# Histogram: Intervals

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

## Idea

Break the range of values into intervals and count how many observations fall into each interval.

### Example: Exam Grades

Here are the exam grades of 15 students:

88, 48, 60, 51, 57, 85, 69, 75, 97, 72, 71, 79, 65, 63, 73

We first need to break the range of values into intervals (also called "bins" or "classes"). In this case, since our dataset consists of exam scores, it will make sense to choose intervals that typically correspond to the range of a letter grade, 10 points wide: 40-50, 50-60, ... 90-100. By counting how many of the 15 observations fall in each of the intervals, we get the following table:

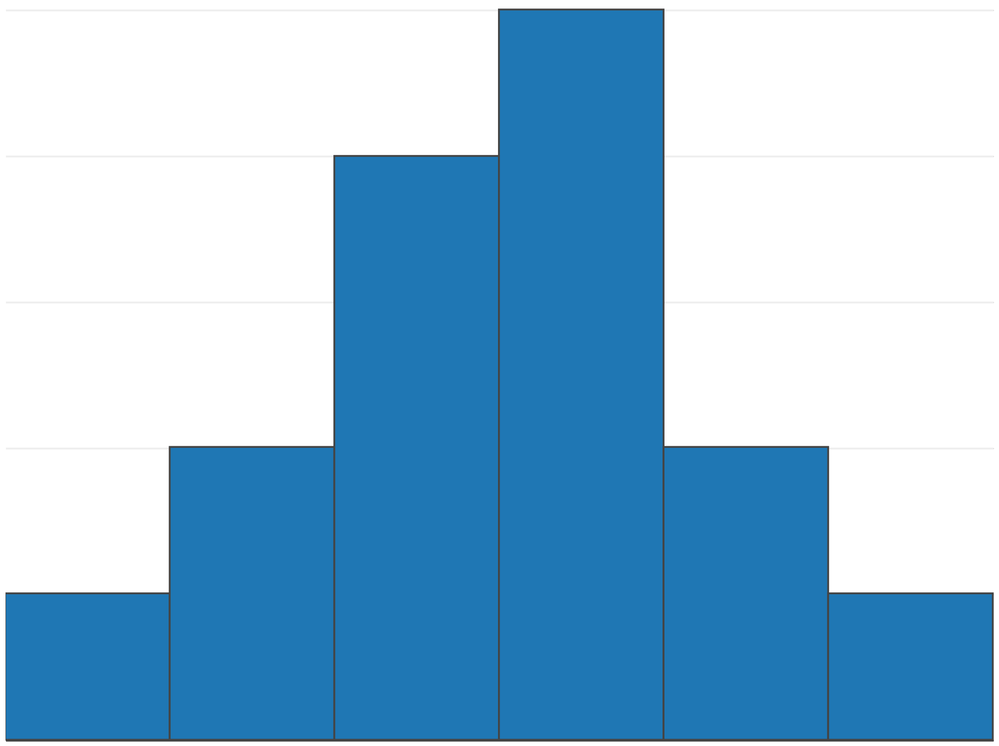
Exam Grades

| Score | Count |
|---|---|
| [40-50) | 1 |
| [50-60) | 2 |
| [60-70) | 4 |
| [70-80) | 5 |

| | |
|---|---|
| [80-90) | 2 |
| [90-100] | 1 |

Note: The observation 60 was counted in the 60-70 interval. See comment 1 below.

To construct the histogram from this table we plot the intervals on the X-axis, and show the number of observations in each interval (frequency of the interval) on the Y-axis, which is represented by the height of a rectangle located above the interval:



EDIT CHART

The table above can also be turned into a relative frequency table using the following steps:

1. Add a row on the bottom and include the total number of observations in the dataset that are represented in the table.
2. Add a column, at the end of the table, and calculate the relative frequency for each interval, by dividing the number of observations in each row by the total number of observations.

