Describing Shapes of Quantitative Variables with Histograms

David Gerard

2017-09-18

Learning Objectives

- Distributions
- Describe center/shape/spread of quantitive variables.
- Understand and use histograms.
- Section 1.6.3 of DBC.

A new dataset

A data frame with 1000 observations on the following 6 variables.

- `sex` Gender of the student.
- `SATV` Verbal SAT percentile.
- `SATM` Math SAT percentile.
- `SATSum` Total of verbal and math SAT percentiles.
- `HSGPA` High school grade point average.
- `FYGPA` First year (college) grade point average.

satGPA

```
library(tidyverse)
data(satGPA, package = "openintro")
glimpse(satGPA)
```

```
Observations: 1,000
Variables: 6
$ sex    <int> 1, 2, 2, 1, 1, 2, 1, 1, 2, 1, 1, 2, 2, 2...
$ SATV   <int> 65, 58, 56, 42, 55, 55, 57, 53, 67, 41, ...
$ SATM   <int> 62, 64, 60, 53, 52, 56, 65, 62, 77, 44, ...
$ SATSum <int> 127, 122, 116, 95, 107, 111, 122, 115, 1...
$ HSGPA  <dbl> 3.40, 4.00, 3.75, 3.75, 4.00, 4.00, 2.80...
$ FYGPA  <dbl> 3.18, 3.33, 3.25, 2.42, 2.63, 2.91, 2.83...
```

email

These data represent incoming emails for the first three months of 2012 for an email account.

Some variables:

- `spam` Indicator for whether the email was spam.
- `to_multiple` Indicator for whether the email was addressed to more than one recipient.
- `viagra` The number of times "viagra" appeared in the email.
- `num_car` The number of characters in the email, in thousands.

email

```
data("email", package = "openintro")
glimpse(select(email, spam, to_multiple,
               viagra, num_char))

Observations: 3,921
Variables: 4
$ spam        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
$ to_multiple <dbl> 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0,...
$ viagra      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
$ num_char    <dbl> 11.370, 10.504, 7.773, 13.256, 1.23...
```

Distribution

- How do we describe variables?
- How do we summarize their characteristics?
- What we are interested in is a variable's *distribution*.

distribution

The distribution of a variable tells us what values it takes and how often it takes these values.

There are two main ways we describe the distribution of a variable: *graphically* or *numerically*.

This lecture, we introduce one graphical way to describe the distribution of quantitative variables.

Histogram

histogram

Histograms plot the frequencies (counts), percents, or proportions of equal-width classes of values.

E.g.

```
x <- c(1, 1.2, 2, 3, 3.5, 3.9)
```

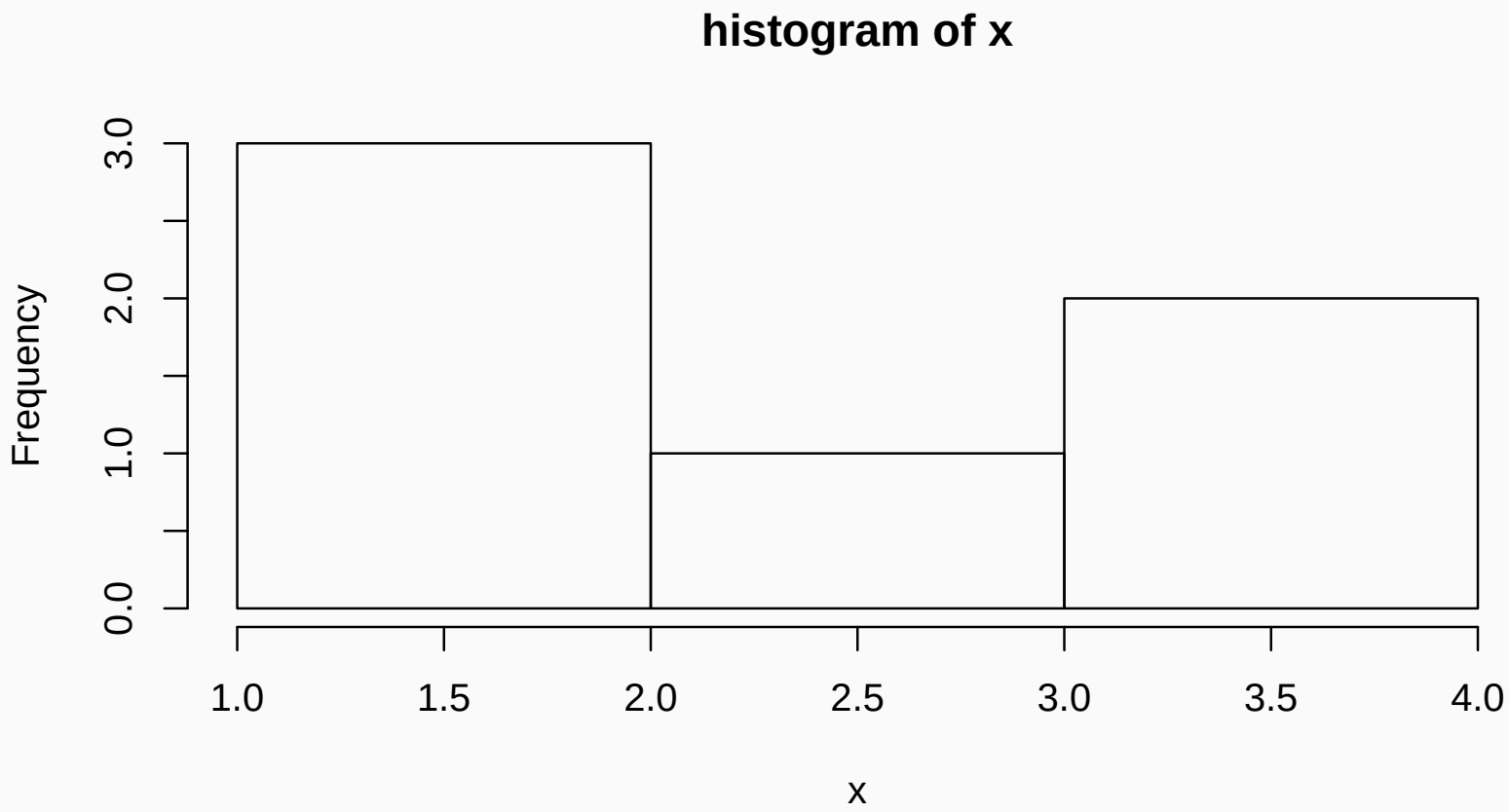
Bin the observations into one of three groups:

- group1 $= x : x \leq 2$
- group2 $= x : 2 < x \leq 3$
- group3 $= x : 3 < x \leq 4$

Then make a plot with bars where the height of each bar is proportional to the counts within each group.

