Chapter 5: Methods for Describing a Numerical Variable

In this chapter, we will consider descriptive methods appropriate for summarizing numerical variables.

Example 5.1: IMDb Movie Reviews

A data set was collected on movies released in 2020. Here is a list of some of the variables collected on the observational units (each movie):

Variable	Description
`Movie`	Title of the movie
`averageRating`	Average IMDb user rating score from 1 to 10
`numVotes`	Number of votes from IMDb users
`Genre`	Categories the movie falls into (e.g., Action, Drama, etc.)
`2020 Gross`	Gross profit from movie viewing
`runtimeMinutes`	Length of movie (in minutes)

```
# A tibble: 6 x 6
  Movie                Genre  `2020 Gross` runtimeMinutes averageRating numVotes
  <chr>                <chr>         <dbl> <chr>                  <dbl>    <dbl>
1 1917                 Thril~    157901466 34                       5.7       23
2 The Invisible Man    Horror     64914050 71                       7.7    29256
3 The Call of the Wild Adven~     62342368 16                       5.3       51
4 Tenet                Action     58044165 150                      8.2    10174
5 Halloween            Horror     47274000 91                       7.8   222169
6 Little Women         Drama      37593127 60                       6.5       34
```

1. What is the variable of interest? What is the data type?

2. What is the observation?

The `favstats` function from the `mosaic` package will provide us with key summary statistics for a numerical variable:

```
library(mosaic)
favstats(~ averageRating, data = movie_ratings)
```

```
 min  Q1 median  Q3 max     mean       sd   n missing
 1.9 6.1      7 7.6 9.2 6.699535 1.252082 215       0
```

Next, let's discuss the summary statistics provided in each piece of the output:

Measures of location and variability

Measures of location These summaries give us an idea of where a data distribution lies.

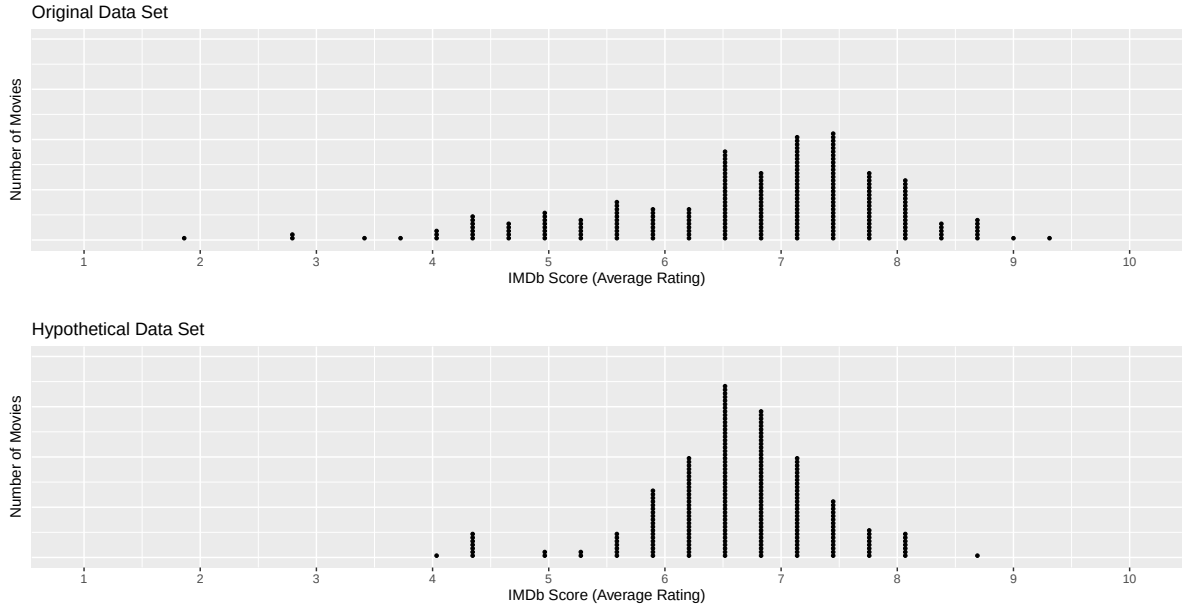
- **Mean** and **median** give us an idea of the center (or middle) of the distribution.
- The percentiles (**Q1** and **Q3**) give us an idea of what percent of the data distribution lies at or below a particular value.