These two steps are illustrated in red in the following frequency distribution table:

| Score | Count (also called Frequency) | Relative Frequency |
|---|---|---|
| [40–50) | 1 | 0.07 |
| [50–60) | 2 | 0.13 |
| [60–70) | 4 | 0.27 |
| [70–80) | 5 | 0.33 |
| [80–90) | 2 | 0.13 |
| [90–100] | 1 | 0.07 |
| Total | 15 | |

Step 1: Add a row at bottom of table. Put in total number of observations in the data set.

In this example, there are 15 (1+2+4+5+2+1= 15) total observations.

Step 2: Add a column to right side of table. Determine the relative frequencies of each interval by dividing the interval count (or frequency) by the total number of observations.

For example, to determine the relative frequency of scores in the [40-50) interval, divide the count (or frequency) by the total number of observations: 1 /15 = .07.

The relative frequency for the [50-60) interval is: 2/15 = .13.

Continue until all of the relative frequencies have been calculated.

To convert each relative frequency into a percentage, multiply it by 100. For example, the percentage of scores for the [40-50) interval would be .07*100= 7, which is 7%.

It is also possible to determine the number of scores for an interval, if you have the total number of observations and the relative frequency for that interval. For instance, if we know that we have 15 scores (or observations) and the relative frequency is 0.13, we can determine the number of scores by multiplying the total number of observations by the relative frequency and rounding up to the next whole number: 15 * 0.13 = 1.95, which rounds up to 2 observations.

A relative frequency table, like the one above, can be used to determine the frequency of scores occurring at or across intervals. Here are some examples, using the above frequency table:

1. What is the percentage of exam scores that were 70 and up to, but not including, 80? To determine the answer, we look at the relative frequency associated with the [70-80) interval. The relative frequency is 0.33; to convert to percentage, multiply by 100 (0.33 * 100 = 33) or 33%.
2. What is the percentage of exam scores that are at least 70? To determine the answer, we need to:
   - Add together the relative frequencies for the intervals that have scores of at least 70 or above. Thus, would need to add together the relative frequencies from [70-80), [80-90), and [90-100] = 0.33 + 0.13 + 0.07 = 0.53.
   - To get the percentage, need to multiple the calculated relative frequency by 100. In this case, it would be 0.53 * 100 = 53 or 53%.

## Learn By Doing

1/1 point (graded)

Recall the table from the exam grades example above:

Exam Grades

| Score | Count |
|---|---|
| [40-50) | 1 |
| [50-60) | 2 |
| [60-70) | 4 |
| [70-80) | 5 |
| [80-90) | 2 |
| [90-100] | 1 |

What percentage of students earned less than a grade of 70 on the exam?

- 9%
- 15%
- 47% ✔
- 53%
- 93%

**Answer**
Correct:
The data displays information about a total of 1 + 2 + 4 + 5 + 2 + 1 = 15 students. Out of these 15 students, 1 + 2 + 4 = 7 earned less than a grade of 70 on the exam. To calculate the percentage, divide 7 (the number of students who earned less than a grade of 70) by 15 (the total number of students), and then multiply by 100 to change the decimal to a percentage: 7/15 = 0.47 * 100 = 47% of the students.

