Histogram: Center, Spread, & Outliers

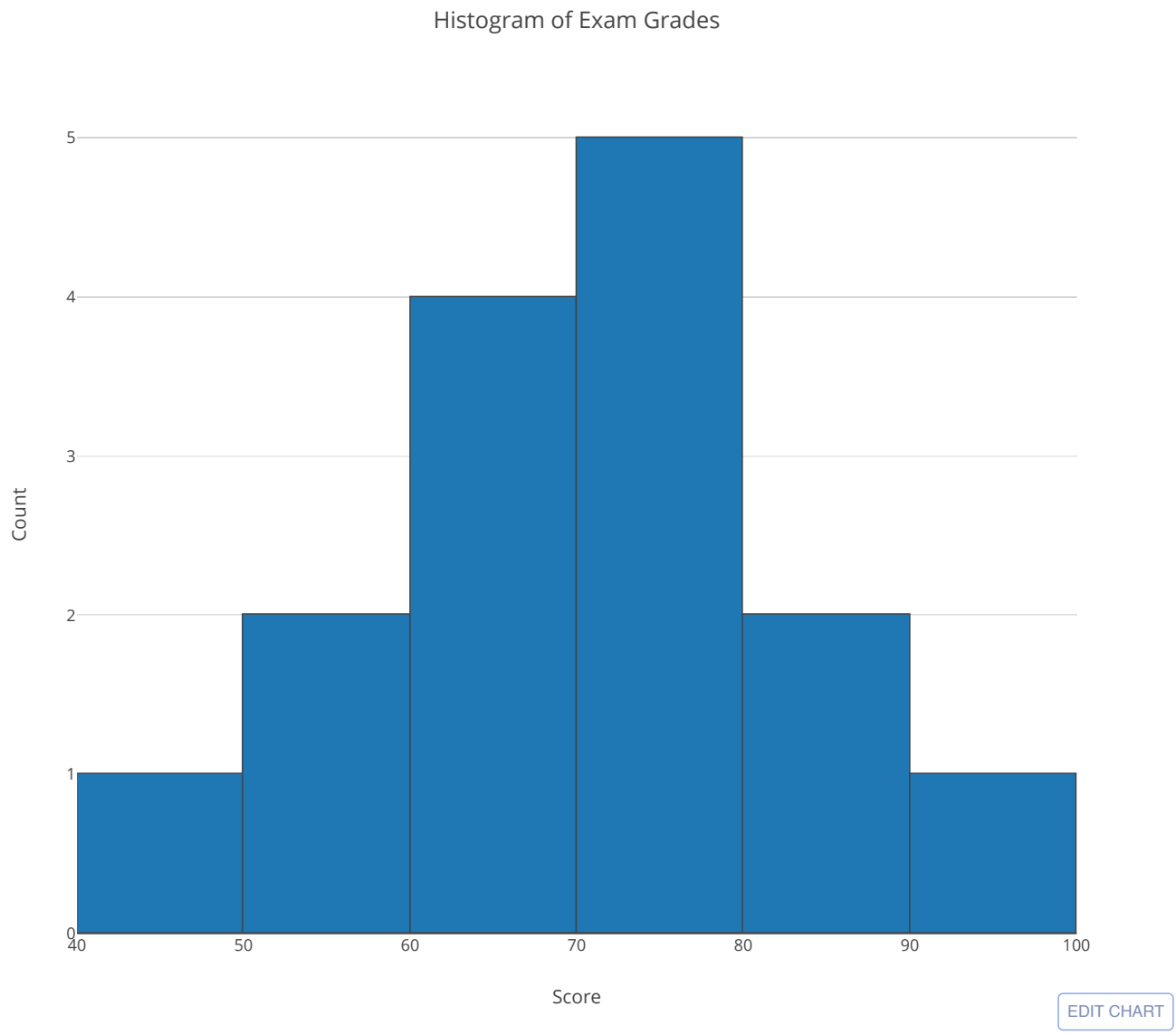
Learning Objective: Generate and interpret several different graphical displays of the distribution of a quantitative variable (histogram, stemplot, boxplot).

Learning Objective: Summarize and describe the distribution of a quantitative variable in context: a) describe the overall pattern, b) describe striking deviations from the pattern.

Center

The center of the distribution is its **midpoint**—the value that divides the distribution so that approximately half the observations take smaller values, and approximately half the observations take larger values. Note that from looking at the histogram we can get only a rough estimate for the center of the distribution. (More exact ways of finding measures of center will be discussed in the next section.)

Recall our grades example:



As you can see from the histogram, the center of the grades distribution is roughly 70 (7 students scored below 70, and 8 students scored above 70).

Spread

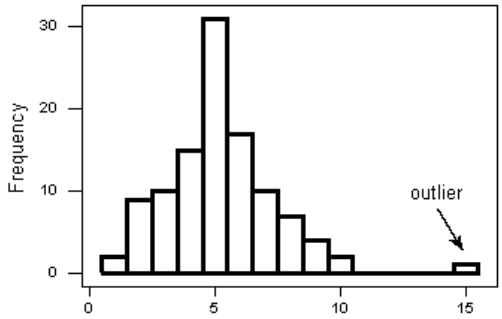
The **spread** (also called **variability**) of the distribution can be described by the approximate range covered by the data. From looking at the histogram, we can approximate the smallest observation (**min**), and the largest observation (**max**), and thus approximate the range. (More exact ways of finding measures of spread will be discussed in the next section.)