What summary (or summaries) we choose to describe the entire data set depends on our objective. If the goal is to describe where a data distribution is centered, then the mean or median may be an appropriate summary statistic. However, if interest lies in what value is exceeded by only 5% of the data distribution, for example, then we would use the 95th percentile.

Measures of variability These summaries help us quantify how much the observations in a data set tend to vary from each other.

- The **sd** (standard deviation) is a measure of variability and quantifies how individual data points vary from the mean.
- The **IQR** (inner-quartile range) is the distance between Q1 and Q3 (the first and third quartiles).

For example, compare and contrast the variability in the following distributions. The first is the actual data distribution of IMDB movie ratings; the second is a hypothetical data set created for purposes of comparison. Which data set has more variability? Why?



3. How many observations are there in the data set?

4. What is the mean IMDb Score for the data set?

5. What is the lowest IMDb Score? The highest?

6. Interpret the value of the standard deviation in the context of these data.

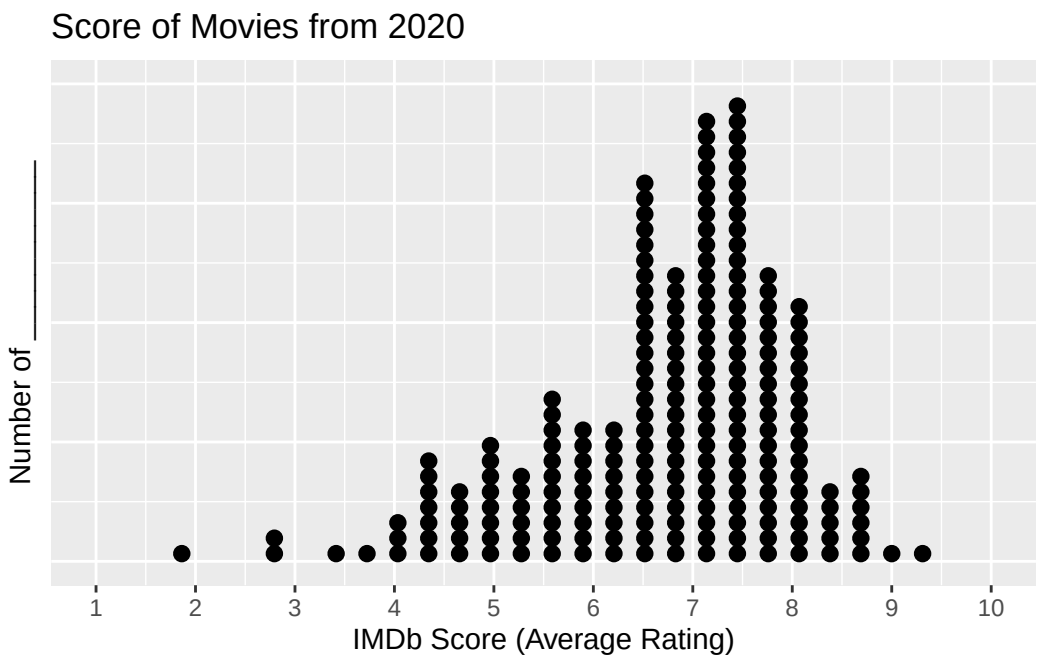
Graphical Summaries of Numerical Data

In this section, we will discuss common methods for graphing numerical data. Graphs conveniently allow us to examine both the location and the variability in a data set. Moreover, we gain insight into the shape of a data distribution.

