```

Speaker notes

Data dictionary for fev, 3

```
vars:
  age:
    scale: ratio
    range: positive integer
    unit: years

  fev:
    label: Forced Expiratory Volume
    scale: ratio
    range: positive real
    unit: liters
```

Speaker notes

This is a small dataset with eight rows and three columns.

Data dictionary for fev, 4


```
  ht:
    label: Height
    scale: positive real
    unit: inches

  sex:
    value:
      0: Female
      1: Male

  smoke:
    value:
      0: Nonsmoker
      1: Smoker
---
```

Speaker notes

The variables are measurements before and after a major overhaul of the air conditioning system. The units are colonies per cubic foot of air. A pump pushes a certain volume of air through a filter and then bacterial colonies are allowed to grow on that filter.

simon-5501-03-fev.qmd, 1

```
---
title: "Analysis of fev data"
format:
  html:
    embed-resources: true
editor: source
---

This program assesses the normality of variables in a study of pulmonary
function in children. There is a [data dictionary][dd] 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.

[dd]: https://github.com/pmean/datasets/blob/master/fev.yaml
```

Speaker notes

The first few lines are the documentation header

simon-5501-03-fev.qmd, 2

````
## Libraries

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

```{r setup}
#| message: false
#| warning: false
library(tidyverse)
```
````

Speaker notes

Here is some additional documentation.

simon-5501-03-fev.qmd, 3

````
## List variable names

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

```{r names}
fev_names <- c(
    "age",
    "fev",
    "ht",
    "sex",
    "smoke")
```
````

Speaker notes

Loads the tidyverese library. No other libraries are needed.

simon-5501-03-fev.qmd, 4

````
## Reading the data

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

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

Speaker notes

Use the read_tsv function when your data uses tab delimiters.

simon-5501-03-fev.qmd, 5

````
## Calculate mean and standard deviation for fev

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

```{r descriptive-fev}
fev |>
  summarize(
    fev_mean=mean(fev),
    fev_stdv=sd(fev))
```
````

Speaker notes

Try to avoid spaces within a variable name. This code changes the space to an underscore.

simon-5501-03-fev.qmd, 6

````
## Histogram for fev, wide bars

```{r histogram-fev-wide}
ggplot(data=fev, aes(x=fev)) +
  geom_histogram(
    binwidth=0.5,
    color="black",
    fill="white")
```
````

Speaker notes

The tolower fuction replaces every uppercase letter with its lowercase equivalent.

simon-5501-03-fev.qmd, 7

````
## Histogram for fev, narrow bars

```{r histogram-fev-narrow}
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.
````

Speaker notes

This code produces a mean and standard deviation for the colony counts before remediation.

simon-5501-03-fev.qmd, 8

```
## 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.

```{r qqplot-fev}
#| fig-width: 5
#| fig.height: 5
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.
```

Speaker notes

This code produces a mean and standard deviation for the colony counts after remediation.

Break #4

- What you have learned
  - Using R to assess normality
- What's coming next
  - Your homework

Summary

- What you have learned
  - The normal distribution
  - Normal probabilities and quantiles
  - Assessing normality
  - Using R to assess normality
  - Your homework