

## Chapter 2: Frequency Distributions and Graphs (or making pretty tables and pretty pictures)

**Example:** Titanic passenger data is available for 1310 individuals for 14 variables, though not all variables are recorded for all individuals. Consider the following variables:

Survival, Sex, Number of relatives on board, Age



Who wants to stare at a big dataset? If you have 1310 people measured for 14 variables, how much information are we going to get by looking at the data set? See for yourself:

<http://lib.stat.cmu.edu/S/Harrell/data/xls/titanic3.xls>

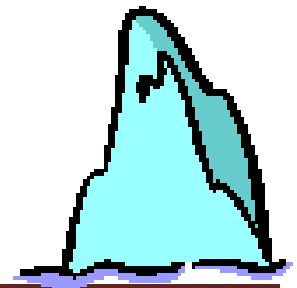
***That's where tables that summarize the data and graphs of these summaries come in handy!***

## Section 2-1 – Organizing Data

Data must be organized in a meaningful way so that we can use it effectively. This is often a pre-cursor to creating a graph.

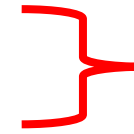
**Frequency distribution** – the organization of raw data in table form, using classes and frequencies.

- **Class** – a quantitative or qualitative category. A class may be a range of numerical values (that acts like a “category”) or an actual category.
- **Frequency** – the number of data values contained in a specific class.

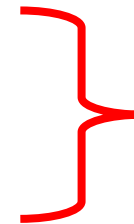


There are 3 types of frequency distributions:

- Categorical frequency distributions
- Ungrouped frequency distributions
- Grouped frequency distributions



**Qualitative  
Variables**



**Quantitative  
Variables**

Let's start with **Categorical frequency distributions** –  
frequency distribution for qualitative data.

*Review: What is qualitative data?*

**Titanic Example:** Survival status and sex are qualitative variables. The following tables give their categorical frequency distributions.

<b><i>Survival Status</i></b>	<b><i>Frequency</i></b>
Yes	500
No	809

<b><i>Sex</i></b>	<b><i>Frequency</i></b>
Female	466
Male	843

We'll come back for graphs which can include a pie graph, bar chart or Pareto chart.

**Example:** Areas of study for students in our class

<b>Area of Study</b>	<b>Frequency</b>
Medical Sciences	
Public Health	
Biology	
Education	
Geography	
Other	

For quantitative variables we have grouped and ungrouped frequency distributions. An **Ungrouped Frequency Distribution** is a frequency distribution where each class is *only one unit wide*.

- Meaningful when the data does not take on many values.
- Each class is constructed using a single data value for each class, e.g., 0, 1, 2, 3, ..., 10
- Class boundaries will be defined to separate the classes (*when graphing*) so there are no gaps in the frequency distribution.
  - **Should have one additional decimal place and end in a 5.**
  - The lower boundary will “round” to the lower class limit.
  - The upper boundary will “round” to the next class
  - *Another way of thinking about this: draw the boundary half way between consecutive classes.*

***Titanic example:*** Number of relatives on board

<b><i>Number of Relatives on Board</i></b>	<b><i>Class Boundaries</i></b>	<b><i>Frequency</i></b>
0	-0.5 – 0.5	790
1	0.5 – 1.5	235
2	1.5 – 2.5	159
3	2.5 – 3.5	43
4	3.5 – 4.5	22
5	4.5 – 5.5	25
6	5.5 – 6.5	16
7	6.6 – 7.5	8
10	9.5 – 10.5	11

***Is this an ungrouped frequency distribution?***

**Grouped frequency distribution** – frequency of a **quantitative variable** with a large range of values, so the data must be grouped into classes that are more than one unit in width.

Class Limits Age Group in Years (Lower, Upper)		Class Boundaries (Lower, Upper)		Frequency	Cumulative Frequency
0	4	-0.5	4.5	51	51
5	9	4.5	9.5	31	82
10	14	9.5	14.5	27	109
15	19	14.5	19.5	116	225
20	24	19.5	24.5	184	409
25	29	24.5	29.5	160	569
30	34	29.5	34.5	132	701
35	39	34.5	39.5	100	801
40	44	39.5	44.5	69	870
45	49	44.5	49.5	66	936
50	54	49.5	54.5	43	979
55	59	54.5	59.5	27	1006
60	64	59.5	64.5	27	1033
65	69	64.5	69.5	5	1038
70	74	69.5	74.5	6	1044
75	79	74.5	79.5	1	1045
80	84	79.5	84.5	1	1046

**Age of  
Passengers on  
the Titanic  
Classified into  
17 Age Groups  
with a class  
wide of 5 years**

## Guidelines:

- There should be between 5 and 20 classes.
- The classes must be mutually exclusive (non-overlapping).
- Makes placing observations into classes unambiguous.
- The classes must be continuous.
- There should be no gaps in the frequency distribution.
- The classes must be exhaustive.
- The classes should accommodate all the data.
- The classes must be equal in width.
- Avoids a distorted view of the data.



## TERMINOLOGY

**Class Limits** – the range of values to be included in a class.

- *Will have the same decimal place value as the data.*

**Class Boundaries** – used to separate the classes (*when graphing*) so there are no gaps in the frequency distribution.

- *Should have one additional decimal place and end in a 5.*
- The lower boundary will “round” to the lower class limit.  
The upper boundary will “round” to the next class
- *Another way of thinking about this: draw the boundary half way between consecutive classes.*

**Class width** – found by subtracting the **lower** class limit of one class from the **lower** class limit of the next class (OR likewise by subtracting an upper class boundary from the corresponding lower class boundary.)