Dotplots

A dotplot will plot a dot for each value in the data set. The code below was used to create a dotplot of the `averageRatings` variable from the movies data set. In a dotplot, the quantitative variable goes on the x-axis, which is why the code says `x = averageRating` inside of the `aes()` function.

```
ggplot(data = movie_ratings,
       mapping = aes(x = averageRating)) +
  geom_dotplot(dotsize = 0.5,
               method = "histodot") +
  labs(title = "Score of Movies from 2020", # Title for plot
       x = "IMDb Score (Average Rating)", # Label for x axis
       y = "Number of __________" # Label for y axis
       ) +
```

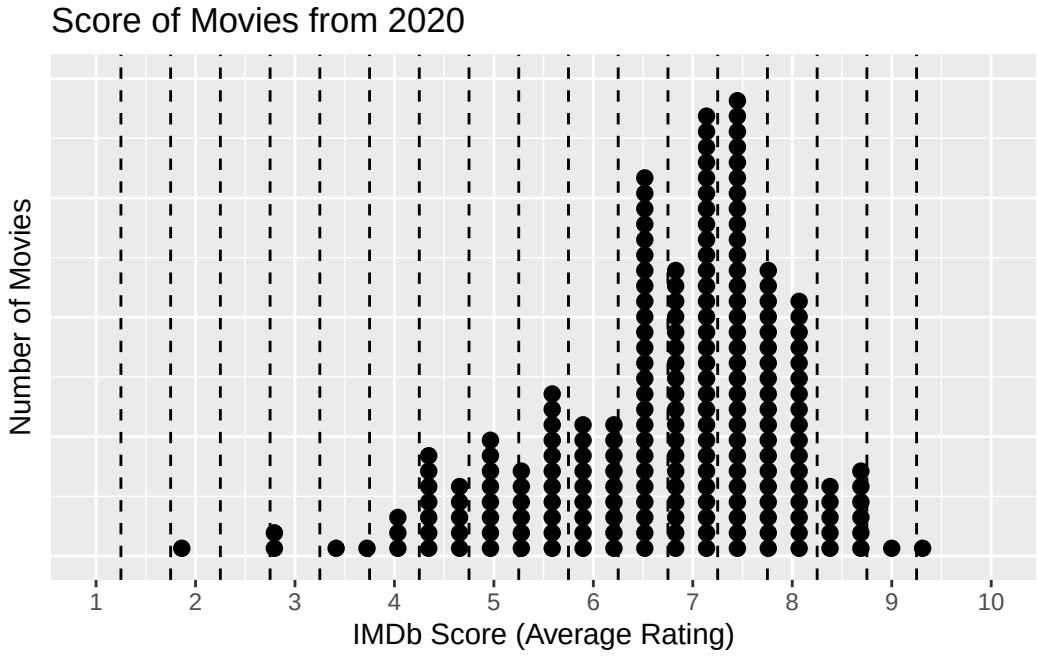


7. What does each dot on the dotplot represent?

8. How would you describe the shape of the distribution of IMDb scores? Think about measures of location and variability.