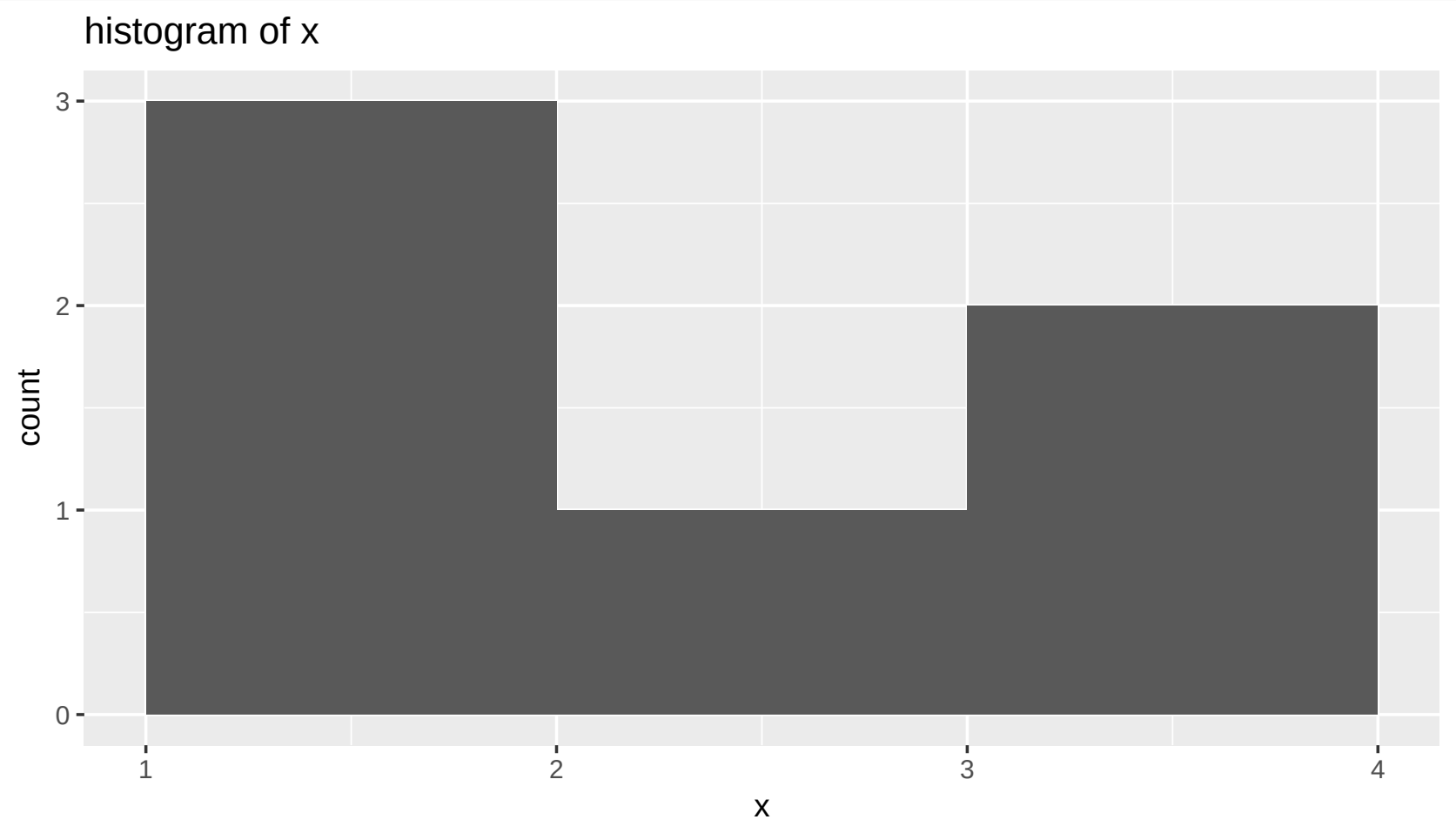
histogram continued

```
hist(x, main = "histogram of x")
```



histogram using ggplot2

```
qplot(x, geom = "histogram", main = "histogram of x",
      breaks = c(1, 2, 3, 4))
```

