Picturing Distributions with Graphs

Diana Mindrila, Ph.D.
Phoebe Balentyne, M.Ed.

Based on Chapter 1 of The Basic Practice of Statistics (6th ed.)

Concepts:

- Categorizing Variables
- Describing the Distribution of a Variable
- Constructing and Interpreting Graphs and Plots

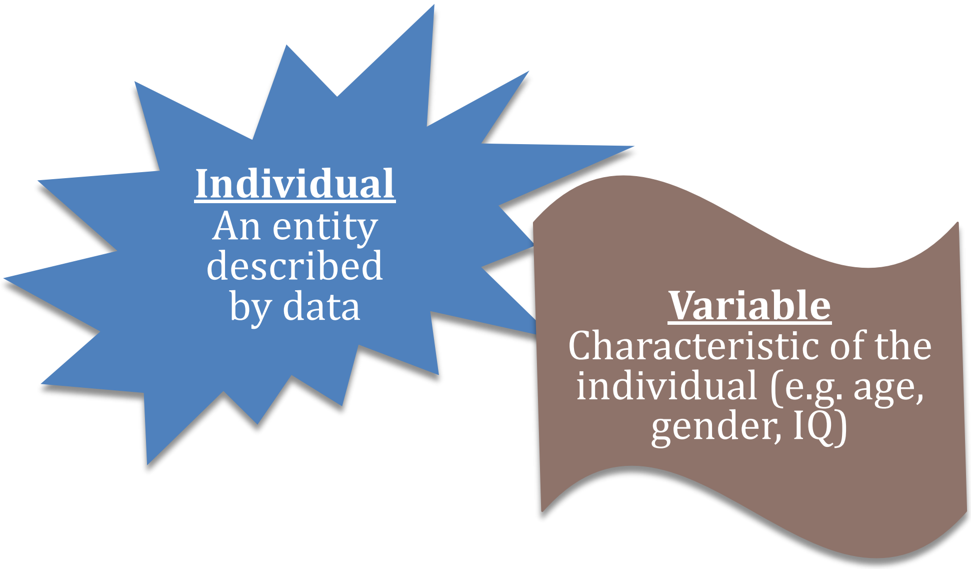
Objectives:

- Define individuals and variables.
- Categorize variables as categorical or quantitative.
- Describe the distribution of a variable.
- Construct and interpret pie charts and bar graphs.
- Construct and interpret histograms and stemplots.
- Construct and interpret time plots.

References:
Moore, D. S., Notz, W. I, & Flinger, M. A. (2013). *The basic practice of statistics* (6th ed.). New York, NY: W. H. Freeman and Company.

Statistics

- **<u>Statistics</u>** is the science of data.
- Each data set includes a collection of information (variables) about a group of individuals.



<u>Individual</u>
An entity described by data

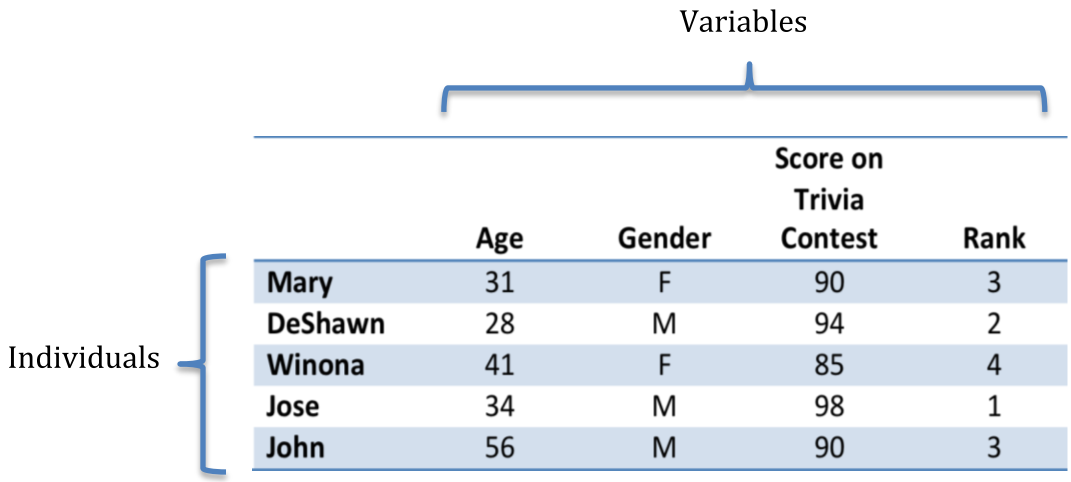
<u>Variable</u>
Characteristic of the individual (e.g. age, gender, IQ)

