1 Variables and Data

Definitions and introduction to terminology used in statistics.

Definition: Data are numbers with a context.
A data set includes *cases*. For each case the data set records information in form of *variables*.

Cases in a particular data set usually possesses many different characteristics (variables):

Example:
Student is case Variables: age, height, weight, sex, hair color, grade in STAT 161.

Definition:

- *Cases* (experimental units) are the objects about which information is desired. Cases may be people, animals, bacteria, companies, bicycles, cars, stocks, etc..
- A *variable* is any characteristic of an individual. A variable can take different values for different cases.
- The entire collection of cases is called the population of interest.
- A sample is a subset of the population, selected for study in some prescribed manner.

1.1 Two types of Statistics

1. In *descriptive statistics* the purpose is to describe a data set with the help of numerical and graphical methods. Usually the interest is just in the sample presented
2. In *inferential statistics* the goal is to draw conclusions about an entire population based on the findings in a sample. We use the sample to estimate, predict and make decisions about the entire population

Example 1

1. The proportion of students at MacEwan who attended high school in the Capital Area. The OUR (Office of the University Registrar) can provide this information. This describes a purely descriptive statistical application.
2. Does attendance matter in student success. In a study a random sample of students enrolled in college courses the attendance rate and the midterm and final exam results are recorded. Then this data is used to test if attendance has a significant influence on the marks obtained in an exam for students enrolled in college courses.

   This describes the setup of a study employing inferential statistics. Data on a sample of cases is used to draw conclusion about all students enrolled in college courses.

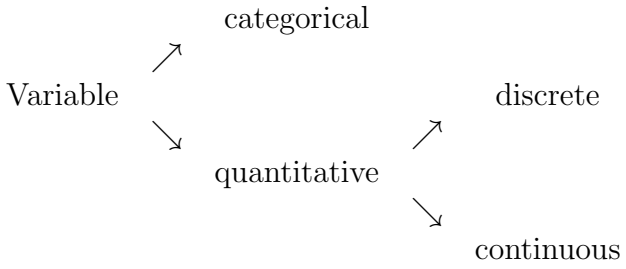
   Warning: This is a very simplistic view, success in a course depends on many other variables (subject, year of study, university, time class is scheduled, etc.)

Definition:

- A *population* is the set of all cases of interest in a study.
- A *sample* is a subset of a population, of which information (data) is obtained.

For inferential statics random samples are required!

1.2 Types of Variables



Example:

1. qualitative: sex, hair color, species, scales (very good ..., very bad), artist
2. quantitative discrete: number of siblings, number of people crossing an intersection (often result of counting), number of band members
3. quantitative continuous: height, weight, distance, length of time (often result of measurement)

Definition:

1. A *categorical* variable places an individual into one of several groups or categories.
2. A *quantitative* variable measures a numerical quantity or amount in each individual.
3. The *distribution* of a variable tells us what values it takes and how often it takes these values.

Definition:

1. A (quantitative) *discrete variable* can assume only a finite or countable number of values.
2. A (quantitative) *continuous variable* can assume the infinitely many values corresponding to the points in an interval.

Quantitative data is discrete if the possible values are isolated points on the number line.

Numerical data is continuous if the set of possible values forms an entire interval on the number line.

Data:
Values of variables for different individuals.

2 Displaying Distributions with Graphs

(Exploratory data analysis)

1. Examine each variable by itself
2. Study the relationships among the variables.

Example 2

1. number of apartments in a building
2. average square meters per apartment
3. gym (yes/no)
4. laundry (yes/no)
5. how many one/two bedroom apartments
6. most widely used flooring products

Different type of graphs for different type of data

2.1 Qualitative data

After qualitative data has been sampled it should be summarized to provide the distribution of the variable, which means report:

1. What values have been observed? (red, green, blue, brown, orange, yellow)
2. How often did every value occur?

The distribution of a qualitative variable is given in form of a table giving the following information:

- each possible category
- count (frequency), or number of individuals who fall into each category
- relative frequency (proportion) or percent of individuals who fall into each category
- percentage of measurements in each category

Definition: The *relative frequency* for a particular category is the fraction or proportion of the frequency that the category appears in the the data set. It is calculated as

$$\text{relative frequency} = \frac{\text{frequency}}{\text{frequence of observations}}$$

$$\text{percent} = 100 \times \text{relative frequency}$$

Example:
sample size=number of observations=n=200

category	frequency	relative frequency	percentage
wood	50	0.25	25%
tiles	10	0.05	5%
linoleum	20	0.1	10%
carpet	120	0.6	60%
total	21	1	100%

Such a table is called the frequency distribution for qualitative data or statistical table.
Once the data is summarized in a frequency distribution table, the data can be displayed in a bar chart or pie chart. The bar chart will effectively show the frequencies or percent in the different categories whereas the pie chart will show the relationship between the parts and the whole.