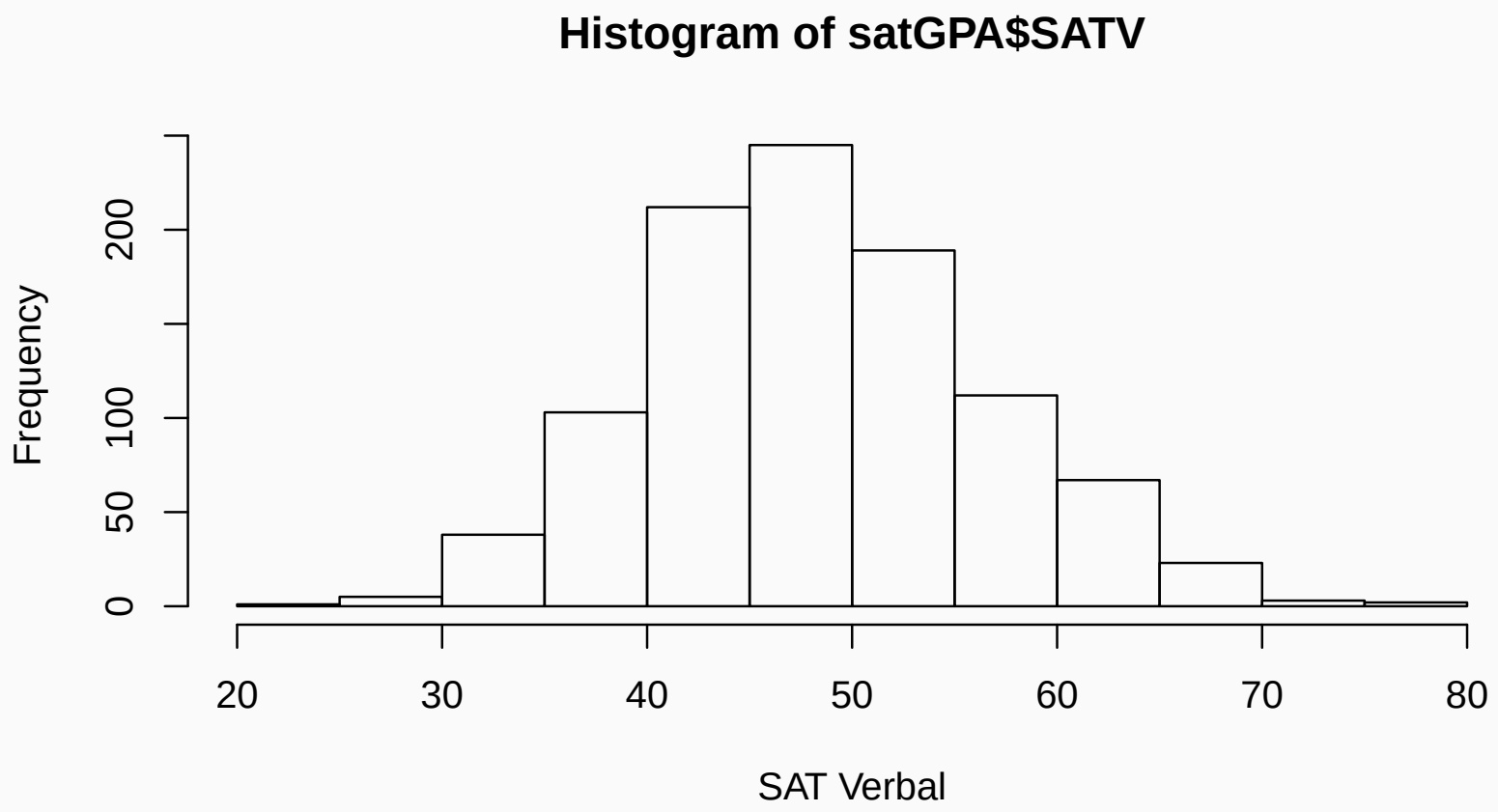


Describing Distributions

- Histograms help us describe the shape of a distribution.
- Symmetric vs skewed left vs skewed right.
- Unimodal, biomodal, multimodal.

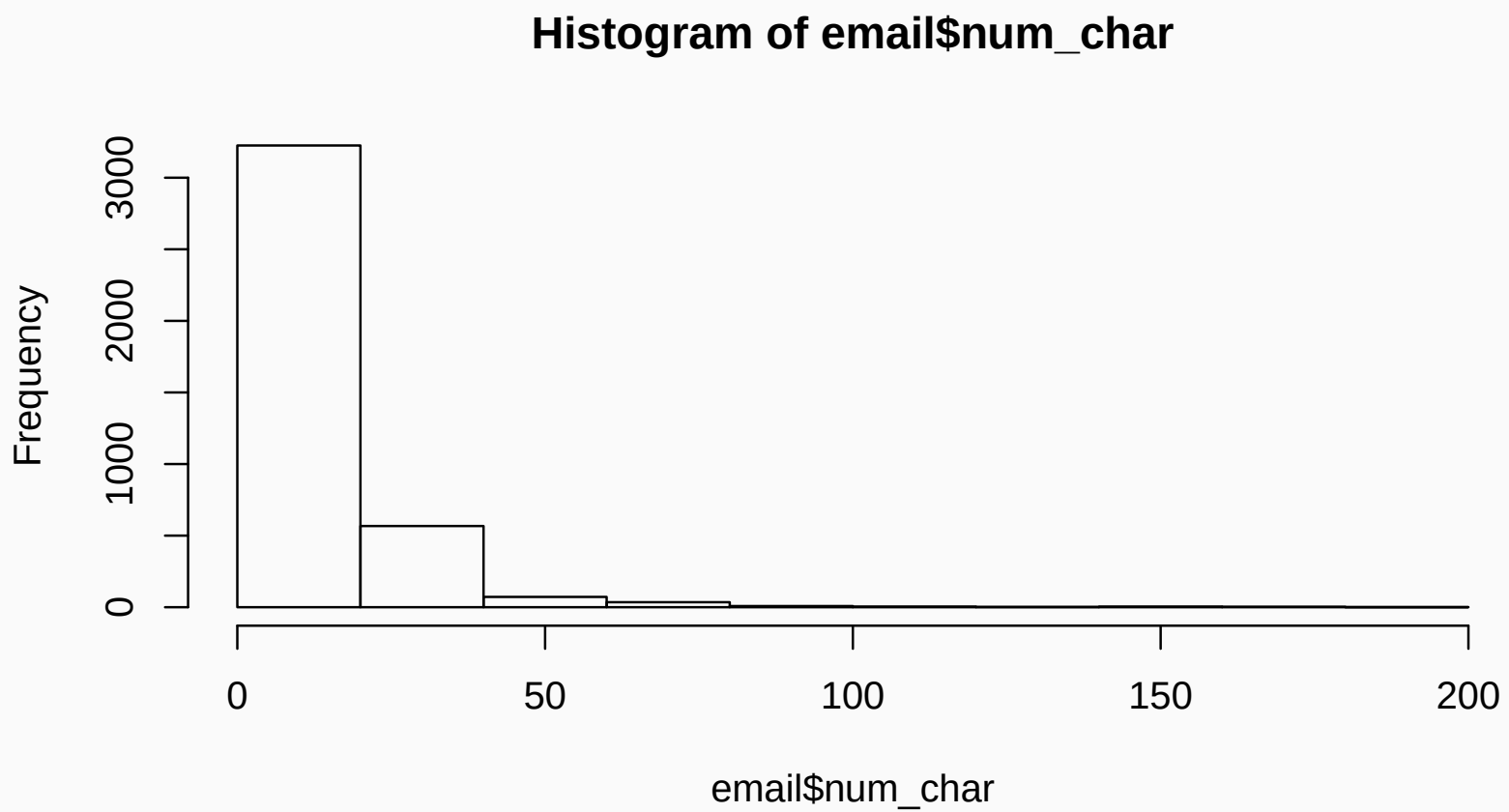
Symmetric — SAT scores

```
hist(satGPA$SATV, xlab="SAT Verbal", breaks = 15)
```