- In any research project, after selecting the sample and choosing a data collection method, the data collection process begins.
- Researchers obtain information about the group of individuals in the sample. This information is called the **<u>data set</u>**.
- Each data set includes a set of individuals, along with the information collected about each individual.
- Information can be collected about a variety of entities (humans, animals, objects, etc.).
- In the social sciences, information is most often collected about human beings, so they are referred to as individuals.
- Statisticians and researchers also use the term **<u>observations</u>**. Each entity, along with the information collected about it, is considered an "observation."
- Each piece of information that is collected is called a **<u>variable</u>** (examples: age, height, weight, score on a test, etc.).
- They are called variables because although the same *type* of information is collected about each individual, the values recorded will most likely *vary* from one individual to another.

Data Sets

- Most data sets list individuals as rows and variables as columns.

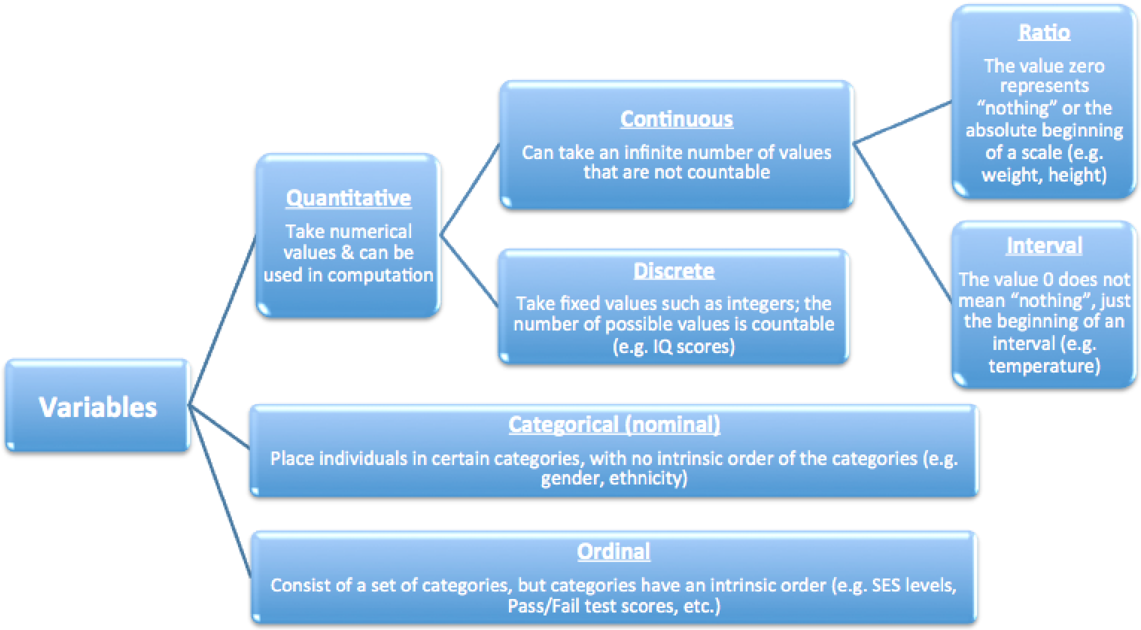
- In the following example, data was collected about a group of participants in a trivia contest:


Variables
Individuals

	Age	Gender	Score on Trivia Contest	Rank
Mary	31	F	90	3
DeShawn	28	M	94	2
Winona	41	F	85	4
Jose	34	M	98	1
John	56	M	90	3

- The name of each individual is listed in a separate row.
- The variable (or the information collected about each individual) is listed in a separate column.
- In this example, age, gender, score on trivia contest, and rank earned based on this score was recorded for each individual.