Example: Why do youth vote? In the National Youth Survey participants were asked why they vote: The following table summarizes the results.
General Attitude towards Politics collapses the following answers:
It is a civic duty to vote
Because I think it is important to vote
It allows me to express my opinions/views
I can/It's my right
Out of habit
It's important that youth vote My vote counts

Political Influencers:
Because a friend, family member, or other person encouraged me to vote

Interest in Politics:
To support or oppose a political party
I want to/I want change
To support or oppose a specific candidate
I care about different issues
sample size=number of observations=n=1014

category	frequency	relative frequency	percentage	angle
Attitude	707	0.70	70%	252
Influencers	34	0.03	3%	11
Interest	266	0.26	26%	94
Other	7	0.01	<1%	3
total	1014	1	100%	360

2.1.1 Bar chart

A bar chart is a graph of the frequency distribution of a qualitative data variable

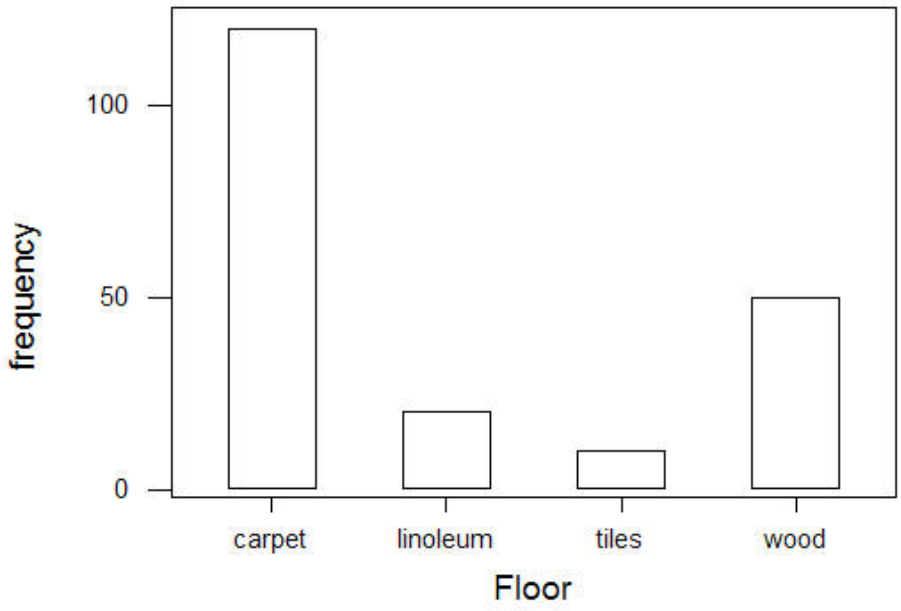
- For every category the x-axis is marked with a tick.
- Each category is represented by a bar, which hight is proportional to the corresponding frequency (relative frequency).

- label the y-axis.