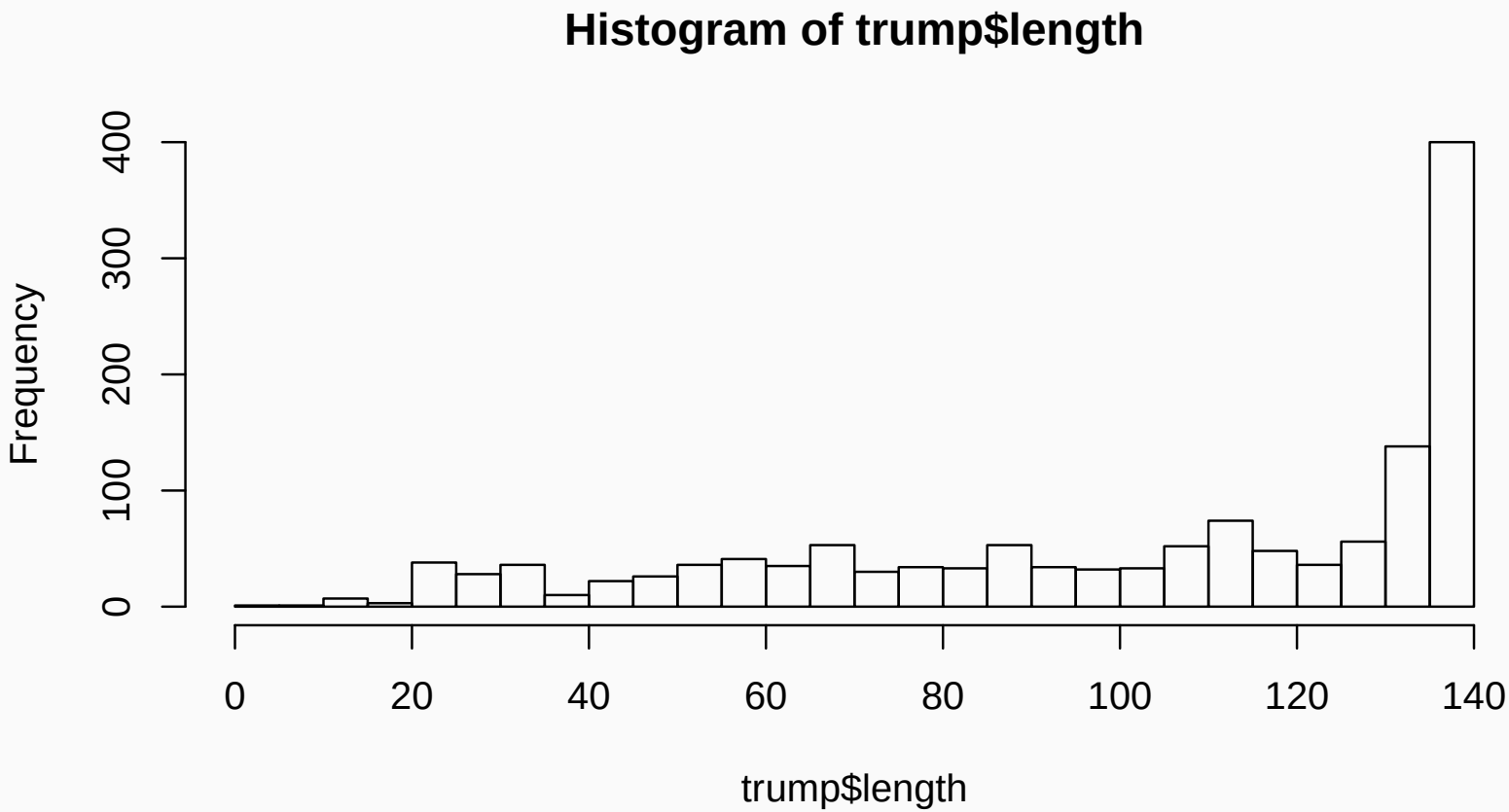


Skewed Right: Email Length

```
data("email", package = "openintro")
hist(email$num_char)
```



Skewed Left: Trump's Tweet Length



On Skew and Symmetry

- Many physical measurements follow symmetric distributions: e.g. height or weight.
- Many variables are specifically designed to follow symmetric distributions: IQ test scores, SAT scores.
- Variables with boundaries tend to be skewed: e.g. income cannot be below zero so tends to be skewed right. Tweets have a max length of 140 characters, so tends to be skewed left.