Variable Types


Variables
Quantitative
Take numerical values & can be used in computation
Continuous
Can take an infinite number of values that are not countable
Discrete
Take fixed values such as integers; the number of possible values is countable (e.g. IQ scores)
Ratio
The value zero represents "nothing" or the absolute beginning of a scale (e.g. weight, height)
Interval
The value 0 does not mean "nothing", just the beginning of an interval (e.g. temperature)
Categorical (nominal)
Place individuals in certain categories, with no intrinsic order of the categories (e.g. gender, ethnicity)
Ordinal
Consist of a set of categories, but categories have an intrinsic order (e.g. SES levels, Pass/Fail test scores, etc.)

- **Quantitative variables** take numerical values and can be used for computations (i.e. test scores). There are multiple types of quantitative variables:
  - **Discrete variables** can only take specific values or rounded values like integers. A discrete variable is a quantitative variable that has a finite number of possible values or a countable number of values.
    - Example: IQ Scores – they range in value, but can only be integers
  - **Continuous variables** are quantitative variables that have an infinite number of possible values between integers (ex: weight, height, speed, etc.).
    - Continuous variables can have an infinite number of decimals.
    - Continuous variables can be divided further into two categories:
      - **Ratio variables** – the value zero represents nothing or the absence of an entity (ex: height or weight – zero does not exist)
      - **Interval variables** – zero represents a point on the scale (ex: temperature – a temperature of zero does exist)

- **Categorical variables**, also called nominal variables, have a certain number of categories, but the categories cannot be ranked in any way.
    - Examples include gender or names of individuals.
    - Categorical variables do not have numerical values.
    - However, when data is entered into a computer the categories often receive a numerical code for practical reasons. For example, males may be denoted as "0" and females may be denoted as "1." This number has no meaning, but giving the category a numerical value enables the researcher to use statistical software to perform descriptive or statistical analyses.

- **Ordinal variables** are similar to categorical variables because they also have a certain number of categories, but they are different in that the categories can be ranked. They have an intrinsic order.
    - One example would be a test with two outcomes, pass or fail. Pass is superior to fail, so even though they are only two categories, they can be ranked.
    - Another example would be survey responses on a Likert-type scale. If there were four categories (Strongly Agree, Agree, Disagree, Strongly Disagree) theses could be ordered or numbers 1 to 4 based on respondents' level of agreement.

Exploratory Data Analysis

An **exploratory data analysis** is the process of using statistical tools and ideas to examine data in order to describe their main features.

Exploring Data

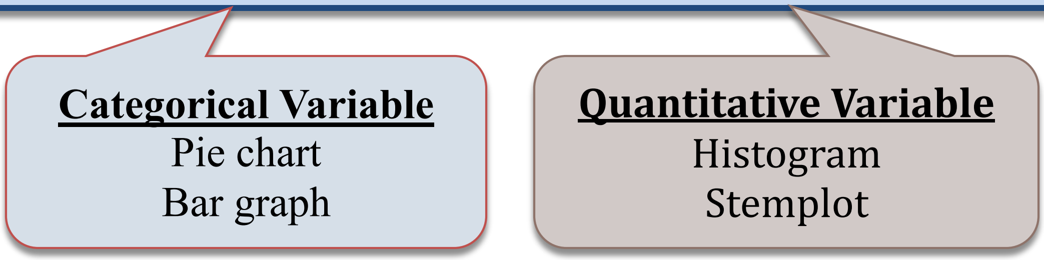
- Begin by examining each variable by itself. Then, move on to study the relationships among the variables.
- Begin with a graph or graphs. Then, add numerical summaries of specific aspects of the data.

- After collecting the data and entering each variable in the data set, the next step is to analyze the data.
- Before conducting more complex statistical analyses, researchers typically employ descriptive analyses.
- This is an exploratory phase during which researchers try to summarize the data using measures of central tendency and graphical representations.
- These notes will focus on the graphical representations.

Distribution of a Variable

- To examine a single variable, graphically display its **distribution**.

