Lec 2

1.2 Pictorial and Tabular Methods in Descriptive Statistics

Histograms

Histograms are bar graph for numerical. data

To construct a histogram, the first step is to "bin" the range of values - that is,

First decide on bins (intervals)

and then count # data pts in each bin

The bins are usually specified as consecutive nonoverlapping intervals

The bins must be adjacent and are often (but not required to be) of same length

If the bins are of equal size, a column is erected over the bin with height

proportional to the frequency - the number of cases in each bin.

Consider data consisting of observations on a discrete variable x. The **frequency** of any

particular x value is # times data was in bin.

The **relative frequency** of a value is the proportion of time value occurs

$$\text{relative frequency of a value} = \frac{\#\text{ times data was in bin}}{\#\text{ observations} \leftarrow \text{sample size}}$$

Relative Frequency Histograms

A histogram may also be normalized to display relative frequency.

It then shows the proportion of cases that fall into each of several categories, with

~~the~~ sum of heights = 1 So relative frequency histograms are

bar charts of the relative frequency.

Example 1. A website gives information on 50 charities. Here is a sample of 20 charities and the amount (in thousands) they spend on fundraisers.

20, 10, 5, 1, 2, 19, 18, 2, 6, 29, 35, 11, 23, 13, 31, 32, 35, 25, 26, 22

Find the histogram and relative frequency histogram

1, 2, 2, 5, 6, 10, 11, 13, 18, 19, 20

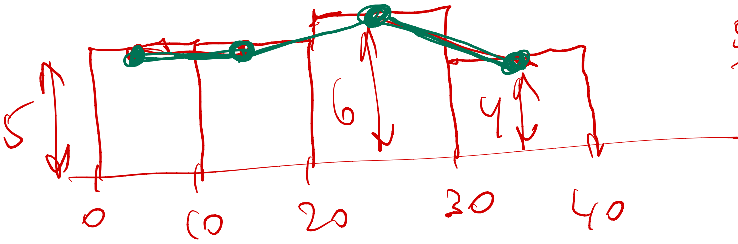
22, 23, 25, 26, 29, 31, 32, 35, 35

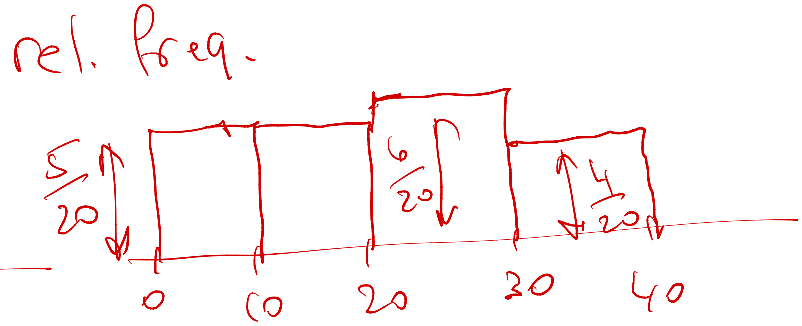
[0,10)

[10, 20)

[20, 30)

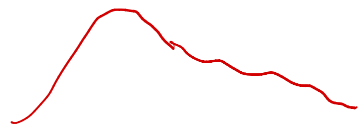
[30, 40)



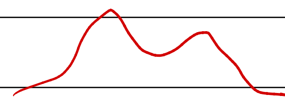


Describing Histogram Shapes

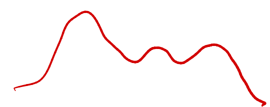
A **unimodal** histogram is one that ______________



A **bimodal** histogram ______________

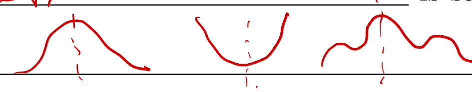


A histogram w/ multiple peaks is said to be **multimodal**.

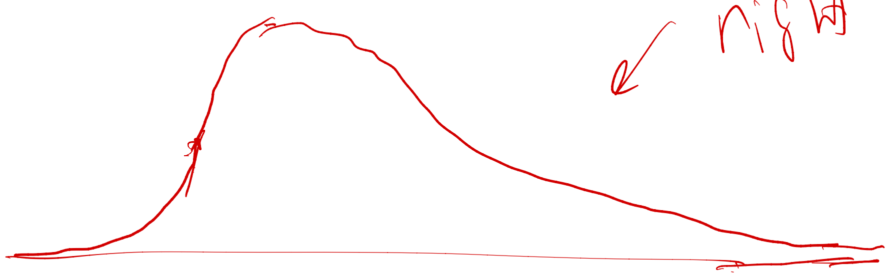


And a histogram with no peaks is said to be **uniform**.