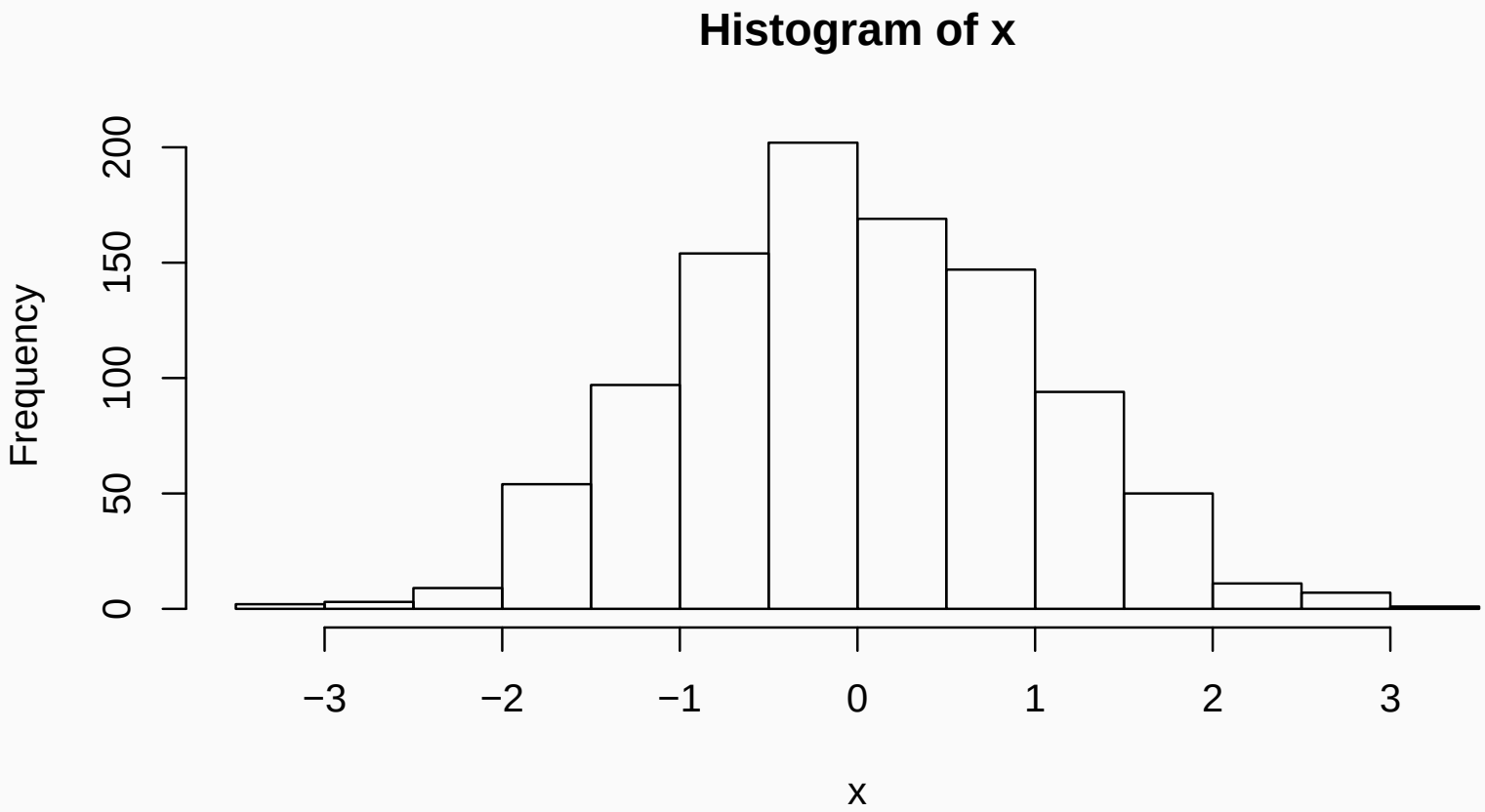
Mode

Mode

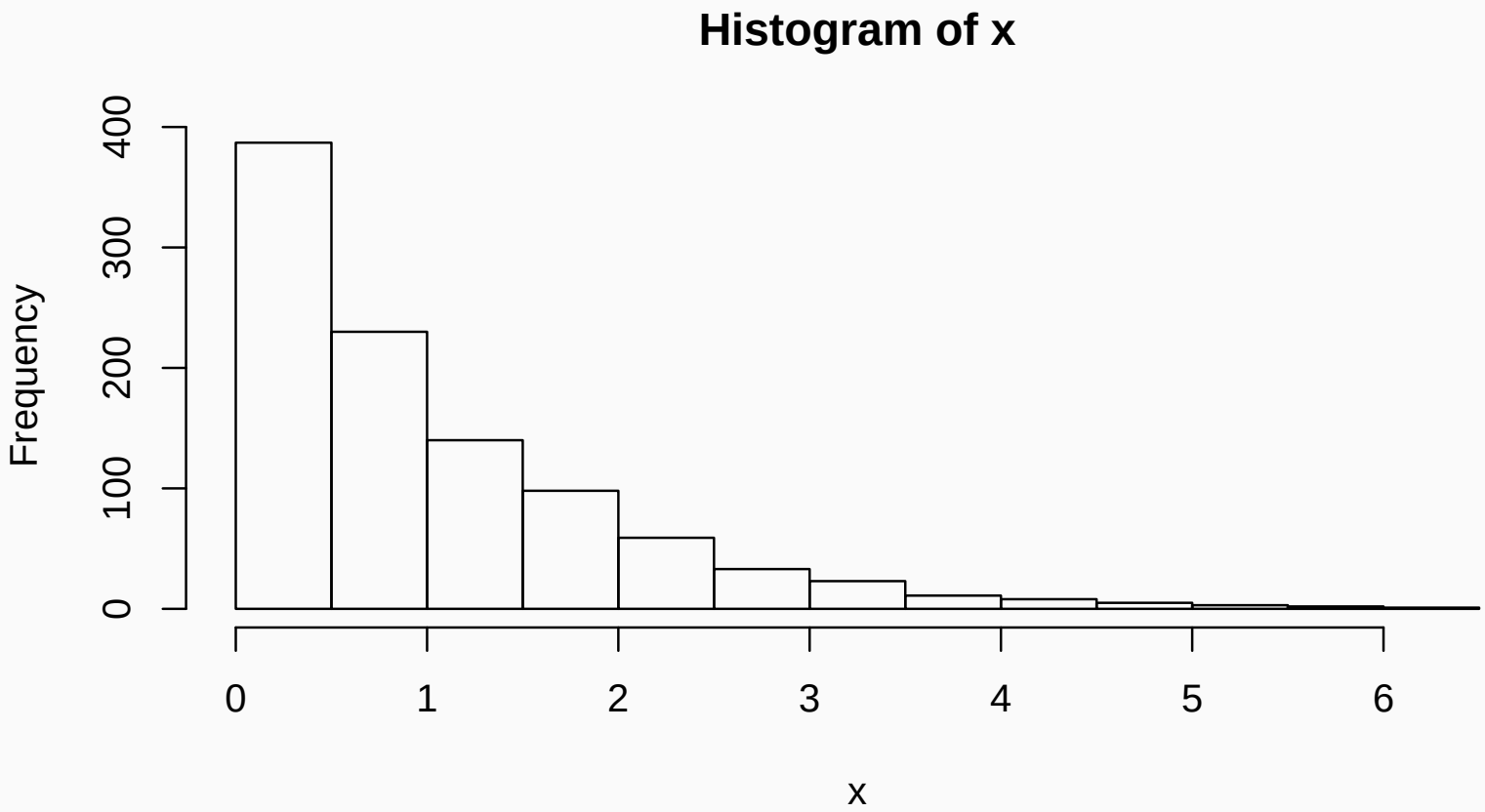
A mode is a prominent peak in a distribution. A distribution with one mode is unimodal. A distribution with two modes is bimodal. A distribution with more than one mode is multimodal.