## Steps for Constructing a Grouped Frequency Distribution:

1. Determine the classes
  - Find the **range** of the data = largest value minus the smallest value
  - Find the class width by dividing the range by the number of classes and rounding up. (Add 1 if this value is a whole number).
  - Select a starting point (usually the lowest value); add the width to get the lower limits of all subsequent classes.
  - Find the upper class limit for each class by subtracting 1 *unit* from the last decimal place of the lower class limit of the next class.
    - If the lower class limit of the next class is 12.5, then the upper class limit for the previous class is 12.4.
    - If the lower class limit of the next class is 11, then the upper class limit of the previous class is 10.
    - If the lower class limit of the next class is 18.55, then the upper class limit of the previous class is 18.54.
  - Find the boundaries.
2. Tally (count) the data in each class.
3. Find the numerical frequencies from the tallies.

## Titanic Example: Ages of passengers in first class

1 2 4 6 **11 13 14 15 16 16 16 17 17 17 17 18 18 18 18 18 18 19 19 19**  
**19 19** 21 21 21 21 21 22 22 22 22 22 22 22 23 23 23 23 23 23 24 24 24 24  
24 24 24 24 24 25 25 25 25 25 26 26 26 27 27 27 27 27 27 27 28 28 28 28  
28 28 29 29 29 29 **30 30 30 30 30 30 30 30 30 30 30 30 31 31 31 31 31 31**  
**31 32 32 32 33 33 33 33 33 33 33 34 35 35 35 35 35 35 35 35 35 35 35 35**  
**36 36 36 36 36 36 36 36 36 36 36 36 37 37 37 37 37 38 38 38 38 38 38**  
**39 39 39 39 39 39 39 39 39 39** 40 40 40 40 40 41 41 41 42 42 42 42 42 42 42  
43 43 43 44 44 44 45 45 45 45 45 45 45 45 45 45 46 46 46 46 46 46 47  
47 47 47 47 47 47 47 48 48 48 48 48 48 48 48 48 49 49 49 49 49 49 49 **50**  
**50 50 50 50 50 50 50 51 51 51 51 52 52 52 52 53 53 53 53 54 54 54 54**  
**54 54 55 55 55 55 55 55 55 56 56 56 56 57 57 58 58 58 58 58 58 59 60**  
60 60 60 60 61 61 61 62 62 62 63 63 64 64 64 64 64 65 65 67 **70 71 71 76**  
80

*Let's use 8 age classes:*

<b><u>Class</u></b> <b><u>Limits</u></b>	<b><u>Class</u></b> <b><u>Boundaries</u></b>	<b><u>Frequency</u></b>	<b><u>Cumulative</u></b> <b><u>Frequency</u></b>
---	---	-------------------------	---

- ☑ Sanity check: make sure that the last cumulative frequency actually matches with the number of observations!

**Cumulative frequency distribution** – a distribution that shows the number of observations less than or equal to a specific value.

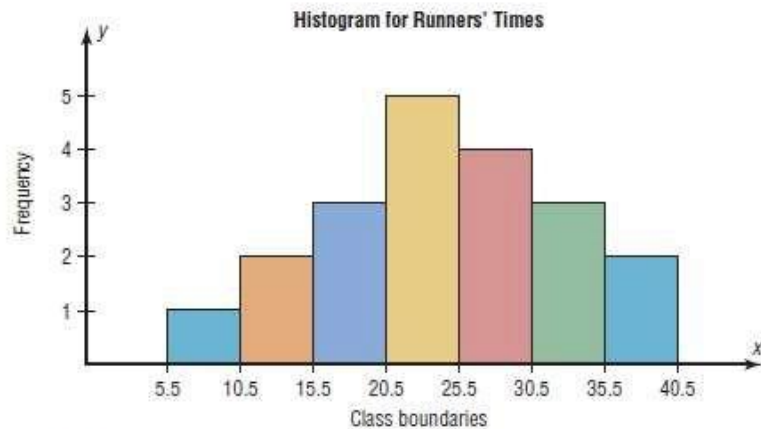
- These are easy to determine from a frequency distribution.
- Simply add the number of observations for each class with the previous classes.
- Start with the first lower class boundary and then use every upper class boundary.

# Continuing the titanic first class passenger age example:

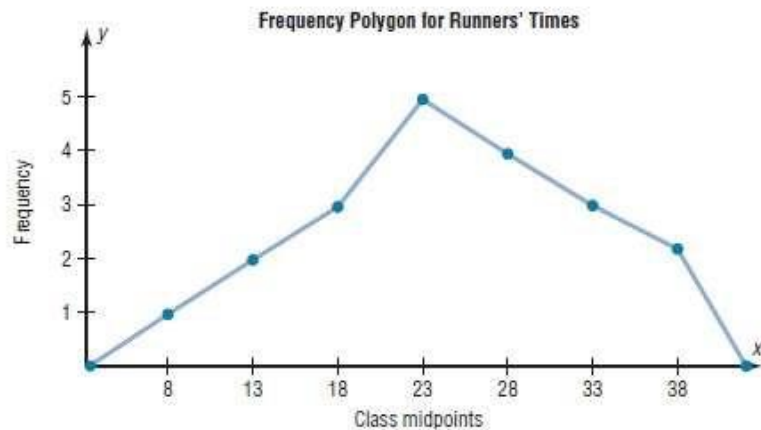
## Cumulative Frequency

Less than	
Less than	
Less than	
Less than	
Less than	
Less than	
Less than	
Less than	
Less than	
Less than	
Less than	

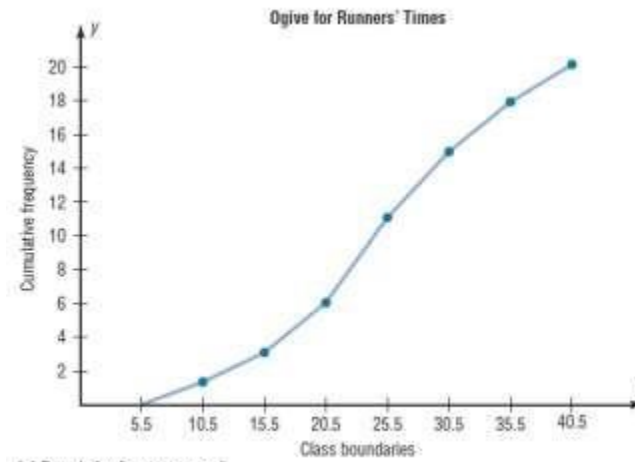
# Section 2-2 Histograms, Frequency Polygons, and Ogives (making pretty pictures for counts from **quantitative** variables)



(a) Histogram



(b) Frequency polygon



(c) Cumulative frequency graph

*Pronounced "O-JIVE"*



**Histogram** – a graph that displays quantitative data by using contiguous vertical bars (unless the frequency of a class is 0) of various heights to represent the frequencies of the classes.