Histogram

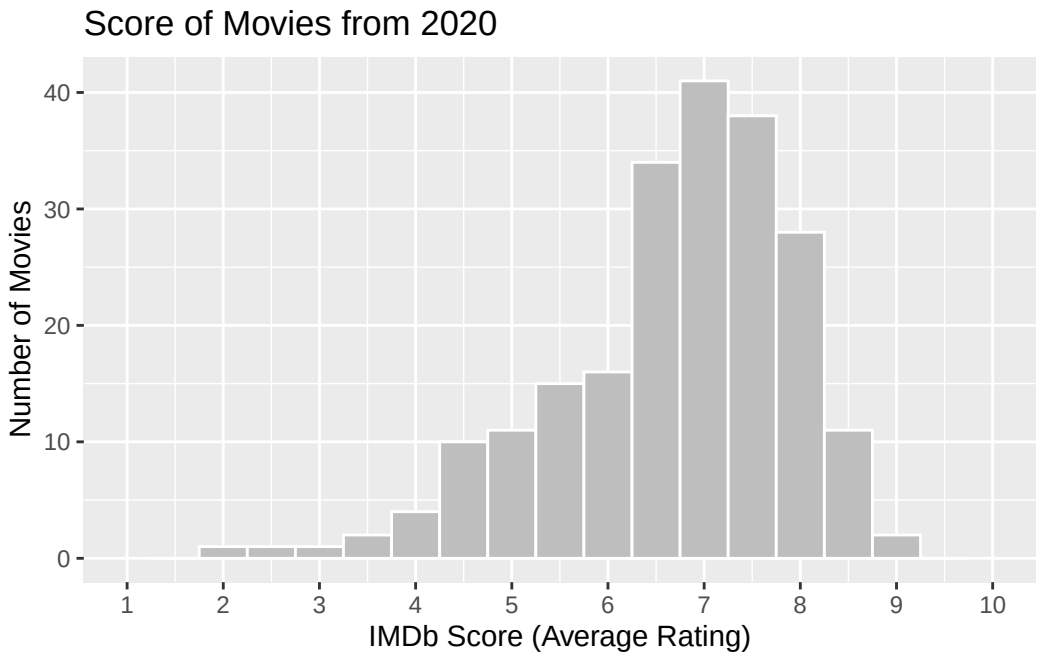
A histogram is created by dividing the range of the data distribution into bins and then counting the number of observations that fall in each bin A rectangular column is plotted in each interval, and the height of the column is proportional to the frequency of observations within the interval. The y-axis can be labeled with either the count or the percentage of the observations that fall in each interval.



To create a histogram of the IMDb scores, all we need to do is change the geometric object we are displaying on our plot! In a dotplot we use dots, but in a histogram we use bars. Notice, in the code below there are two changes:

- I am using `geom_histogram()` instead of `geom_dotplot()`
- I am specifying how wide the bins of the histogram should be using `binwidth = 0.5`

```
ggplot(data = movie_ratings,
       mapping = aes(x = averageRating)) +
  geom_histogram(binwidth = 0.5) +
  labs(title = "Score of Movies from 2020", # Title for plot
       x = "IMDb Score (Average Rating)", # Label for x axis
       y = "Number of Movies" # Label for y axis
       )
```



4. Which range of IMDb scores have the *highest* frequency (number of movies)?

5. What IMDB scores are movies *rarely* rated?

6. Are there IMDB scores that were possible but no movies in this sample were given those ratings?

Boxplot

The procedure for constructing a boxplot is as follows:

1. Draw horizontal lines at Q1 (25th percentile), Q2 (median / 50th percentile), and Q3 (75th percentile). Enclose these horizontal lines in a box.
2. Find the lower and upper whiskers:

- The endpoint of the lower whisker is the larger of the minimum and (Q1 – 1.5*IQR).
- The endpoint of the upper whisker is the smaller of the maximum and (Q3 + 1.5*IQR).

> **i** Outliers
>
> Any measurement beyond the endpoint of either whisker is classified as a potential outlier (an extreme observation).

Summary Statistics

```
 min  Q1 median  Q3 max     mean       sd   n missing
 1.9 6.1      7 7.6 9.2 6.699535 1.252082 215       0
```

Bottom 6:

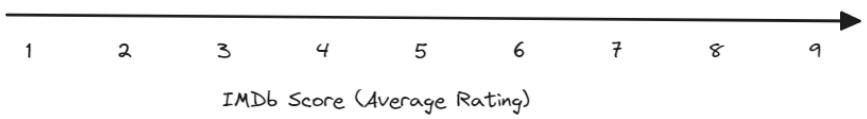
```
# A tibble: 7 x 1
  averageRating
          <dbl>
1           1.9
2           2.7
3           2.9
4           3.3
5           3.7
6           3.9
7           3.9
```

Top 6:

```
# A tibble: 6 x 1
  averageRating
          <dbl>
1           9.2
2           8.9
3           8.7
```

```
4           8.7
5           8.7
6           8.7
```