- The distribution of a variable shows what values it takes and how often it takes these values.
- Distributions can be displayed using a variety of graphical tools. The proper choice of graph depends on the nature of the variable.



Categorical Variable
Pie chart
Bar graph

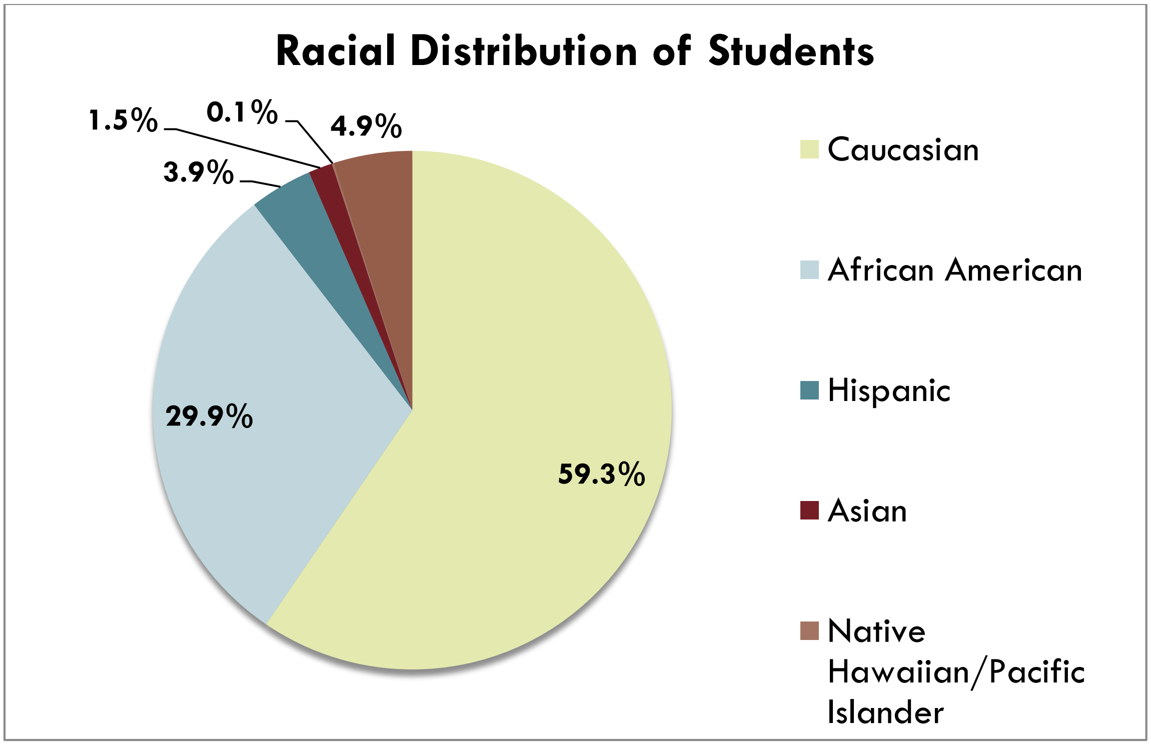
Quantitative Variable
Histogram
Stemplot

- The goal of using graphs is to illustrate the distribution of each variable.
- This means showing the value the variable takes across the individuals in the sample.
- The objective is to visually display the range of values as well as which values occur most often.
- ❖ The type of graph that should be used depends on the type of variable being described.
- Categorical variables should be displayed using pie charts or bar graphs.
- Quantitative variables are usually displayed using histograms or stemplots.
- Variables that change over time should be displayed using time plots.

Categorical Data – Pie Charts

- **<u>Pie charts</u>** show the distribution of a categorical variable as a "pie" whose slices are sized by the counts or percentage of individuals belonging to that category.