Steps:

1. Draw and label the  $x$  and  $y$  axes.
2. Represent the frequency on the  $y$  axis and the class boundaries on the  $x$  axis.
3. Draw vertical bars which have height corresponding to the frequency of each class.



**Example:** Use the frequency distribution for the ages of titanic first class passengers to create a histogram.

**Frequency Polygon** – a graph that displays the data by using lines that connect points plotted for the frequencies at the **midpoints of the classes**. The frequencies are represented by the heights of the points.

***Steps:***

1. Find the midpoint of each class (by averaging the **class boundaries**).
2. Draw the  $x$  and  $y$  axes. Label the  $x$  axis with the midpoints of each class and use a suitable scale for the  $y$  axis frequencies.
3. Using the midpoints of the classes for the  $x$  value and the frequencies for the  $y$  value, plot the points.
4. Connect the adjacent points with line segments Draw a line back to the  $x$ -axis at the beginning and end of the graph, at the same distance that the previous and next midpoints would be located.
  - a. To calculate these points, subtract the class width from the midpoint of the first class and add the class width to the midpoint of the last class.

**Example:** Draw a frequency polygon for the ages of titanic first class passengers.

## Frequency polygons vs histograms:

- They are two different ways to represent the same data set.
- The choice is at the researcher's discretion.
- They show the overall shape of the distribution of the variable.

**Ogive (“o-jive”)** – a graph that represents the cumulative frequencies for the classes in a frequency distribution.

- Ogives are used to visually represent how many values are below a certain upper class boundary.

***Steps:***

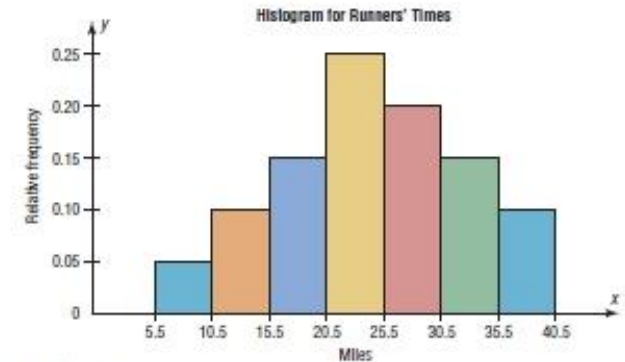
1. Draw the  $x$  and  $y$  axes. Label the  $x$  axis with the class boundaries. Use an appropriate scale for the  $y$  axis.
2. Plot the cumulative frequency at each class boundary.
3. Connect the points with lines.

**Example:** Draw an ogive for the ages of titanic first class passengers.

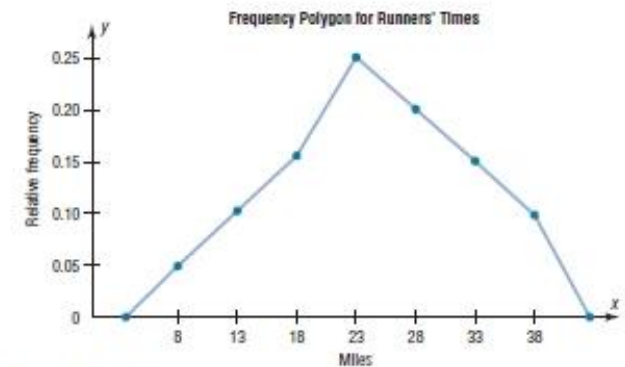
**Relative frequency graph** – a distribution graph that uses proportions instead of frequencies.

Histograms, frequency polygons, and ogives can be converted into relative frequency graphs by converting the frequencies into proportions/percents.

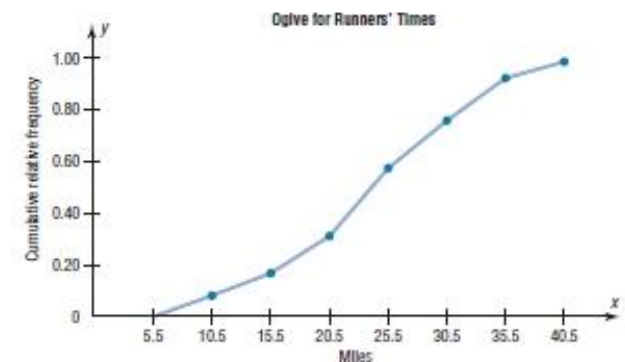
- Divide the frequency (or cumulative frequency) by the total number of observations.
- Use the relative frequencies on the y-axis rather than the frequencies