- Multimodality often occurs when (and is usually interesting because) there are subgroups within the sample.

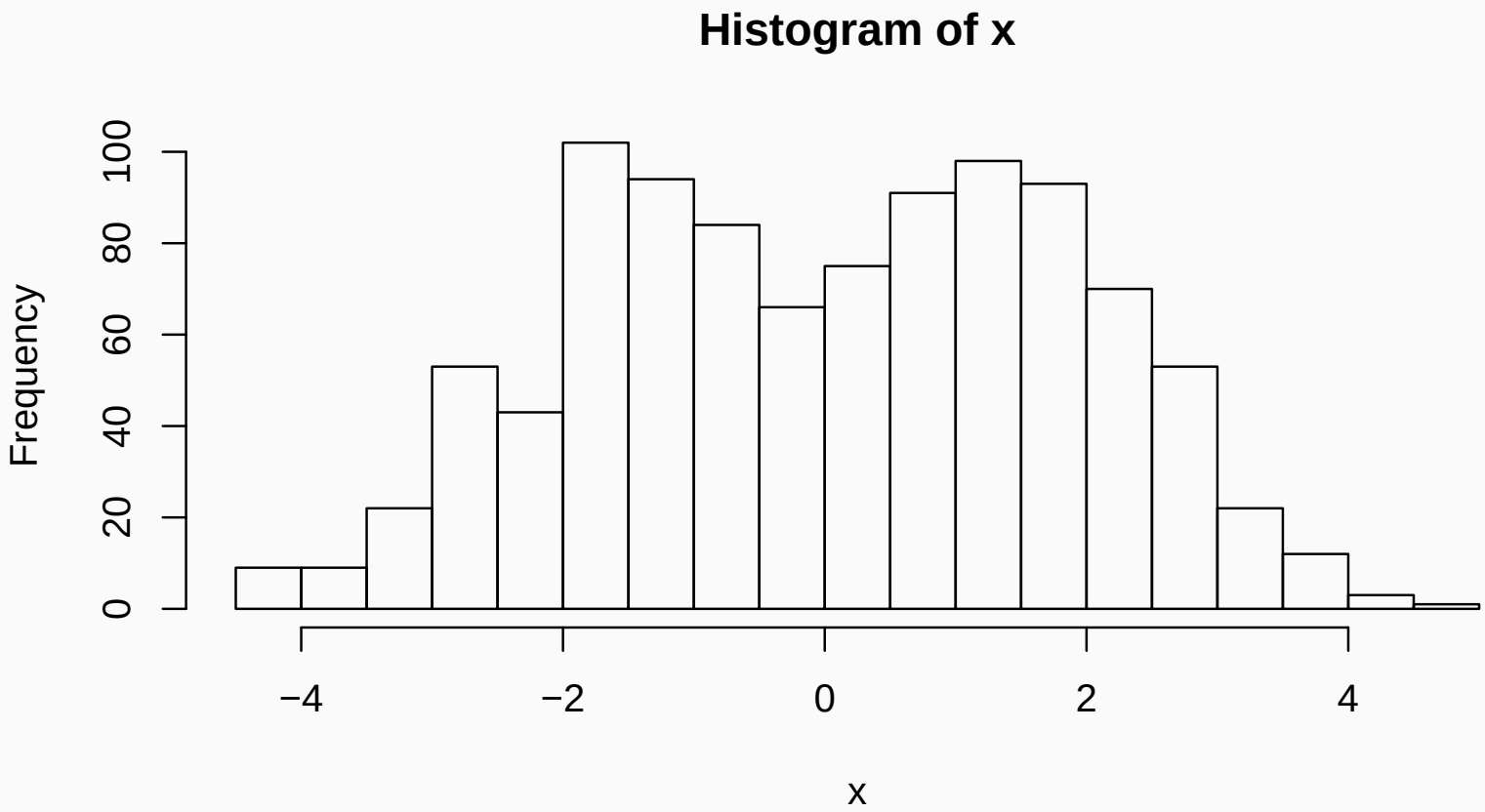
Unimodal



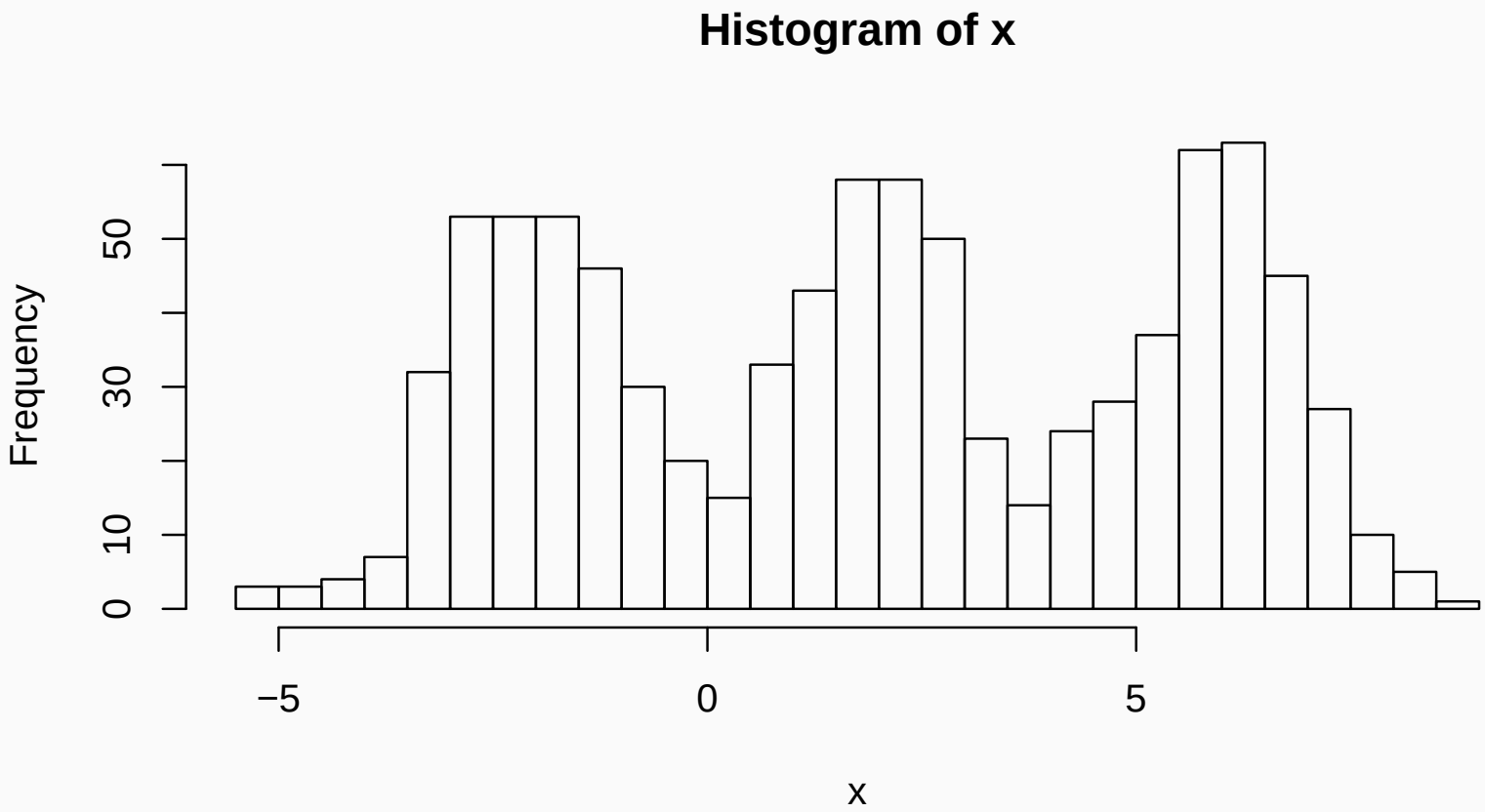
Unimodal



Bimodal



Multimodal



bin width

- Bin width can drastically change how you see the shape of the distribution.
- Always make multiple plots with multiple bin widths to get different views of a distribution.

A new dataset

Observational units: Movies that sold tickets in 2015.

Variables:

- `rt` Rotten tomatoes score normalized to a 5 point scale.
- `meta` Metacritic score normalized to a 5 point scale.
- `imdb` IMDB score normalized to a 5 point scale.
- `fan` Fandango score.

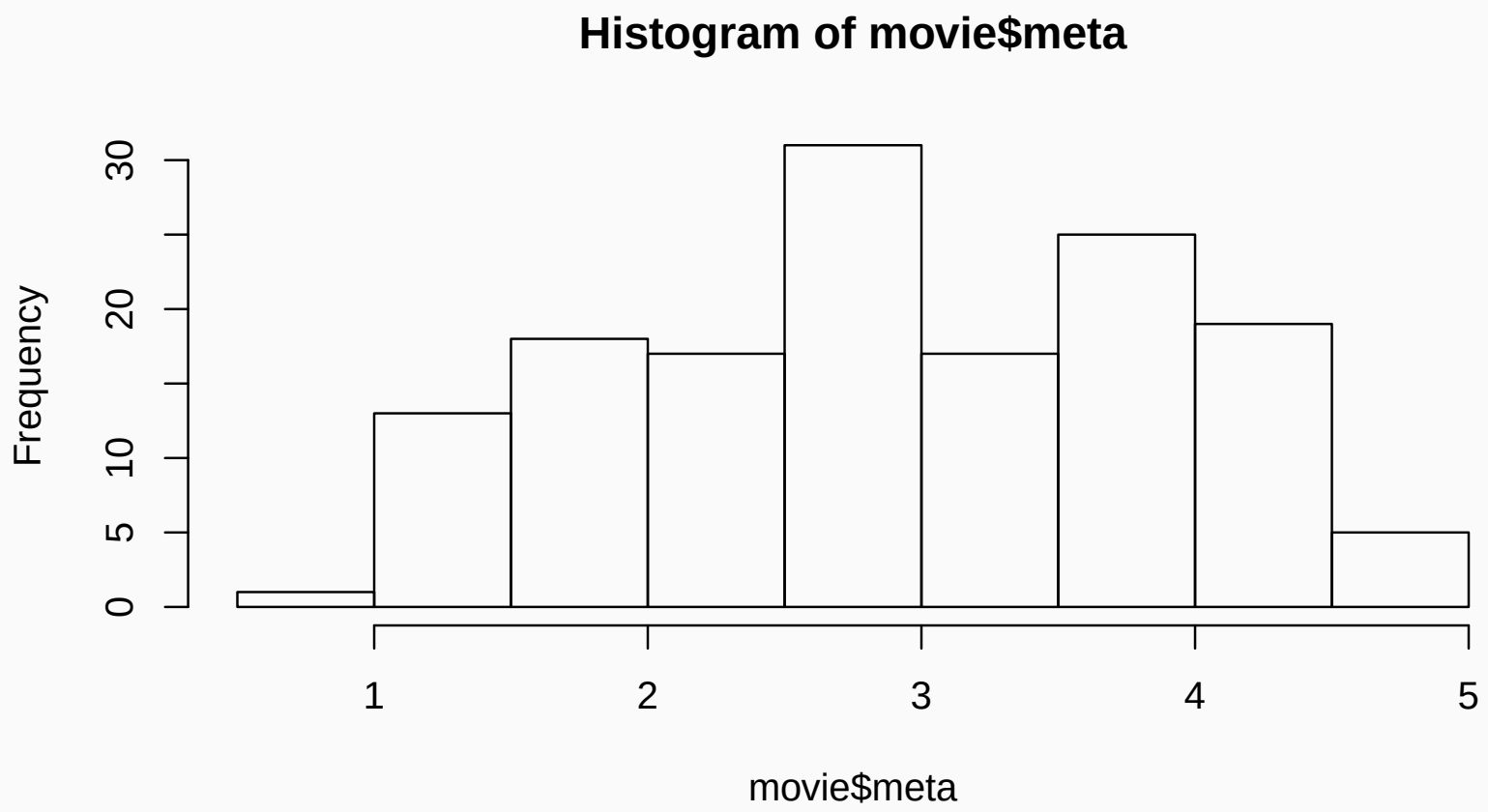
Movie Scores

```
library(tidyverse)
read_csv("../../data/movie.csv") %>%
  select(FILM, RT_norm, Metacritic_norm,
         IMDB_norm, Fandango_Stars) %>%
  transmute(film = FILM, rt = RT_norm, meta = Metacritic_norm,
            imdb = IMDB_norm, fan = Fandango_Stars) ->
  movie
head(movie)
```

```
# A tibble: 6 x 5
                                 film    rt  meta  imdb   fan
                                <chr> <dbl> <dbl> <dbl> <dbl>
1 Avengers: Age of Ultron (2015)  3.70  3.30  3.90   5.0
2              Cinderella (2015)  4.25  3.35  3.55   5.0
3                 Ant-Man (2015)  4.00  3.20  3.90   5.0
4         Do You Believe? (2015)  0.90  1.10  2.70   5.0
5  Hot Tub Time Machine 2 (2015)  0.70  1.45  2.55   3.5
6       The Water Diviner (2015)  3.15  2.50  3.60   4.5
```