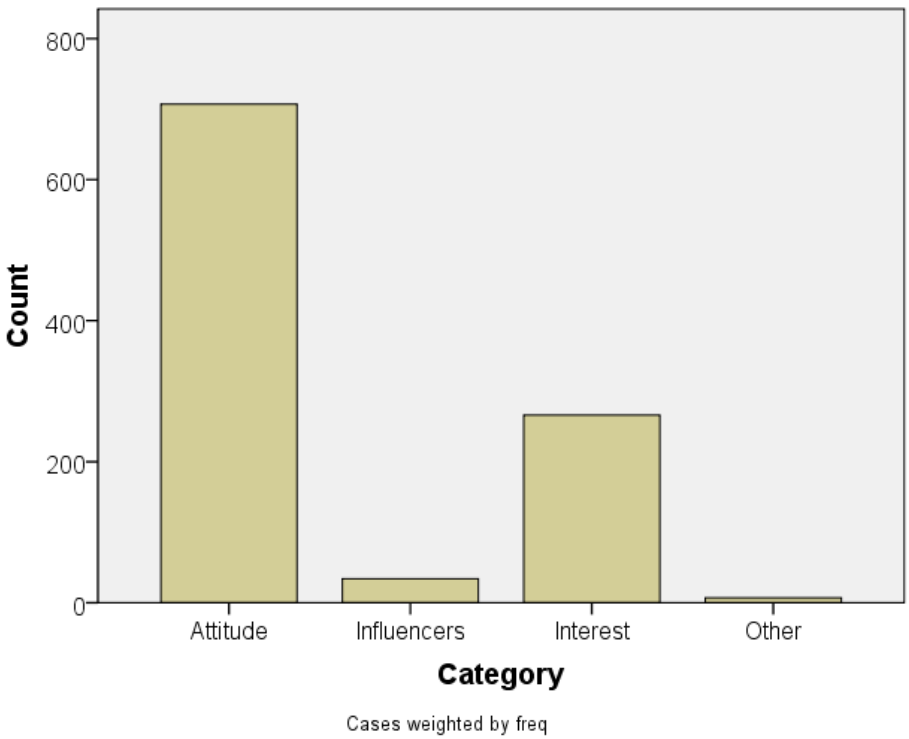
Remark: The width of each bar should be the same, so the height is proportional to the corresponding frequency (or proportion, or percentage).

Example 3
Suppose the frequency distribution of the mainly used flooring products is:

	frequency	relative freq
wood	50	0.25
tiles	10	0.05
linoleum	20	0.1
carpet	120	0.6



Example 4
For the Question from the National youth Survey the bargraph looks like



2.1.2 Pie charts

Pie charts provide an alternative kind of graph for qualitative data:
In a pie chart a circle is used to represent the sample. The size of the slice representing a particular category is proportional to the corresponding frequency (relative frequency).

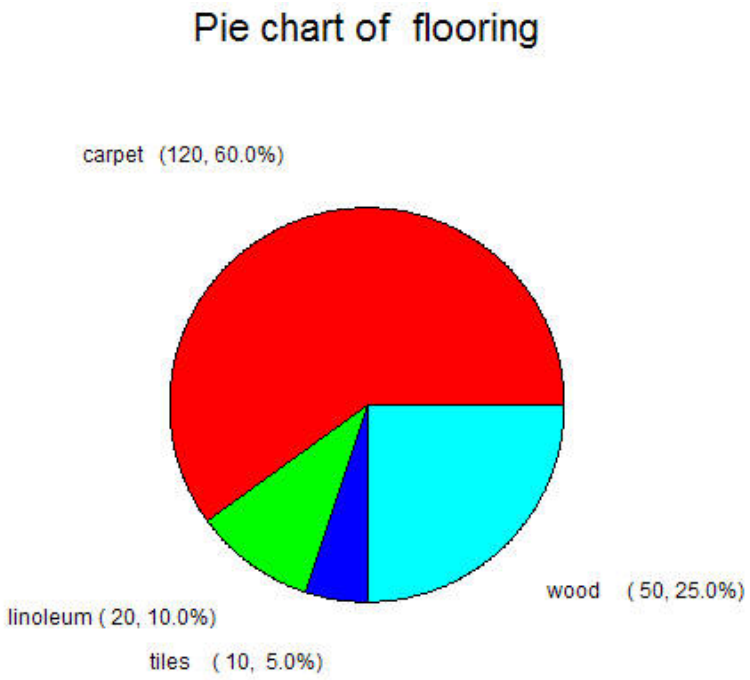
How to create a pie chart:

- Draw a circle
- Calculate the slice size (angle)

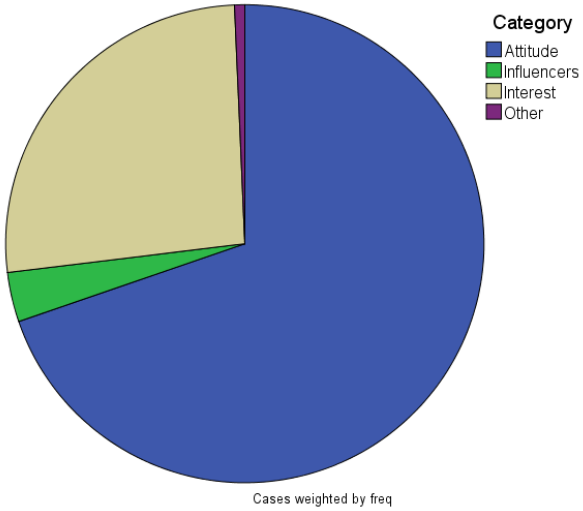
  $$\text{slice size}=\text{category relative frequency} \cdot 360$$