Example:

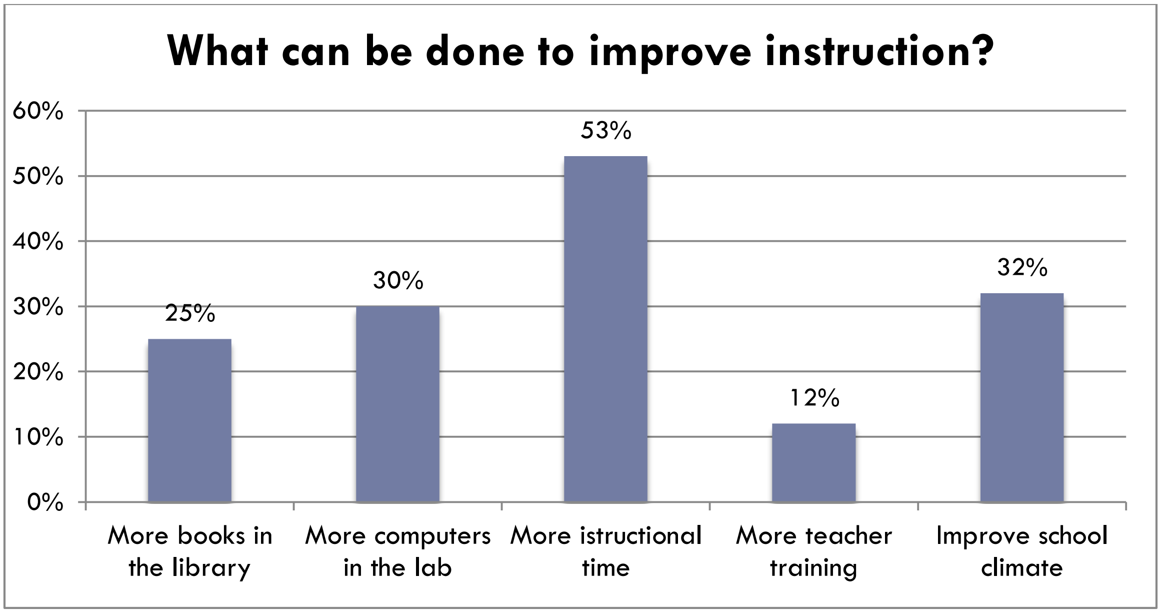


- The above example illustrates the racial distribution of a student body.
- Each slice of the pie indicates the percentage of individuals in each racial group.
- Pie charts are appropriately used <u>only</u> when all the categories put together form a whole (all the percentages add up to one hundred percent).
- Pie charts should only be used when each individual can belong to only one category.

Categorical Data – Bar Graphs

- **Bar graphs** show the distribution of a categorical variable by displaying each variable as its own bar whose height represents the number of individuals belonging to that category.

Example:



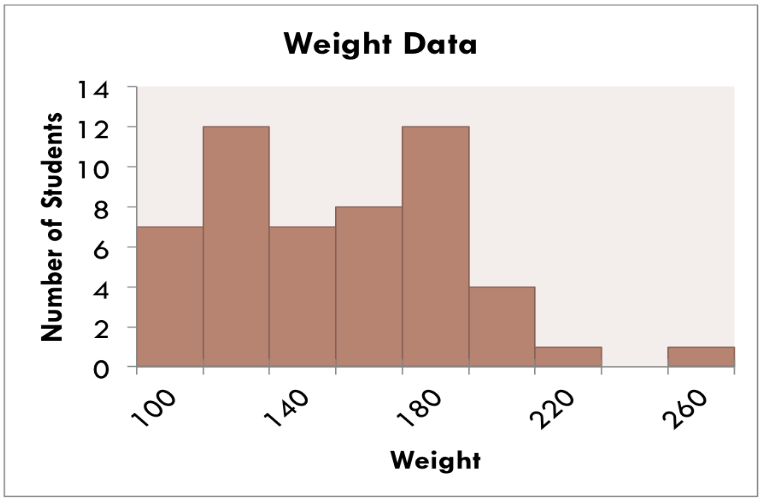
- The above example shows how teachers responded to the question, "What can be done to improve instruction?"
- This type of question allowed teachers to select all responses that applied to them, so they were able to choose more than one response.
- In this instance the percentages do not add up to 100, so a pie chart would be inappropriate.
- Unlike pie charts, bar graphs can be used when participants belong to more than one category.