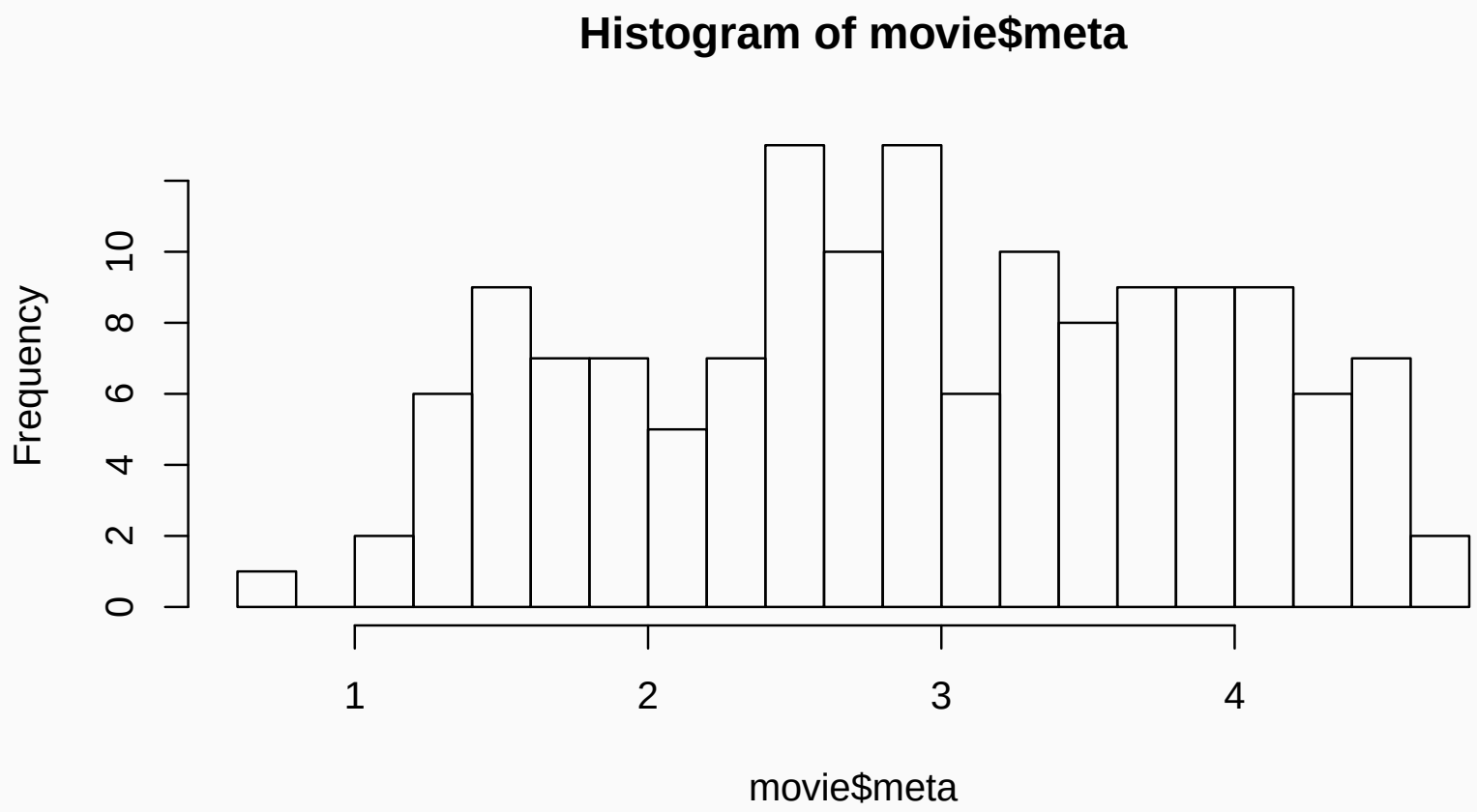
Metacritic Score: Mostly Symmetric?

```
hist(movie$meta, breaks = 10)
```



Metacritic Score: Maybe some Modality?

```
hist(movie$meta, breaks = 20)
```



Outliers

outliers

Outliers are observations that lie outside the overall pattern of a distribution. Always look for outliers and try to explain them.