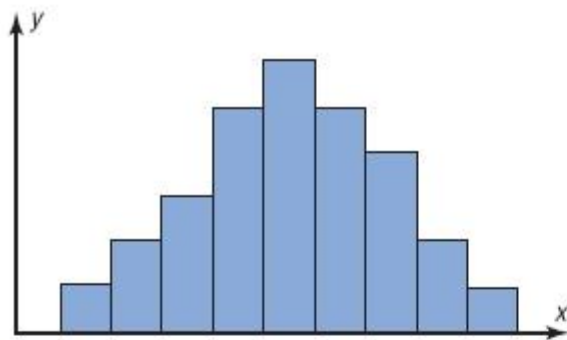


(a) Histogram

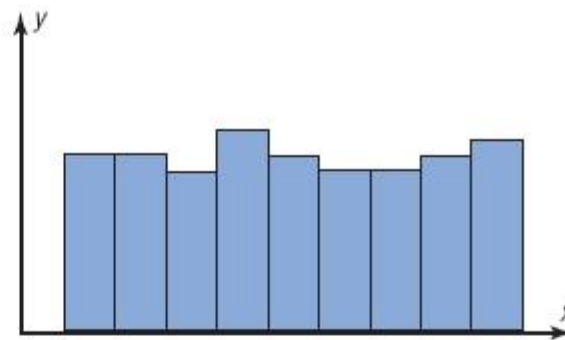


(b) Frequency polygon

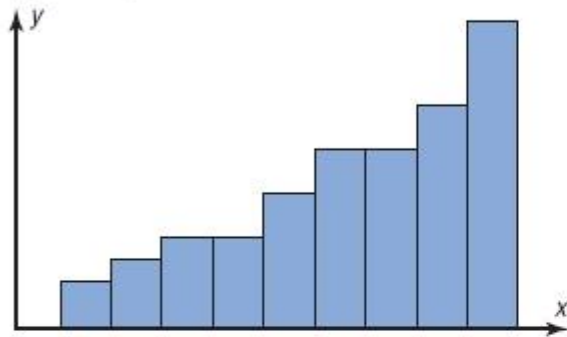




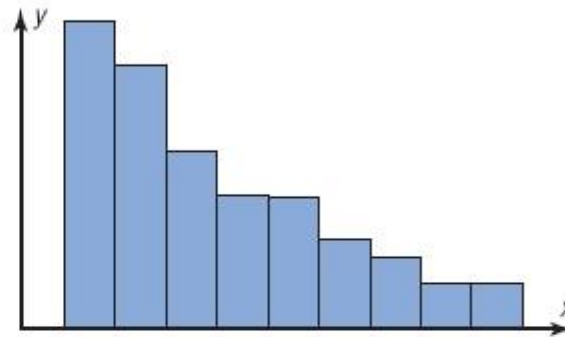
(a) Bell-shaped



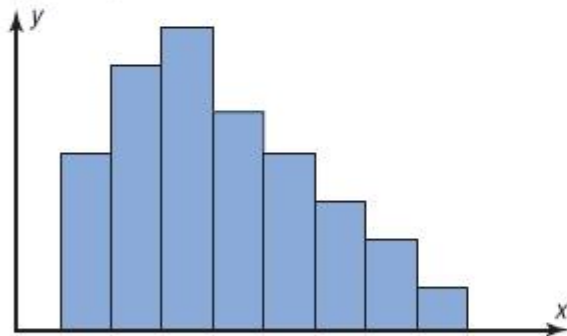
(b) Uniform



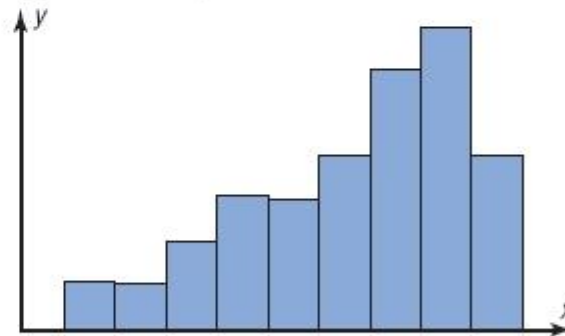
(c) J-shaped



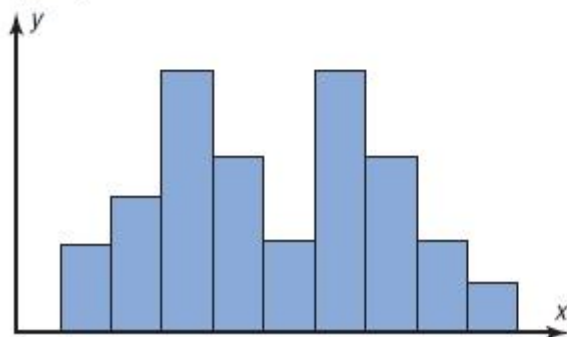
(d) Reverse J-shaped



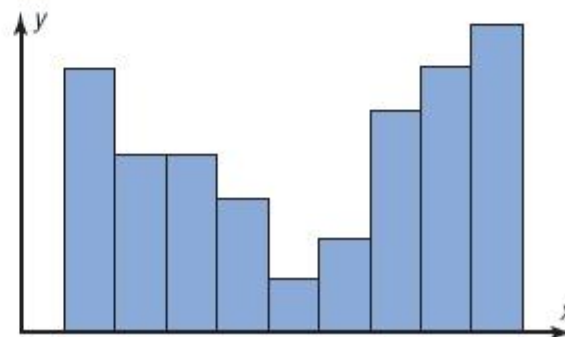
(e) Right-skewed



(f) Left-skewed



(g) Bimodal



(h) U-shaped

## DISTRIBUTION SHAPE

- When describing data, it is important to be able to recognize the shapes of the distribution values.
- Being able to identify the shape of the distribution will be important later when we are choosing appropriate statistical tests.
- Histograms and frequency polygons allow us to see the distribution of the data.