  (fraction of the circle for the category)
- use protractor to mark the angles

	frequency	relative freq	angle
wood	50	0.25	90
tiles	10	0.05	18
linoleum	20	0.1	36
carpet	120	0.6	216



Example 5
For the Question from the National Youth Survey the piechart looks like



M&M's example:
On the M&M's webpage the following information on the distribution of colors in peanut M&M's is provided

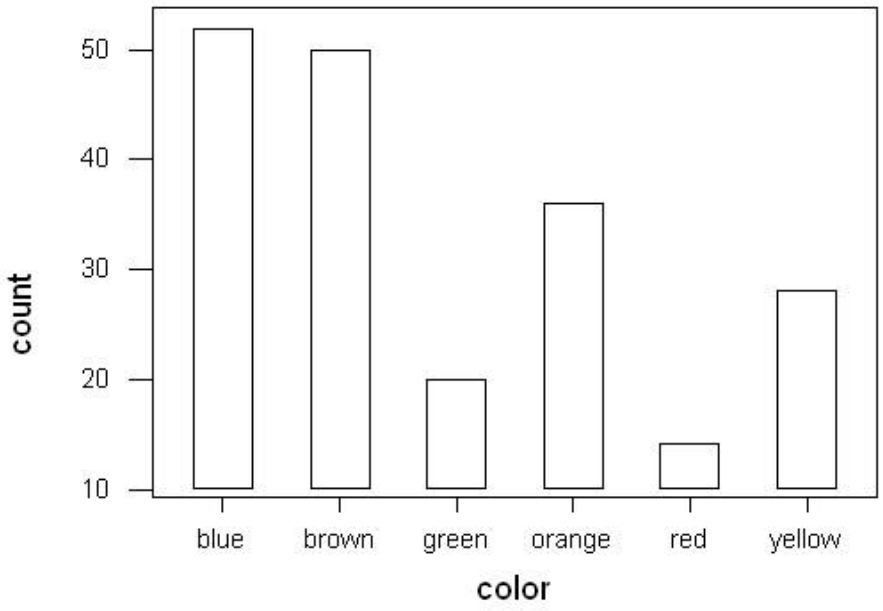
color	brown	yellow	red	blue	orange	green
percent	12%	15%	12%	23%	23%	15

In order if this distribution is a "true" description of what is in a bag, someone bought a bag with 200 peanut M&M's and wants to describe the colors of the contents.

Color is a qualitative variable, so a relative frequency table shall be obtained.

color	count	rel. freq.	percentage
brown	50	0.25	25%
yellow	28	0.14	14%
red	14	0.07	7 %
blue	52	0.26	26%
orange	36	0.18	18%
green	20	0.1	10%
Total	200	1	100%

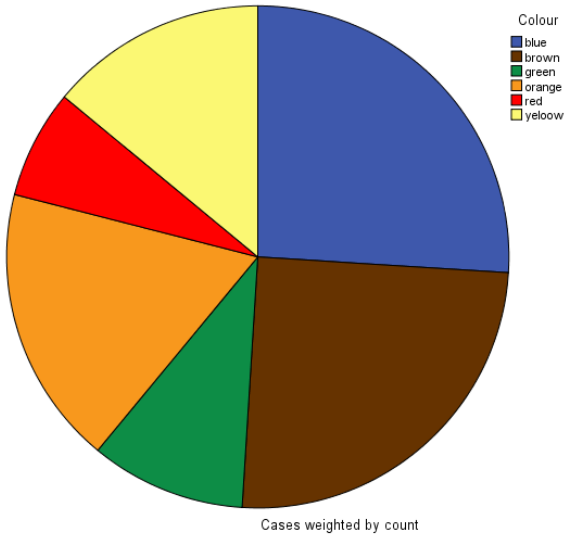
And a bar chart would look like this:



For the pie chart the angles of the slices have to be determined

color	count	rel. freq.	angle
brown	50	0.25	90^o
yellow	28	0.14	50.4^o
red	14	0.07	25.2^o
blue	52	0.26	93.6^o
orange	36	0.18	64.8^o
green	20	0.1	36^o
Total	200	1	360^o

This results in the following pie chart



2.2 Graphs for Quantitative Data

Graphs from this section display the data for a quantitative variable in a fashion so that the distribution of the data becomes apparent. Remember: The distribution shows **which** values were observed and **how often** they occurred.