Submit

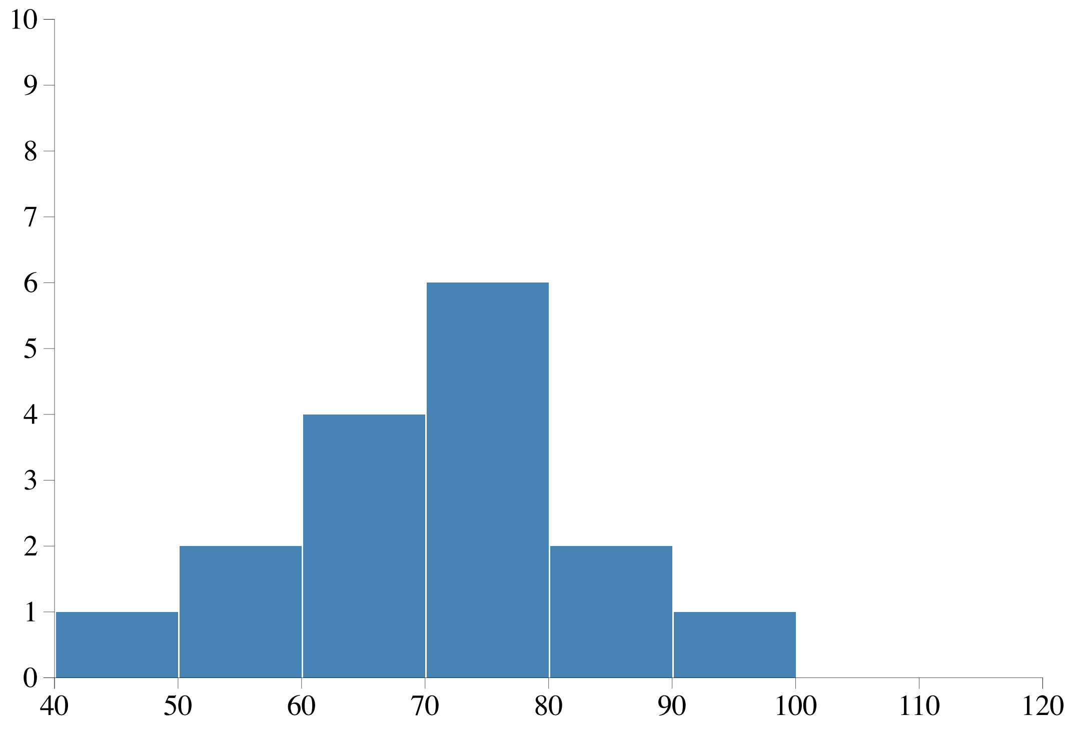
## Comments

1. It is very important that each observation be counted only in one interval. For the most part, it is clear which interval an observation falls in. However, in our example, we needed to decide whether to include 60 in the interval 50-60, or the interval 60-70, and we chose to count it in the latter. In fact, this decision is captured by the way we wrote the intervals. If you'll scroll up and look at the table, you'll see that we wrote the intervals in a peculiar way: [40-50), [50,60), [60,70)

etc. The square bracket means "including" and the parenthesis means "not including". For example, [50,60) is the interval from 50 to 60, including 50 and not including 60; [60,70) is the interval from 60 to 70, including 60, and not including 70, etc. It really does not matter how you decide to set up your intervals, as long as you're consistent.

2. When data are displayed in a histogram, some information is lost. Note that by looking at the histogram we *can* answer: "How many students scored 70 or above?" (5+2+1=8) But we *cannot* answer: "What was the lowest score?" All we can say is that the lowest score is somewhere between 40 and 50, and therefore we can approximate that it is around 45.

3. Obviously, we could have chosen to break the data into intervals differently (for example: 45-50, 50-55, 55-60 etc.). To see how our choice of bins or intervals affects the histogram, you can use the interactive simulation that lets you change the intervals dynamically. Try changing the bin width by dragging the slider underneath the bin width scale.




Move slider to select bin width: 10

or choose a value: 1 2 5 10 20 25

Show individual grades

## Many Students Wonder ...

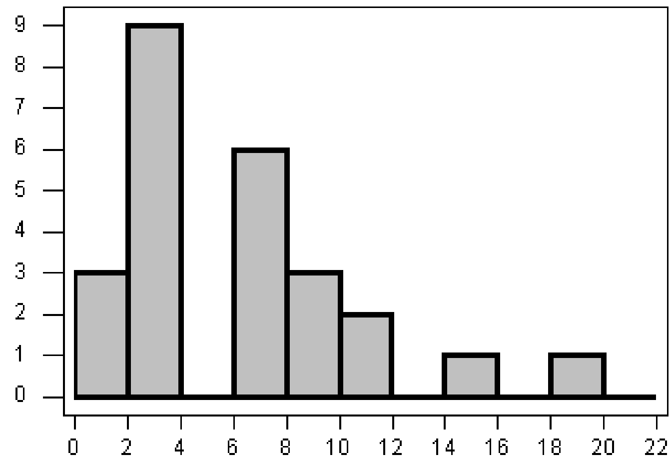
**Question:** How do I know what interval width to choose?