## Section 2-3 Other Types of Graphs

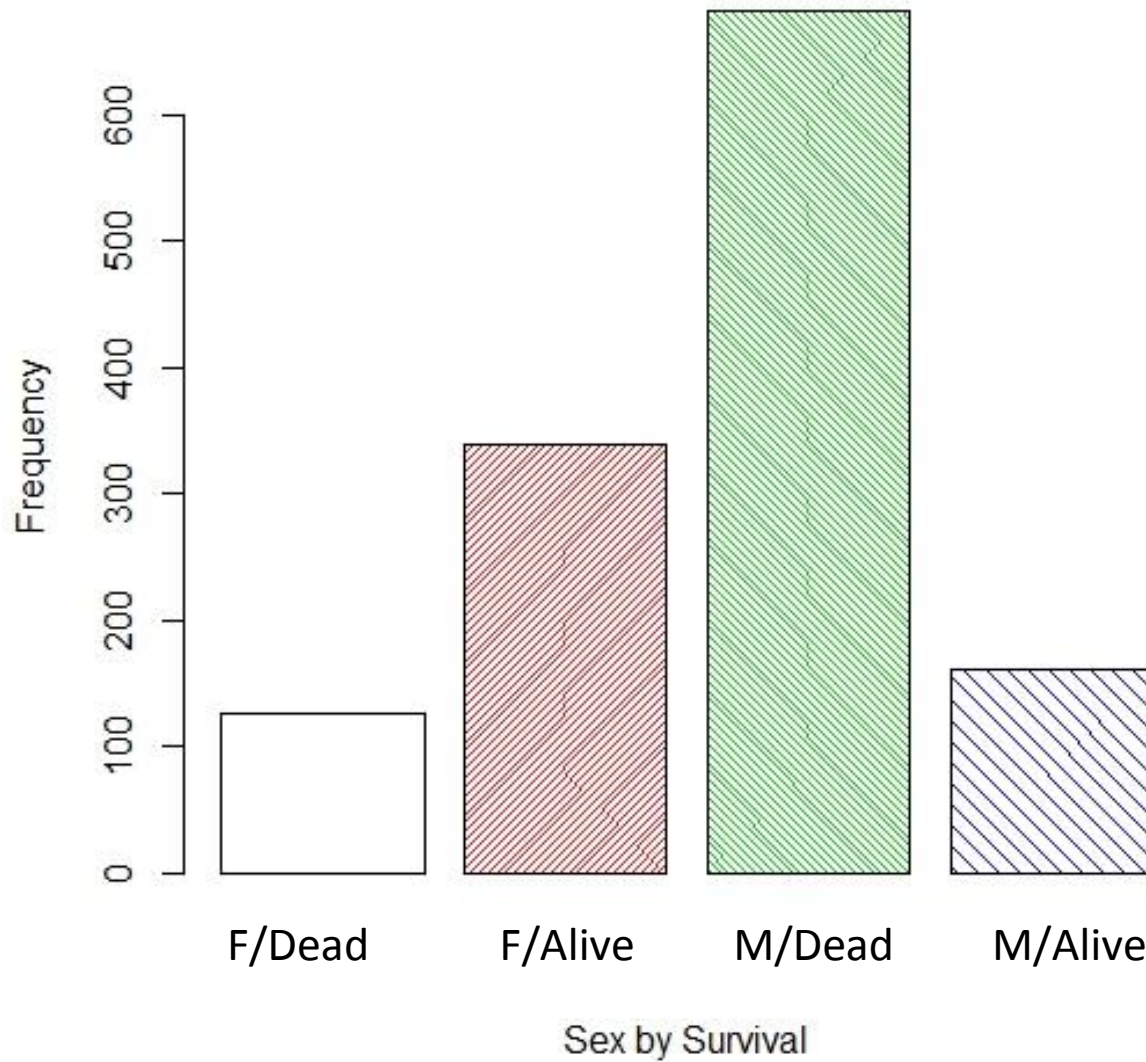
**Bar Graph** – represents the data by using vertical or horizontal bars whose heights or lengths represent the frequencies of the data. Used for **qualitative** data.

- The bars of a horizontal bar graph are horizontal.
- Relevant to qualitative variables.

### *Steps:*

1. Draw the  $x$  and  $y$  axes. Label the  $x$  axis and  $y$  axis.
2. Draw of the appropriately sized bars for each category.

The following bar graph is for the survival by sex groups on the titanic:

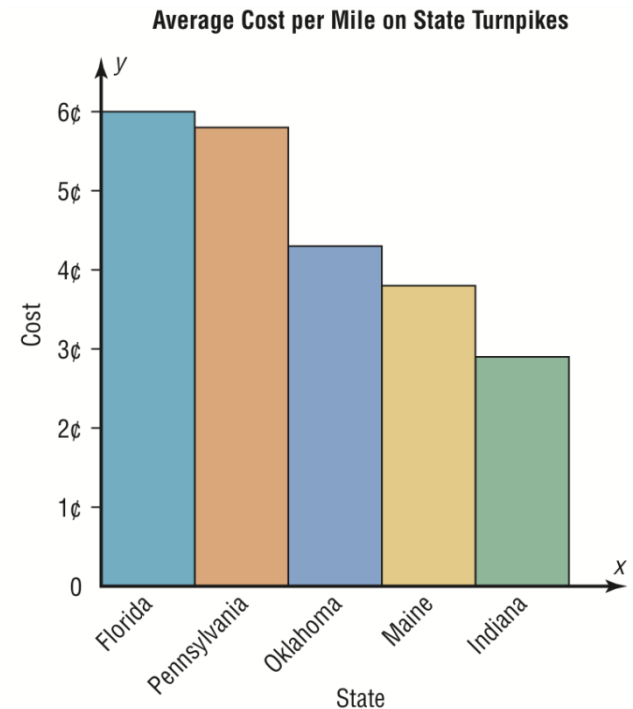


**Pareto Chart** – used to represent a frequency distribution for a categorical variable, and **the frequencies are displayed by the heights of vertical bars, which are arranged in order from highest to lowest.**

- This is a bar chart with a special ordering.
- The Pareto chart has the qualitative variable on the x axis.
- The highest frequency categories are on the left and the lowest frequency categories are on the right.

### **Suggestions for Drawing Pareto Charts**

- Make the bars the same width.
- Arrange the data from largest to smallest according to frequency.
- Make the units that are used for the frequency equal in size.

