STAT 108 Lecture 3 Tu, Sept 1, 2015

2.2 Organizing and Graphing Quantitative Data

Recall that a **quantitative** variable is measured numerically, and arithmetic operations on these numbers make sense.

Examples. Height, weight, income, age, blood pressure, cholesterol level, temperature, lifespan.

Definition. **Quantitative data** are observations of a quantitative variable.

Definition. **Ungrouped data** contain information on each observational unit individually.

Definition. **Grouped data** contain information on groups of observational units.

Examples.

- Ages of 10 patients in a hospital are (ungrouped data): 27 42 36 48 67 42 58 54 34 56
- Ages of 10 patients in a hospital fall into several groups (called **classes** or **bins**):

Age class	Frequency
20 – 29	1 patient
30 – 39	2 patients
40 – 49	3 patients
50 – 59	3 patients
60 – 69	1 patient

These are grouped data.

<u>Definition</u>. A **class** (or a **bin**) is an interval that includes all values that fall within two numbers.

<u>Definition</u>. A **frequency distribution** for quantitative data lists all the classes and corresponding frequencies.

Definition. The **midpoint** between the upper limit of the first class and the lower limit of the second class is the average between the two values.

Example. Consider two classes: 20 – 29 and 30 – 39. The midpoint is (29+30)/2=29.5.

Definition. The midpoint is called the **upper boundary** of the first class and the **lower boundary** of the second class.

Definition. The lower and the upper boundaries of a class are called **class boundaries**.

Definition. The **class width** (or **class size**) is the difference between the two boundaries of the class:

Class Width=Upper Boundary–Lower Boundary

Definition. The **class midpoint** is the average of the two limits of this class:

Class Midpoint = (Lower Limit + Upper Limit)/2

Example. In our example,

Class Limits	Class Boundaries	Class Width	Class Midpoint
20 – 29	19.5 to less than 29.5	29.5-19.5=10	(20+29)/2=24.5
30 – 39	29.5 to less than 39.5	39.5-29.5=10	(30+39)/2=34.5
40 – 49	39.5 to less than 49.5	49.5-39.5=10	(40+49)/2=44.5
50 – 59	49.5 to less than 59.5	59.5-49.5=10	(50+59)/2=54.5
60 – 69	59.5 to less than 69.5	69.5-59.5=10	(60+69)/2=64.5

Definition. A **relative frequency of a class** is the frequency of that class divided by the total number of observations.

Definition. A **relative frequency distribution** is the list of all classes and their respective relative frequencies.

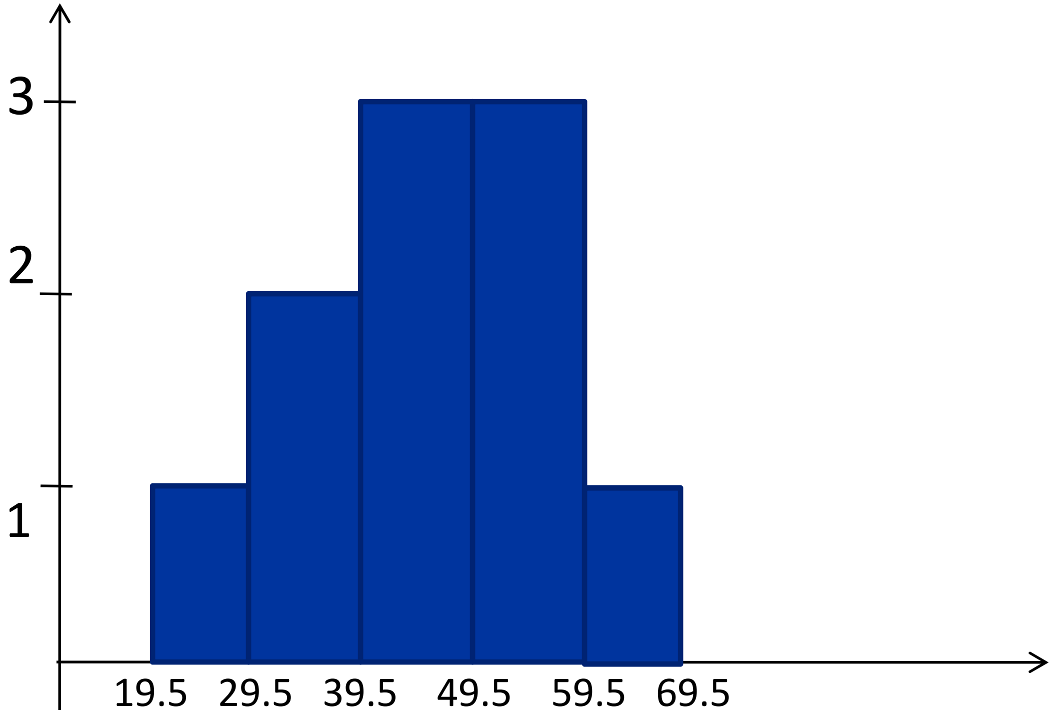
Definition. A **percentage of a class** is the relative frequency of that class multiplied by 100%.