7. Using the summary statistics (same as we saw earlier) and top/bottom values of the data set provided, sketch a box plot of IMDb scores. *Hint: You will need to determine if there are any outliers.*



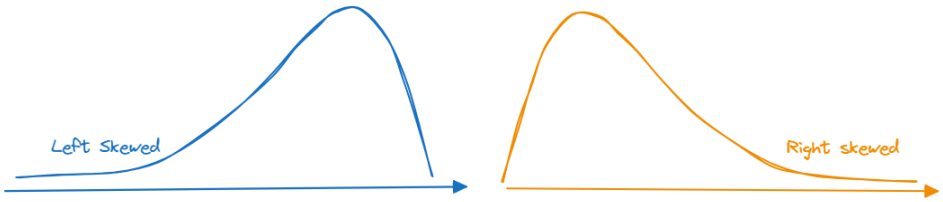
8. Are there any movies that are rated unusually low? If so, which ones?

9. Are there any movies that are rated unusually high? If so, which ones?

A Discussion of Skewness

A data distribution is said to be symmetric if it has the same shape on both sides of the center. Skewness measures the amount of asymmetry in a data distribution.

The distribution is said to be skewed to the right if the measurements tend to trail off to the right. Similarly, a distribution is skewed to the left if the measurements trail off to the left.



Describing distributions of quantitative variables

When describing distributions of quantitative variables we look at the **shape**, **center**, **spread**, and for **outliers**.

- There are two measures of center: mean and the median
- There are two measures of spread: standard deviation and the interquartile range, IQR = Q3 − Q1.

10. Compare the three graphs of IMDb scores created above.

- Which graph(s) show the shape of the distribution?

- Which graph(s) show the outliers of the distribution?

- Which graph plots the raw data (individual observations)?

Z-SCORES

A *Z-score*, often called a standardized value, measures the number of standard deviations a single observation is away from the mean of the data distribution. The z-score can be used to transform observations to a dimensionless scale; in addition, it can be used to measure the position of an observation. Z-scores are calculated as shown below:

$$\text{Z-score} = \frac{\text{Observation - Mean}}{\text{Standard Deviation}}$$

Interpretation of Z-Scores:

- As mentioned, the standardized values transform the data so that the data is placed on a standard, dimensionless scale that has a mean of 0 and a standard deviation of 1.
- If a Z-Score is negative, then the observation is that many standard deviations below the mean.
- If the Z-Score is positive, then the observation is that many standard deviations above the mean.
- If the Z-Score is 0, then the data value is the same as the mean.
- If the Standard Deviation is 0, then the Z-Score is not defined and thus cannot be computed.

```
# A tibble: 6 x 3
  Movie                averageRating Zscore
  <chr>                        <dbl>  <dbl>
1 1917                           5.7 -0.798
2 The Invisible Man              7.7  0.799
3 The Call of the Wild           5.3 -1.12
4 Tenet                          8.2  1.20
5 Halloween                      7.8  0.879
6 Little Women                   6.5 -0.159
```

11. Show how the Z-score for "Halloween" was calculated:

12. What does this tell you about the relative position of "Halloween" in the data set?

13. Show how the Z-score for "The Call of the Wild" was calculated:

14. What does this tell you about the relative position of "The Call of the Wild" in the data set?

The Identification of Outliers

We have already discussed using boxplots to identify outliers. In addition, we can use Z-scores.

- Any data value whose Z-Score is below -2 or above 2 is considered a potential outlier.
- Any data value whose Z-Score is below -3 or above 3 is considered an outlier and warrants further investigation.

These guidelines come from the Empirical Rule: If the probability distribution is bell-shaped and symmetric, then the Empirical Rule applies. This rule says that APPROXIMATELY…

- 68% of the data values fall within one standard deviation of the mean.
- 95% of the data values fall within two standard deviations of the mean.
- 99.7% of the data values fall within three standard deviations of the mean.