2.2.1 Stemplots

One way of displaying quantitative data is in a stem and leaf plot.
Each observed number is broken into two pieces called the *stem* and the *leaf*.
How to do a stemplot:

1. Divide each measurement into two parts:
   - The first digit(s) of the number are the stems.
   - The last digit(s) of the number are the leaves.
2. List the stems in a column, with a vertical line to their right.
3. For each measurement, record the leaf portion in the same row as its corresponding stem.
4. Order the leaves from lowest to highest in each stem.
5. Provide a key to your stem and leaf coding so that the reader can recreate the actual measurements.

Example 6
Acceptance rates at a random set of business schools:

16.3, 12.0, 25.1, 20.3, 31.9, 20.7, 30.1, 19.5, 36.2, 46.9, 25.8, 36.7, 33.8, 24.2, 21.5, 35.1, 37.6, 23.9, 17.0, 38.4, 31.2, 43.8, 28.9, 31.4, 48.9

Stemplot:

1	20 63 70 95
2	03 07 15 39 42 51 58 89
3	01 14 19 38 51 62 67 76 84
4	38 69 89

It shows: center, range, concentration, nature of distribution (unimodal, bimodal, multimodal), unusual values, skewed to the right/left.
Sometimes the available stem choices result in a plot that contains too few stems and a large number of leaves within each stem. In this situation you can *stretch* the stems by dividing each into several lines.

The two common choices for dividing stems are:

- Into two lines, with leaves starting with 0 to 4 and 5 to 9
- into 5 lines, with leaves 0-1, 2-3, 4-5, 6-7, 8-9

Example 7

1	20
1	63 70 95
2	03 07 15 39 42
2	51 58
3	01 14 19 38
3	51 62 67 76 84
4	38
4	69 89

You also can use stem and leaf plots for the comparison of the distribution of two groups:

Example 8

Acceptance rates of law schools: 32.9, 34.2, 33.6, 34.0, 37.3, 34.0, 31.5, 31.9, 32.4, 31.1, 29.7, 31.6, 33.0, 30.5, 29.2, 29.6, 31.6

Back-to-back or comparative stem and leaf plot:

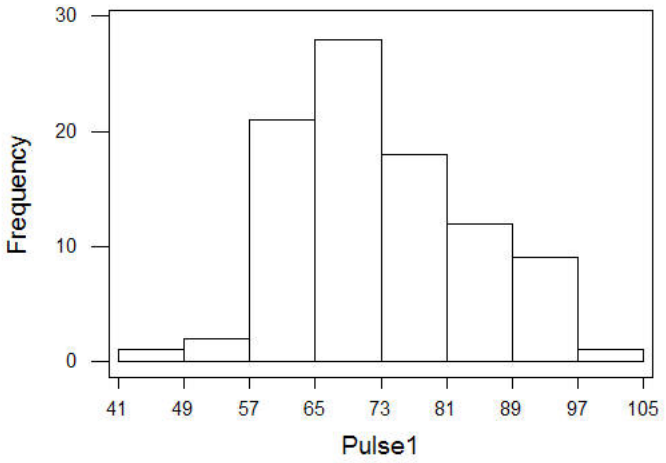
```
---------------------
  1 | 2: represents 12, leaf unit: 1
BusSchools   LawSchools
 ---------------------
         | 1 |
        2| 1 |
         | 1 |
       76| 1 |
        9| 1 |
```

```
      100| 2 |
        3| 2 |
      554| 2 |
         | 2 |
        8| 2 |999
     1110| 3 |011111
        3| 3 |2233
        5| 3 |444
      766| 3 |7
        8| 3 |
         | 4 |
        3| 4 |
         | 4 |
        6| 4 |
        8| 4 |
         | 5 |
---------------------
n:     25      17
---------------------
```

Interpretation: Centres of the distributions seem to be similar, but much more spread in the acceptance rates for Business Schools than for Law Schools.