Definition. A **percentage distribution** is the list of classes and their respective percentages.

Definition. A **histogram** is a graph in which classes are marked on the horizontal axis and the heights of vertical bars represent the frequencies (or relative frequencies, or percentages).

As a rule, the classes are chosen to be of the same size, and the bars are drawn between the class boundaries, adjacent to each other.

Example. The histogram for class frequencies for our data are

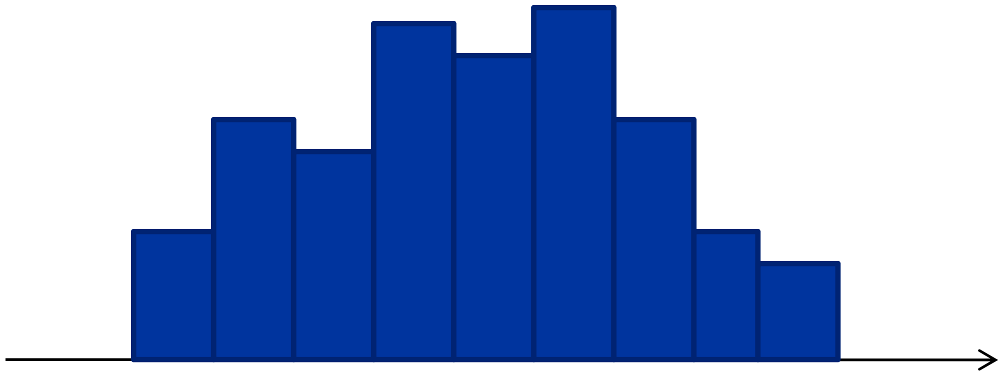


2.2 Shapes of Histograms

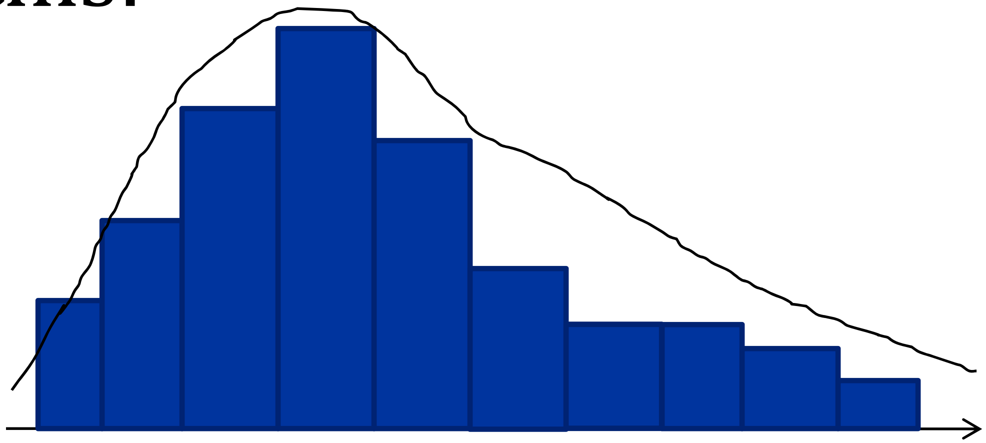
To describe the shape of a histogram, we don't look at particular bars but at overall shape. The following terms may be used to describe a distribution represented by the histogram.

<u>Definition</u>. A **symmetric** histogram is almost identical on both sides of its central point.

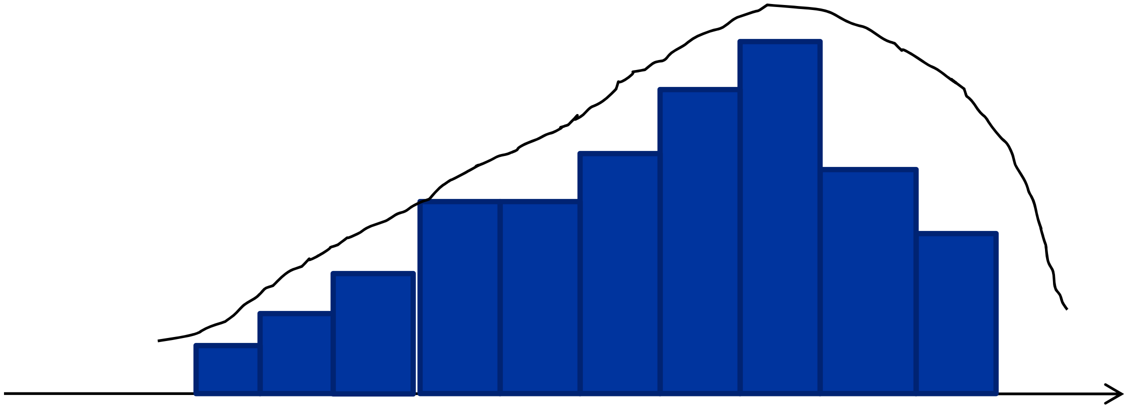
Schematically a symmetric histogram looks like this:



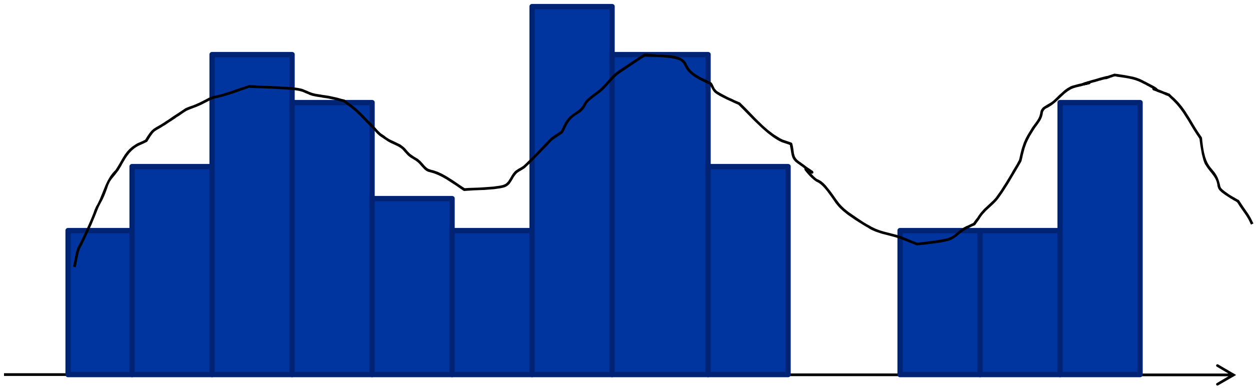
Definition. A histogram that is **skewed to the right** (or **right-skewed**) has a long right tail. Schematically it looks like this:



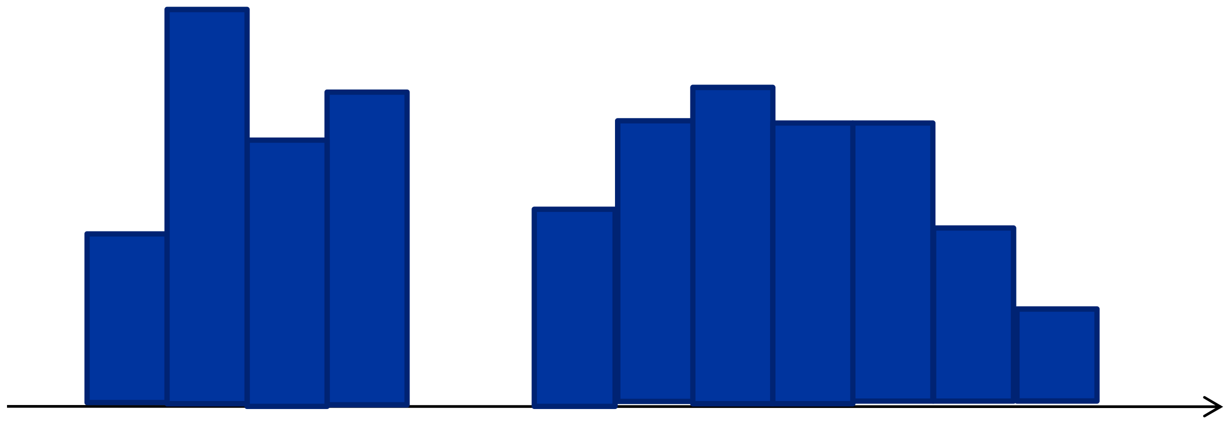
Definition. A histogram that is **skewed to the left** (or **left-skewed**) has a long left tail. Schematically this histogram looks like this:



Definition. A **unimodal** histogram has one peak. A **bimodal** histogram has two peaks. A **multimodal** histogram has several peaks. A multimodal histogram looks like this:



Exercise. Describe the histogram shown in terms of skewness and modality.



3.1 Measures of Central Tendency

Definition. A **measure of central tendency** for a data set gives the center of frequency distribution or histogram.