Quantitative Data – Histograms

- **Histograms** show the distribution of a quantitative variable by using bars whose height represents the number of individuals whose values fall within a specific range.

Example:

Weight Group	Count
100 – <120	7
120 – <140	12
140 – <160	7
160 – <180	8
180 – <200	12
200 – <220	4
220 – <240	1
240 – <260	0
260 – <280	1



- The above example displays the weights of the individuals in the sample.
- The first step in creating a histogram is to divide the range of possible values into equal intervals.
- Statistical software will create the ranges automatically, but these intervals can be changed if desired.
- In this example, weights range from 100 lbs to 280 lbs and the intervals are set at 20 lbs.
- The first column represents students who weigh between 100 and 120 lbs (100 or more lbs, but less than 120 lbs – an individual weight of 120 lbs would belong in the next range of values).
- A histogram displays the distribution of a quantitative variable.
- It shows what values the variable takes and how often it takes those values.
- However, a histogram does not display individual values.

Quantitative Data – Stemplots

- ***Stemplots*** separate each observation into a stem and a leaf that are then plotted to display the distribution while maintaining the original values of the variable.
    - Stemplots are also considered graphical representations, but they use numbers to show how variables are distributed.
    - A stemplot looks like a histogram that is turned on end.

Example:

Distribution of Student IQ

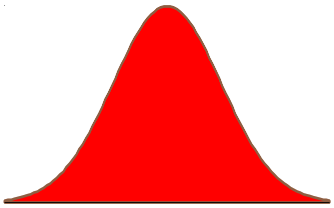
Stem	Leafs													
3	8													
4	5													
5	1	3												
6	2	2	5											
7	1	2	2	5	8									
8	2	2	3	4	8	9								
9	0	0	1	3	5	8	8	9						
10	0	0	1	1	1	2	2	3	5	5	7	8	9	9
11	0	1	1	2	3	3	4	5	5	7	8			
12	0	1	3	5										
13	2	5												
14														
15	2													

- The above example shows a distribution of student IQs.
- The first column, labeled "stem," shows the first digit or two digits of the IQ scores (the tens digit and hundreds digit).
- The second column, labeled "leaf," shows the last digit (the ones) of the IQ scores.
- For instance, the first row has a stem of 3. There is only one leaf and it is 8. This means that one of the IQ scores in the sample is 38.
- In the third row, the stem is 5 and the leaves are 1 and 3. This means there are two IQ scores in the 50s: 51 and 53.

- One benefit to using a stemplot is that every value is displayed.
- Steps to creating a stemplot:
    1) Separate each observation into a ***stem*** (first part of the number) and a ***leaf*** (the remaining part of the number).
    2) Write the stems in a vertical column; draw a vertical line to the right of the stems.
    3) Write each leaf in the row to the right of the stem it belongs with.

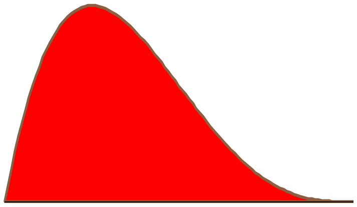
Describing Distributions

- A distribution is **<u>symmetric</u>** if the right and left sides of the graph are approximately mirror images of each other.



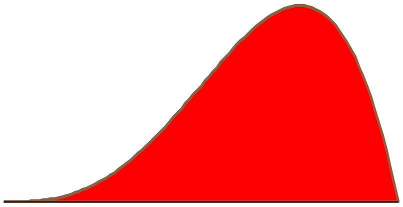
  - One example of a symmetric distribution is IQ scores. Many individuals have average IQ scores, and the frequency of individuals decreases as the scores depart from the average.

- A distribution is **<u>skewed to the right</u>** (right-skewed) if the right side of the graph (containing the half of observations with larger values) is much longer than the left side. (A few very large values skew the mean).



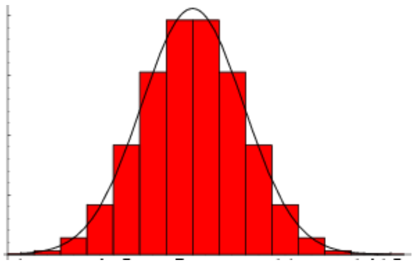
  - One example of a right-skewed distribution is income. Income levels cannot fall below a certain point, but they can be extremely large. Extremely large incomes (located to the right on the graph) occur infrequently, but skew the mean income to the right. This is why median income, rather than mean income, is usually reported. The median is more resistant to outliers.