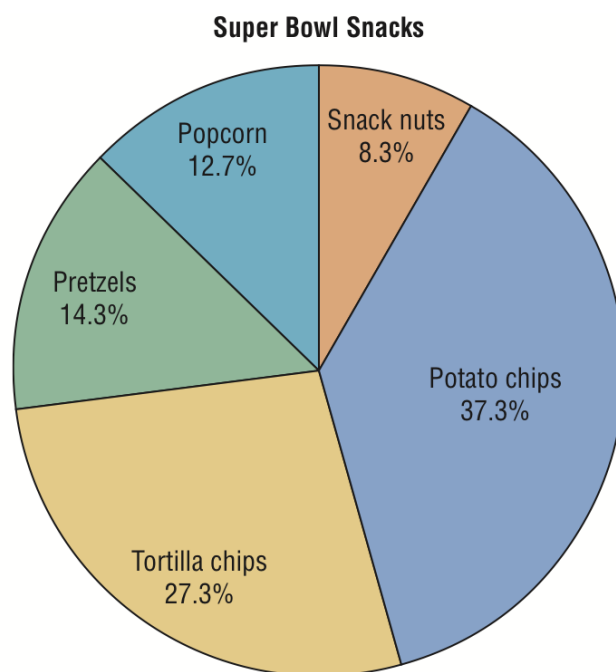


**Example:** Construct a horizontal bar graph and a Pareto chart for the passenger class by survival categories

<b>Group</b>	<b>Frequency</b>
1 <sup>st</sup> Class, Survived	200
1 <sup>st</sup> Class, Dead	123
2 <sup>nd</sup> Class, Survived	119
2 <sup>nd</sup> Class, Died	158
3 <sup>rd</sup> Class, Survived	181
3 <sup>rd</sup> Class, Dead	528



**Pie Graph** – a circle that is divided into sections or wedges according to the percentage of frequencies (relative frequency) in each category of the distribution.



When analyzing a pie graph, examine the size of the sections of the pie graph and compare them to other sections and how they compare to the whole.

## **TWO OTHER USEFUL GRAPHS:**

**Time series graph** – represents data that occur over a specific period of time (time is your x-axis).

### ***Steps:***

1. Draw and label the  $x$  and  $y$  axes.
2. Plot a point for each time at the corresponding value in the  $y$  direction.
3. Connect the points.

### ***Notes on Time Series Graphs:***

1. Do not extend the ends of the times series graph to the  $x$ - axis unless it makes sense that the data would fall to zero (which is rarely the case)!
2. Two data sets can be compared on the same graph, called a compound time series graph.
3. When you analyze a time series graph, look for patterns or trends that occur over time.

**Example:** Draw a time series graph to represent the data for the number of airline departures (in millions) for the given years. Over the years, is the number of departures increasing, decreasing, or about the same?

<b>Year</b>	<b># of Departures (in millions)</b>
1996	7.9
1997	9.9
1998	10.5
1999	10.9
2000	11.0
2001	9.8
2002	10.1





**Stem and leaf plot** – a data plot that uses part of the data value as the stem and part of the data value as the leaf to form groups or classes.

***Steps:***

1. Order the data.
2. Separate the data by the first digit.
3. Use the first digit as the stem and trailing digit as a leaf.

**Notes on Stem and Leaf Plots:**

- When you analyze a stem and leaf plot, look for peaks and gaps in the distribution.
- See if the distribution is symmetric or skewed.
- Check to see how variable the data is by examining how spread out the data is.
- A back-to-back stem and leaf plot can be created to compare two data sets using the same digits as stems.

**Example:** Consider the following data on car thefts in large cities and draw a stem and leaf plot:

38 53 53 56 69 89 94 41 58 68 66 69  
89 52 50 70 83 81 80 90 74 50 70 83

**Stem (leading digit)                      Leaf (trailing digit)**



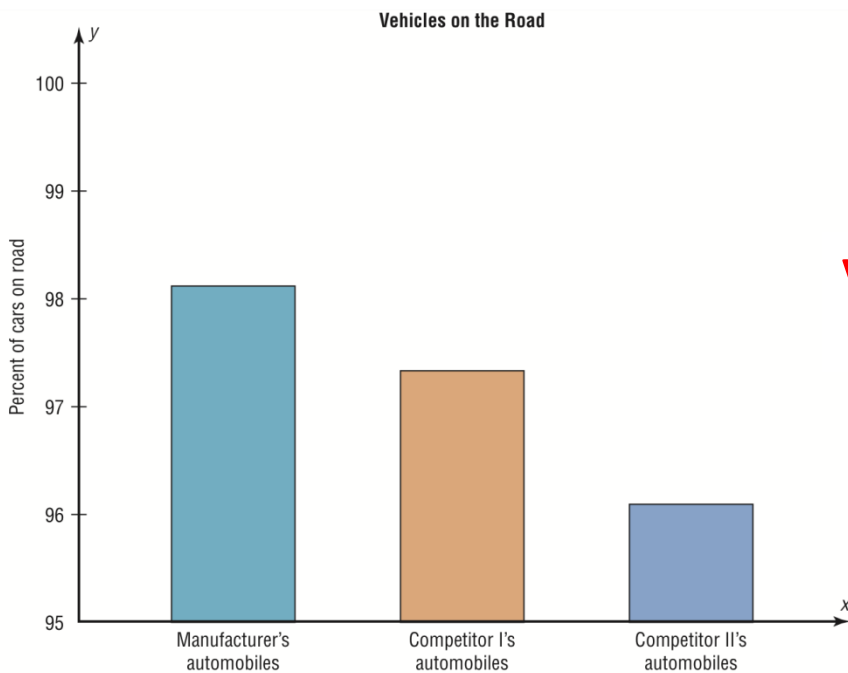
**You can construct all of the graphs we have discussed in Excel. See pages 91-93 of the textbook for a description.**