Three measures of central tendency are defined: **sample mean**, **median**, and **mode**.

Definition. The **sample mean** (or **average**) of observations $x_1, x_2, \ldots, x_n$ is the arithmetic average computed according to the formula:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum x}{n}$$

<u>Example</u>. Suppose 9 observations are:

1 3 1 2 5 4 15 4 1

The sample mean is equal to

$$\bar{x} = \frac{1+3+1+2+5+4+15+4+1}{9} = \frac{36}{9} = 4$$

Definition. The **median** is the middle value for data put in increasing order.

Example. Our data in increasing order are

1 1 1 2 3 4 4 5 15

The median is the middle observation, median = 3.

Note that we have an odd number of observations, so that there is always a middle observation, which is defined to be the median.

Example. Suppose now we have an even number of observations in increasing order: 1 1 1 2 3 4 4 4 5 15

There are now two middle observations (3 and 4), and we define the median as the average of the two:

$$\text{median} = (3+4)/2 = 3.5$$

Definition. The **mode** is the most frequent observation.

Example. In the set of 9 observations

1 3 1 2 5 4 15 4 1

value 1 occurs the largest number of times, so it is the mode, mode = 1.

Example. In the set of 10 observations

1 1 1 2 3 4 4 4 5 15

values 1 and 4 occur most often, so the data are bimodal, mode = 1 and 4.

Example. In the set 3 5 6 8 12 , all observations occur equally often, so we say that there is no mode.

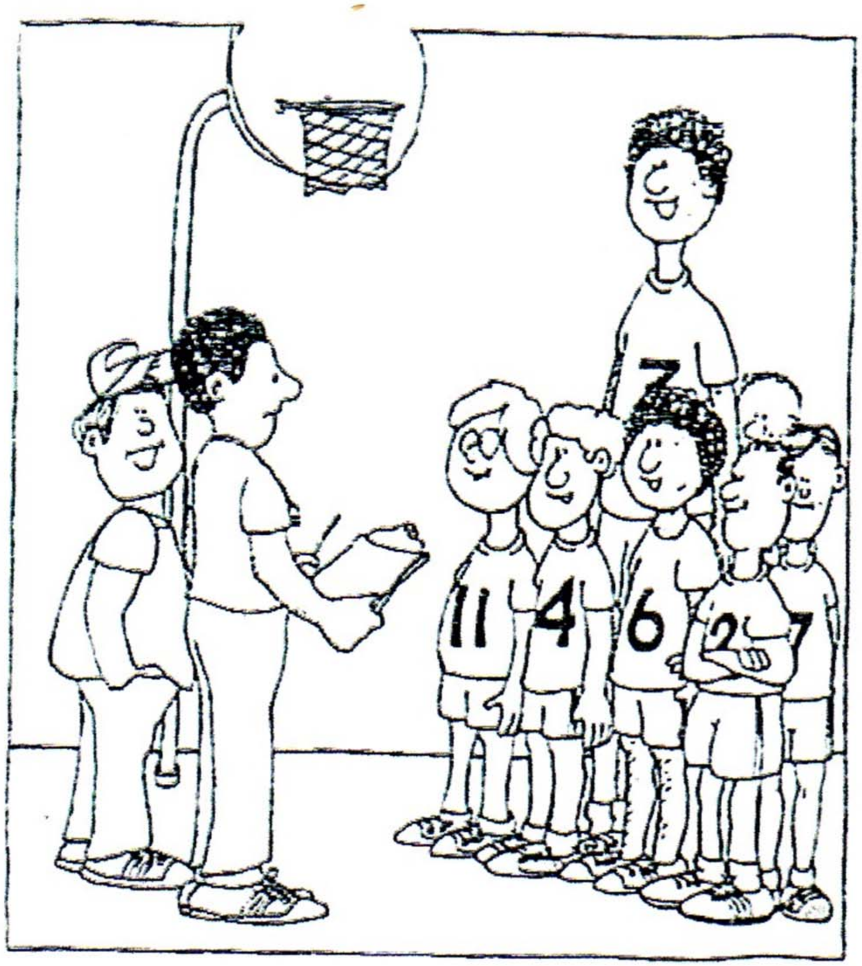
Sometimes, it may be defined as “every observation is the mode”.

Example. Consider two data sets:

1 2 3 4 4 10 (mean=4, median=3.5, mode=4) and

1 2 3 4 4 70 (mean=14, median=3.5, mode=4)

Conclusion: mean is heavily influenced by large observations (outliers), while median and mode do not change (are robust to outliers).



"Should we scare the opposition by announcing our mean height, or lull them by announcing our median height?"