2.2.2 Histograms

The most common graph for describing numerical continuous data is the histogram.
It visualizes the distribution of the underlying variable very well.
How a histogram looks like:



Definition:

A relative frequency histogram for a quantitative variable is a bar graph in which the hight of the bar shows "how often" (measured as a relative frequency) measurements fall in a particular interval. The classes or intervals are plotted along the horizontal axis.

Grouping
Before constructing a histogram the numerical data has to be grouped and the frequency distribution needs to be found.

1. Decide which *classes* (intervals of equal length) to use for the histogram.

   The number of *classes* used should be approximately the square root of the sample size.

   Use sensible cutpoints: The intervals should have the same width and the cutpoints should be, if possible whole numbers or tenth. The lower cutpoint of a class is included with the class, but the upper cutpoint is included with the next higher class.

2. Obtain the frequency (and relative frequency) distribution table for the classes.

   Count how many values fall within each class, giving the frequency for this class, and then find the relative frequencies by dividing the frequencies by the sample size.

With the information from the distribution table the histogram can now be constructed

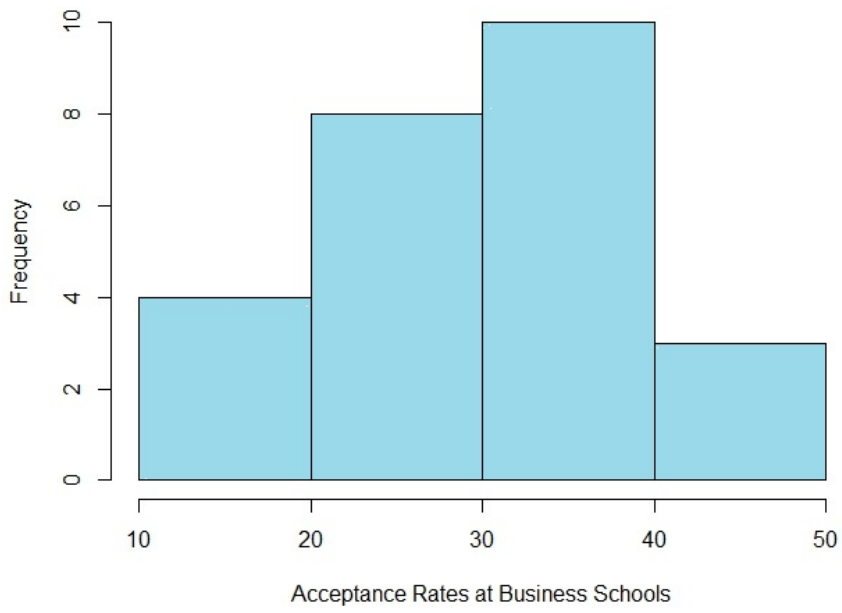
1. Mark the boundaries of the class intervals on a horizontal axis.
2. Use the frequency (or relative frequency) on the vertical axis.
3. Draw a bar for each class, with heights according to the frequency (or relative frequency) of the corresponding interval.

Histogram for Acceptance Rates:

1. The sample size is 25, the square root is 5, but the interval can easily be cut into 4 intervals, so will go with intervals. The range is 50 - 10 = 40. 40/4=10 $[10, 20)$, $[20, 30)$, $[30, 40)$, $[40, 50)$
2.

class intervals	frequency	relative frequency
$[10, 20)$	4	0.16
$[20, 30)$	8	0.32
$[30, 40)$	10	0.4
$[40, 50)$	3	0.12

3.



It shows: center, range, concentration, nature of distribution (unimodal, bimodal, multimodal), unusual values, skewed to the right/left.

Histogram shapes

1. number of peaks: unimodal, bimodal (often occurs if you have observation from two groups (men, women), multimodal
2. symmetry: if you can draw a vertical line so that the part to the left is a mirror image of the part to the right, then it is symmetric.
3. nonsymmetric graphs are skewed. If the upper tail of the histogram stretches out farther than the lower tail, then is the histogram positively skewed, or skewed to the right.

   Is the lower tail longer than the upper tail the histogram is negatively skewed.

Check a distribution for shape, center, and spread and look for deviations from the overall pattern.