**Answer:** There are no right or wrong choices of interval widths. In this course, we will rely on a statistical package to produce the histogram for us, and we will focus instead on describing and summarizing the distribution as it appears from the histogram.

## Did I Get This

1/1 point (graded)

An instructor asked her students how much time (to the nearest hour) they spent studying for the midterm. The data are displayed in the following histogram:




What do the numbers on the horizontal axis represent?

- (selected) The values of the number of hours studied. ✔
- The count of students falling in each of the intervals.

**Answer**

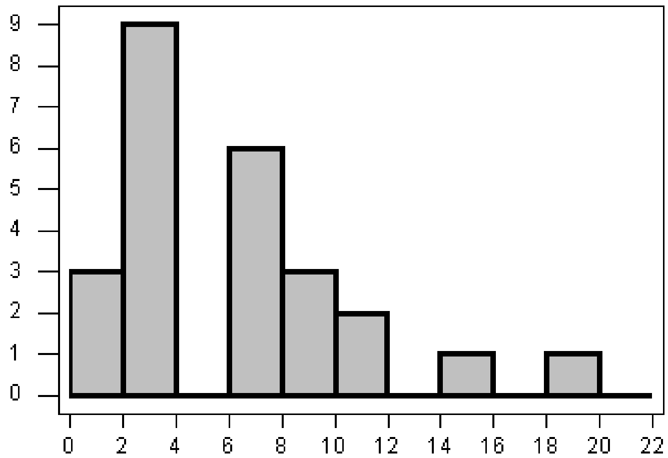
Correct: The horizontal axis represents the number of hours studied.

Submit

## Did I Get This

1/1 point (graded)

An instructor asked her students how much time (to the nearest hour) they spent studying for the midterm. The data are displayed in the following histogram:




What do the numbers on the vertical axis represent?

- ○ The values of the number of hours studied.
- ◉ The count of students falling in each of the intervals. ✔

**Answer**

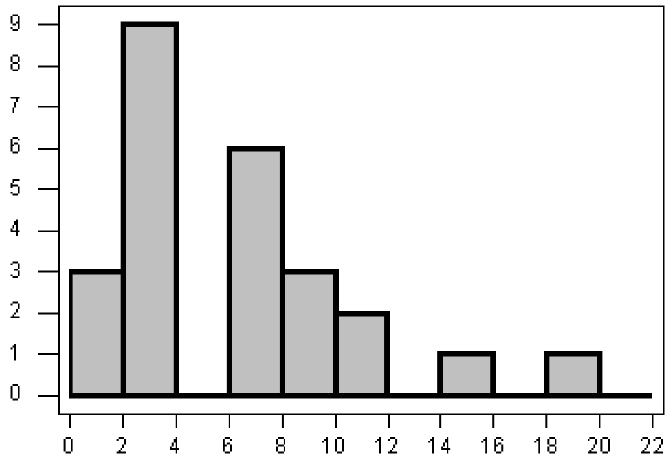
Correct: The vertical axis represents the count of students falling in each of the intervals.

Submit

## Did I Get This

1/1 point (graded)

An instructor asked her students how much time (to the nearest hour) they spent studying for the midterm. The data are displayed in the following histogram:




What percentage of students study 6 or more hours for the midterm?

- 24%
- 48%
- 52% ✔
- 72%
- 88%

**Answer**

Correct:

The histogram displays information about 3 + 9 + 6 + 3 + 2 + 1 + 1 = 25 students. Out of these 25 students, 6 + 3 + 2 + 1 + 1 = 13 studied 6 or more hours for the exam, so 13/25 = 0.52 * 100 = 52% of the students. Note that it might have been easier to count the bars representing less than 6 hours. Thus, the number of students who studied 6 or more hours is 25 - (3 + 9) = 25 - 12 = 13 students.

Submit