## Misleading Graphs

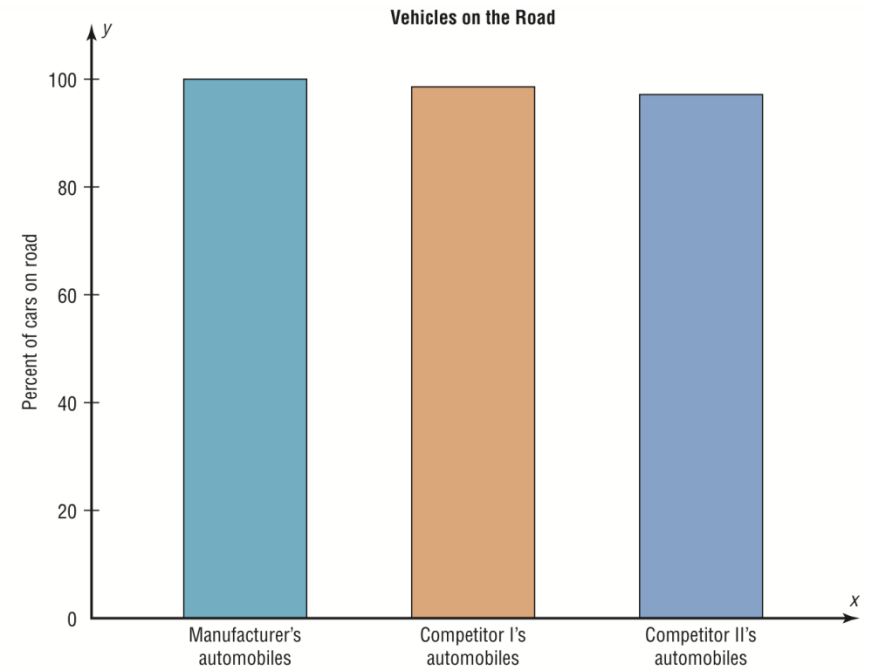
When creating graphs we must be careful not to misrepresent the data by inappropriately drawing the graphs. When the graphs misrepresent the data they can lead the reader to draw false conclusions.

Let's look at a few examples. Why are they misleading and how should they be correctly portrayed?

**Example 1: Using a scale that doesn't go from 0 to 100%.** The graph below shows the percentage of the manufacturer's automobiles still on the road and the percentage of its competitors' automobiles still on the road. Is there a large difference?

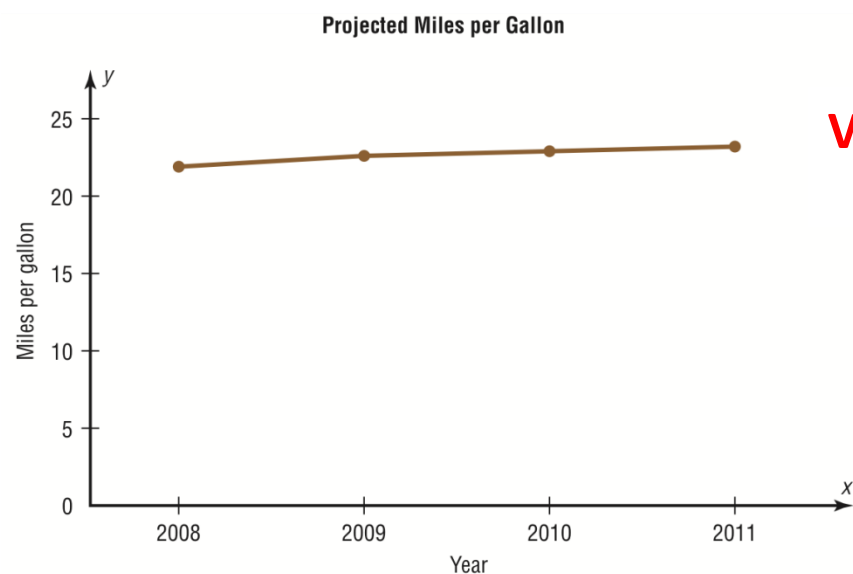


Versus

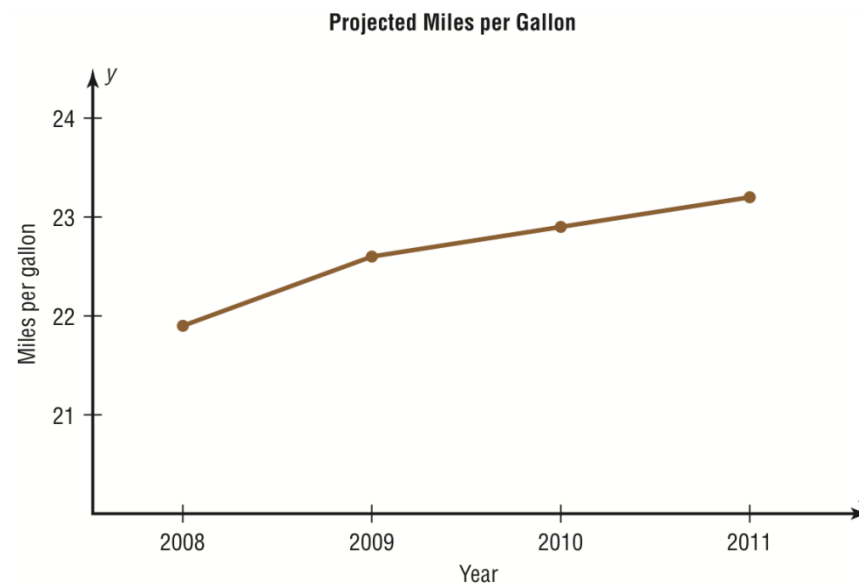


On the full scale, it is clear that the manufacturer's numbers are comparable!