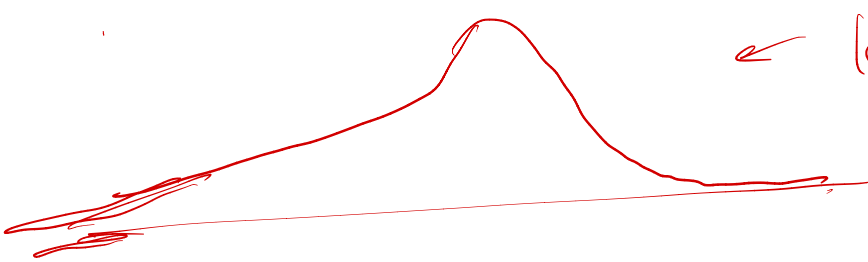
A histogram is **symmetric** if ______________



A unimodal histogram is Positively skewed (right skewed) if the right or upper tail is stretched out compared with the left or lower tail, and negatively skewed or left skewed if the stretching is to the left.



right skewed



left skewed

Example 2. Let's take a look at the problem on HW1

1. [7.5 points] A small survey was conducted in which each respondent was asked how many times, in the previous two-week period, they had eaten at a fast food restaurant. The data appear below.

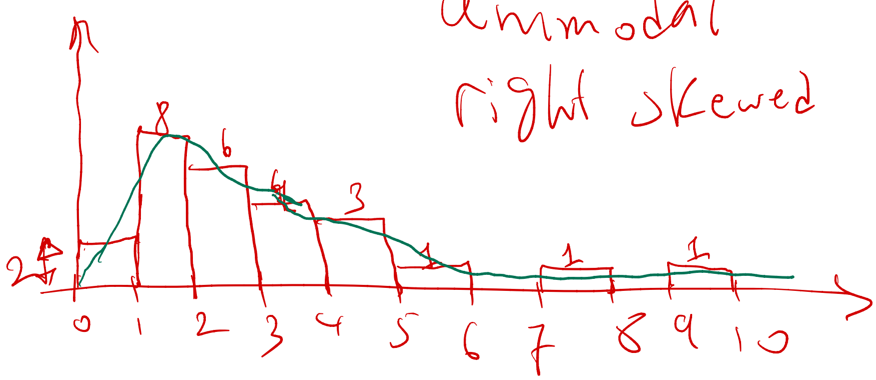
0, 2, 1, 5, 2, 2, 3, 4, 1, 2, 7, 1, 3, 4, 1, 0, 1, 4, 2, 1, 3, 3, 2, 1, 9, 1

(a) Construct a frequency histogram. The histogram should be neat, accurate, and well-labeled. [3.5 points]
(b) How would you describe the shape of the distribution? [1 point]

bins of size 1

unimodal
right skewed

rel. freq.
$8 \leftrightarrow \frac{8}{26}$



1.3 Measures of Location

Suppose, that our data set is of the form $x_1, x_2, \cdots, x_n$, where each x_i is a number. One important characteristic of such a set of numbers is its location, and in particular its center.

1.3.1 Sample Mean

For a given set of numbers $x_1, x_2, \cdots, x_n$, the most familiar and useful measure of the center is the mean, or arithmetic average of the set.

The **sample mean** $\bar{x}$ of observations $x_1, x_2, \cdots, x_n$ is given by

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

A physical interpretation of the sample mean demonstrates how it assesses the center of a sample. Think of each dot in the dotplot below representing a 1-lb weight. Then a fulcrum placed with its tip on the horizontal axis will balance precisely when it is located at $\bar{x}$. So the sample mean can be regarded as the balance pt of the distribution of observations.

EXAMPLE 1.14 Here are the 24-hour water-absorption percentages for the specimens:

$x_1 = 16.0$	$x_2 = 30.5$	$x_3 = 17.7$	$x_4 = 17.5$	$x_5 = 14.1$
$x_6 = 10.0$	$x_7 = 15.6$	$x_8 = 15.0$	$x_9 = 19.1$	$x_{10} = 17.9$
$x_{11} = 18.9$	$x_{12} = 18.5$	$x_{13} = 12.2$	$x_{14} = 6.0$	

With $\Sigma x_i = 229.0$, the sample mean is

$$\bar{x} = \frac{229.0}{14} = 16.36$$

Dotplot of the data from Example 1.14

1.3.2 Sample Median

The **sample median** is the middle value

The sample median $\tilde{x}$ is obtained by ______.

1) order the data

Then, 2) n odd: median is the $\frac{n+1}{2}$ th pt

n even: median is the average of the $\frac{n}{2}$ & $\frac{n}{2}+1$ th pts

Example 3.

(1) Suppose we have a data set as: 1.6, 3.0, 1.9, 0.6, 3.8. What are the mean and median of this sample?

(a)

$\bar{x} = 2.18$

(b)

0.6, 1.6, 1.9, 3, 3.8

$\tilde{x} = 1.9$

(2) Suppose we have a data set as: 6.9, 16.3, 32.8, 41, 47.7, 48.9. What is the median of this sample?

$$\tilde{x} = \frac{32.8+41}{2} = 36.9$$