In our example:

- approximate min: 45 (the middle of the lowest interval of scores)
- approximate max: 95 (the middle of the highest interval of scores)
- approximate range: 95-45=50

Outliers

Outliers are observations that fall outside the overall pattern. For example, the following histogram represents a distribution that has a high probable outlier:

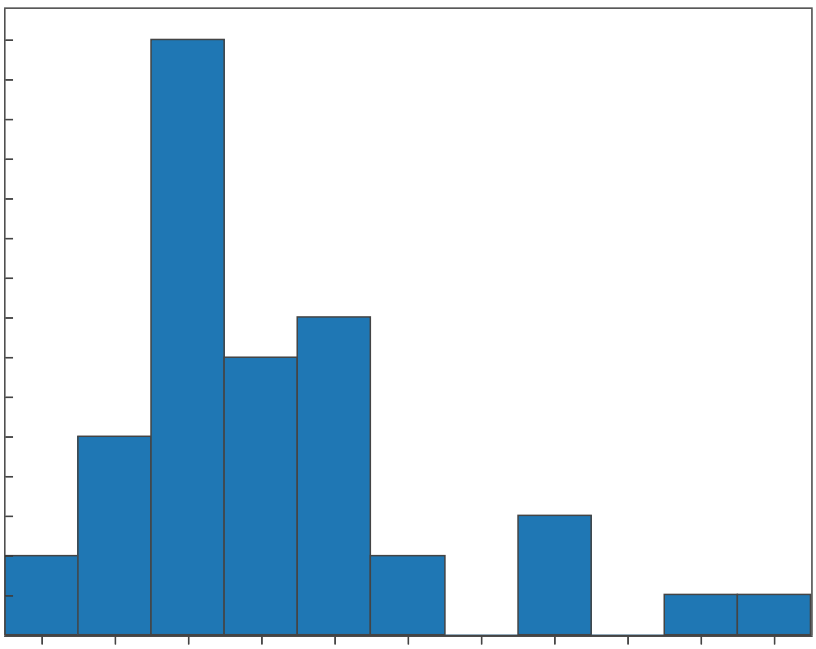


Go back and check the histogram of scores at the top of this page. As you can see, there are no outliers.

Example: Best Actress Oscar Winners

To provide an example of a histogram applied to actual data, we will look at the ages of Best Actress Oscar winners from 1970 to 2013 (To see the full dataset, click here).

The histogram for the data is shown below.



EDIT CHART

We will now summarize the main features of the distribution of ages as it appears from the histogram:

Shape: The distribution of ages is skewed right. We have a concentration of data among the younger ages and a long tail to the right. The vast majority of the "best actress" awards are given to young actresses, with very few awards given to actresses who are older.

Center: The data seem to be centered around 34 or 35 years old. Note that this implies that roughly half the awards are given to actresses who are less than 34 years old.

Spread: The data range from about 20 to about 80, so the approximate range equals 80 - 20 = 60.

Outliers: There seem to be two probable outliers to the far right and possibly three around 62 years old.

You can see how informative it is to know "what to look at" in a histogram. If there is one conclusion that we can make here, it is that Hollywood likes its actresses young.

We will use the Best Actor Oscar winners (1970-2013) dataset to practice what you've learned about describing the histogram. Below is the histogram of Best Actor Oscar winners from 1970 - 2013 grouped by age .

Edit chart
Loading graph

Learn By Doing

1/1 point (graded)
What is the shape of this histogram?

- ○ Symmetric—Uniform
- ○ Skewed Left
- ● Skewed Right ✔
- ○ Symmetric—Unimodal

Symmetric—Bimodal

Answer
Correct: The bulk of Best Actor Oscar winners are from younger age groups.

Submit

Did I Get This

1/1 point (graded)
What is the most likely shape of the distribution of age of death from trauma (accident, murder, suicide, drug overdose, etc.) when represented by a histogram?

Recall that we talked earlier about the shape of the distribution of age of death from natural causes (heart disease, cancer, etc.). Use a similar type of reasoning for the age of death from trauma.

Symmetric—Uniform

Skewed Left

Skewed Right ✔

Symmetric—Unimodal

Symmetric—Bimodal

Answer
Correct:
The bulk of deaths from trauma (accidents, suicide, drug overdose, etc.) happen at a younger age, and fewer at an older age. Therefore, we expect the distribution of age of death from trauma to be skewed right.

Submit

Did I Get This

1/1 point (graded)
What is the most likely shape of the distribution of the age at which a child takes its first steps?

Symmetric—Uniform

Skewed Left

Skewed Right

Symmetric—Unimodal ✔

Symmetric—Bimodal

Answer
Correct:
Most children start to walk at about the same age, so the distribution is centered at about 18 months. It has some variability, but is unlikely to have outliers. Thus, it is symmetric and unimodal.

Submit

Let's Summarize

- The histogram is a graphical display of the distribution of a quantitative variable. It plots the number (count) of observations that fall in intervals of values.
- When examining the distribution of a quantitative variable, one should describe the overall pattern of the data (shape, center, spread), and any deviations from the pattern (outliers).
- When describing the shape of a distribution, one should consider:
    - Symmetry/skewness of the distribution
    - Peakedness (modality)—the number of peaks (modes) the distribution has.

  Not all distributions have a simple, recognizable shape.
- Outliers are data points that fall outside the overall pattern of the distribution and need further research before continuing the analysis.
- It is always important to interpret what the features of the distribution (as they appear in the histogram) mean in the context of the data.