## Example 2: Changing the units or starting point on they-axis.

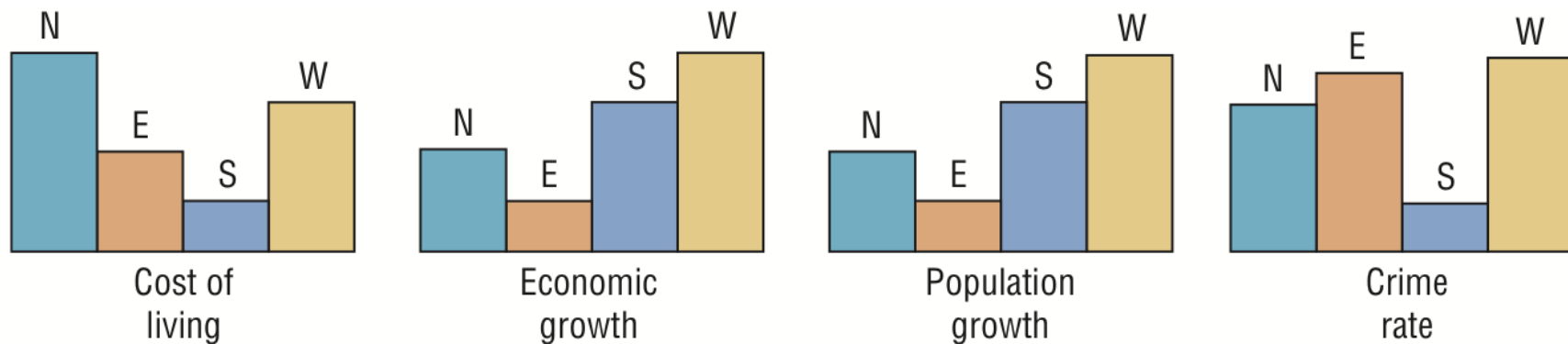


Versus



Looks like much more change as occurred. That said, which way is the “correct” way to plot the data?

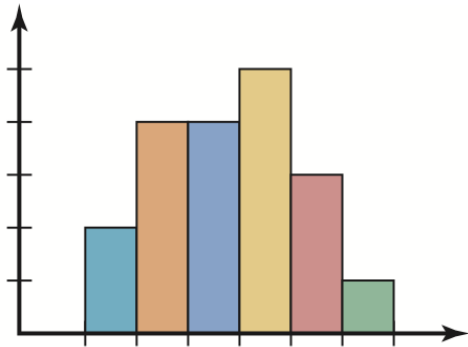
### Example 3: Not labeling or not putting units on the y-axis.



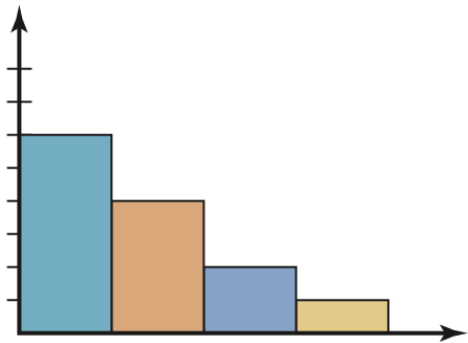
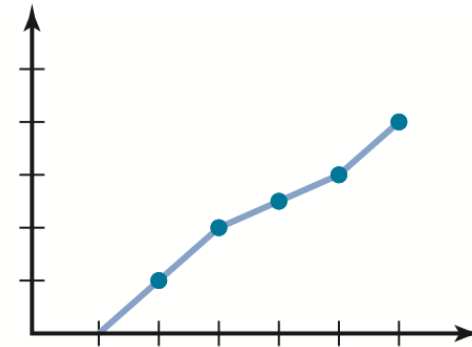
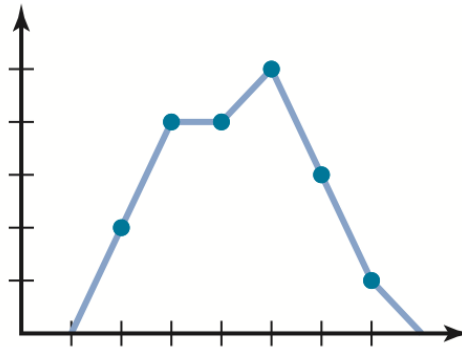
With no given scale it is impossible to tell if the differences are meaningful. Not to mention, what do N, E, S and W represent?



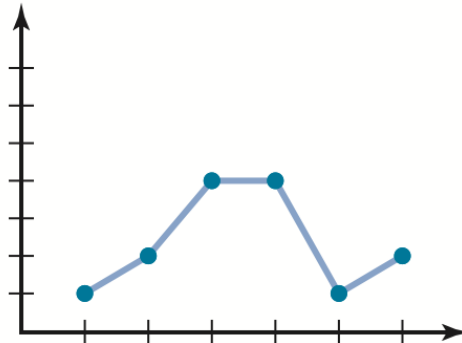
# Summary of Graphs



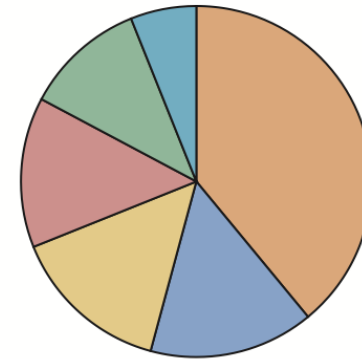
**(a)** Histogram; frequency polygon; ogive  
Used when the data are contained in a grouped frequency distribution.



**(b)** Pareto chart  
Used to show frequencies for nominal or qualitative variables.



**(c)** Time series graph  
Used to show a pattern or trend that occurs over a period of time.



**(d)** Pie graph  
Used to show the relationship between the parts and the whole.  
(Most often uses percentages.)

What graphs are used for qualitative data?

What graphs are used for quantitative data?

**Histogram**

**Bar chart**

**Frequency polygon**

**Pareto chart**

**Ogive**

**Stem and leaf plot**

**Time series graph**

**Pie chart**