- A distribution is **<u>skewed to the left</u>** (left-skewed) if the left side of the graph is much longer than the right. (A few very small values skew the mean).



  - One example of a left-skewed distribution is GPA. There is a maximum value of GPA, but a few individuals can have very small GPAs. These values can skew the mean to the left.

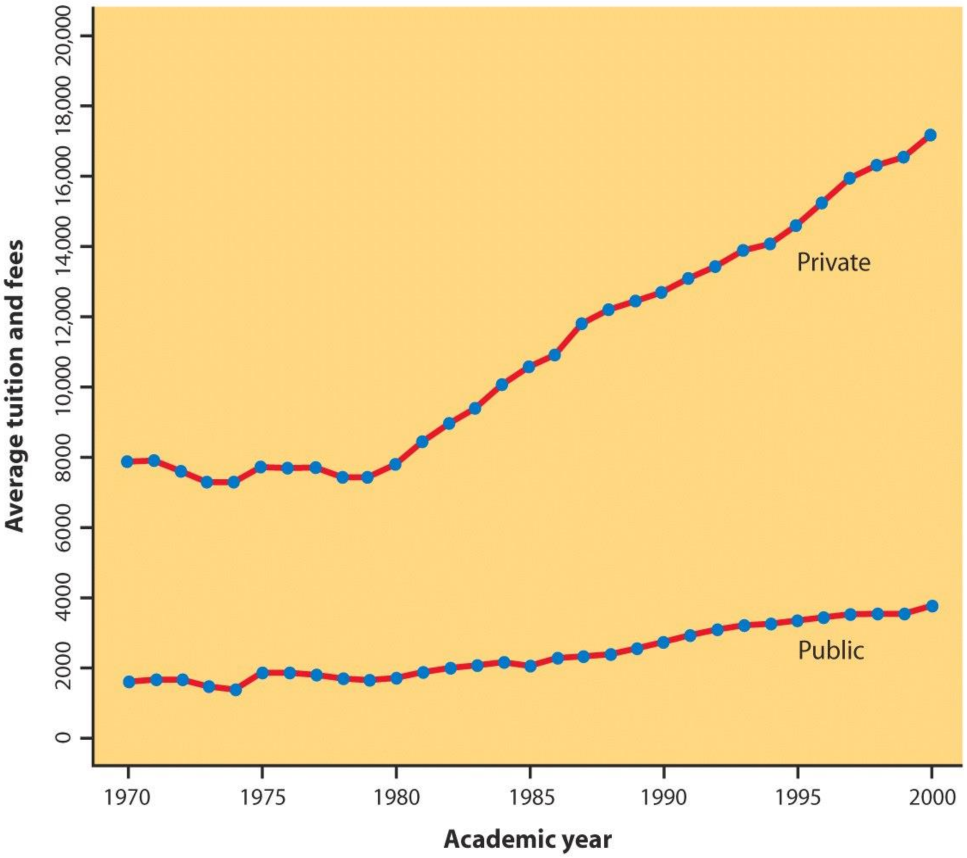
- The curves shown above are symbolic ways to show the distribution of data.
- Distributions would be shown using a histogram, which could be closely approximated by a curve, like the example shown below:



Time Plots

- A **time plot** is a graphical representation of how a certain variable changes over time.
  - Data is collected several times, not just once, and values are recorded for the same variable at different points in time.
  - Time is always on the horizontal axis, and the variable being measured is always on the vertical axis.
  - Look for an overall pattern (trend) and deviations from this pattern. Connecting the data points with lines may emphasize this trend.
  - Look for patterns that repeat at known regular intervals (seasonal variations).

Example:



- The above example displays the changes in college tuition from 1970 to 2000 for public and private universities.
- The graph shows that tuition has increased over time for both public and private schools, but the rate of increase is much higher for private schools after 1980.