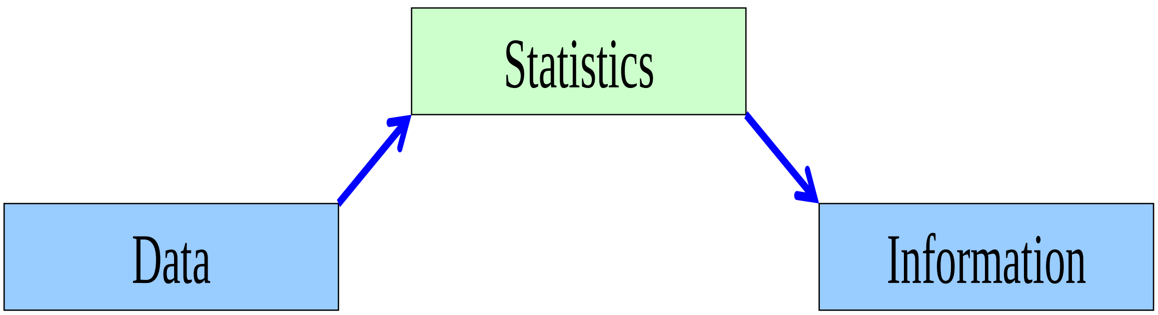
Graphical and Tabular Summarization of Data

OPRE 6301

Introduction and Re-cap. . .

Descriptive statistics involves arranging, summarizing, and presenting a *set of data* in such a way that useful *information* is produced.



It makes use of graphical techniques and numerical descriptive measures (such as averages) to summarize and present the data.

The graphical and tabular methods presented here apply to both entire populations and samples drawn from populations.

Definitions...

A **random variable**, or simply **variable**, is a characteristic of a population or sample.

Examples: Student grades, which *varies* from student to student; and stock prices, which *varies* from stock to stock as well as over time.

Typically denoted by a capital letter: X, Y, Z...

The **values** of a variable are possible observations or realizations of that variable. The possible values of a variable usually land in a specified range. Examples:

Student Grades: the interval $[0, 100]$.

Stock Prices: nonnegative real numbers.

Data are the *observed* values of a variable. Examples:

Grades of a sample of students: $\{34, 78, 64, 90, 76\}$

Prices of stocks in a portfolio: $\{\$54.25, \$42.50, \$48.75\}$

Types of Data...

Data fall into three main groups:

- Interval Data
- Nominal Data
- Ordinal Data

Details...

Interval Data...

Interval Data are:

- real numbers, e.g., heights, weights, prices, etc.
- also referred to as **quantitative** or **numerical** data.

Arithmetic operations can be performed on interval data, thus it is meaningful to talk about:

2 * Height, or

Price + $1,

and so on.

Nominal Data...

Nominal Data are:

- **names** or **categories**, e.g., {Male, Female} and {single, Married, Divorced, Widowed}.
- also referred to as **qualitative** or **categorical** data.

Arithmetic operations do *not* make sense for nominal data (e.g., does Widowed / 2 = Married ?!).

Ordinal Data...

Ordinal Data are also categorical in nature, but their values have an *order*. Example:

Course Ratings: Poor, Fair, Good, Very Good, Excellent.

Student Grades: F, D, C, B, A.

Taste Preferences: First Choice, Second Choice, Last Choice.

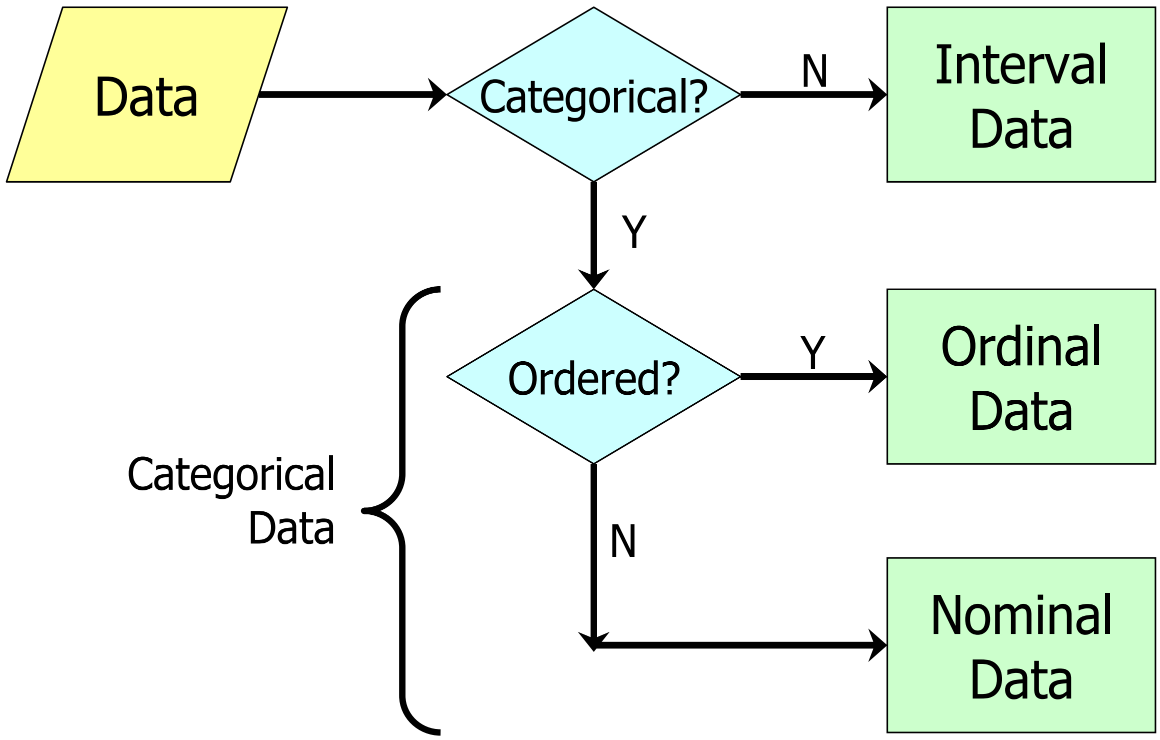
Thus, while it is still not meaningful to do arithmetic on ordinal data (e.g., does 2 * fair = very good?!), we can say things like:

$$\text{Excellent} > \text{Poor, or}$$

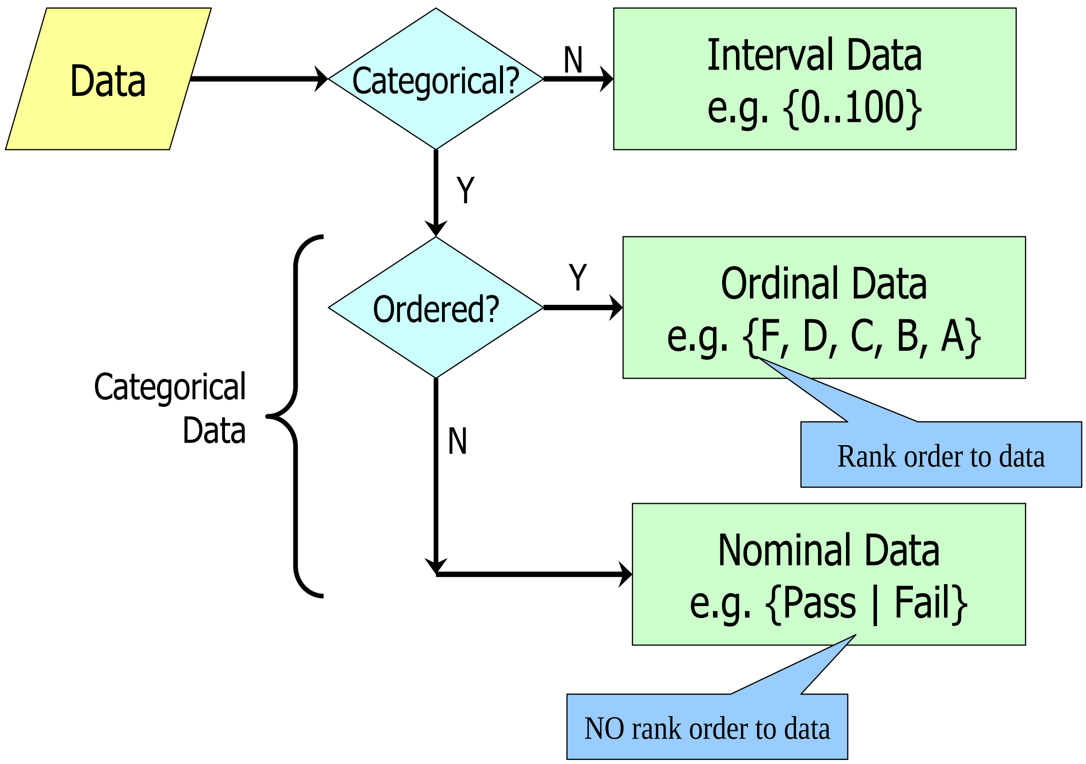
$$\text{Fair} < \text{Very Good}$$

That is, order is maintained no matter what numeric values are assigned to each category.

Information Hierarchy. . .



Example: For student grades, we have



Thus, information is lost as we move down this hierarchy.

In terms of calculations, we also have:

- All calculations are permitted on *interval* data.
- Only calculations involving a ranking process, or comparison, are allowed for *ordinal* data.
- No calculations are allowed for *nominal* data, other than counting the number of observations in each category.

Nominal Data — Tables and Graphs. . .

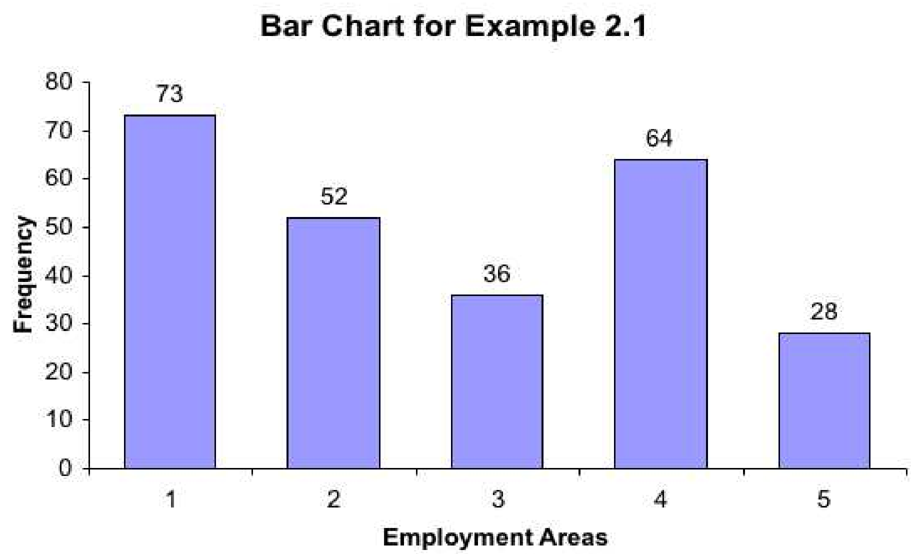
Nominal (and ordinal) data can be summarized in a table that lists individual categories and their respective frequency counts, i.e., a **frequency distribution**.

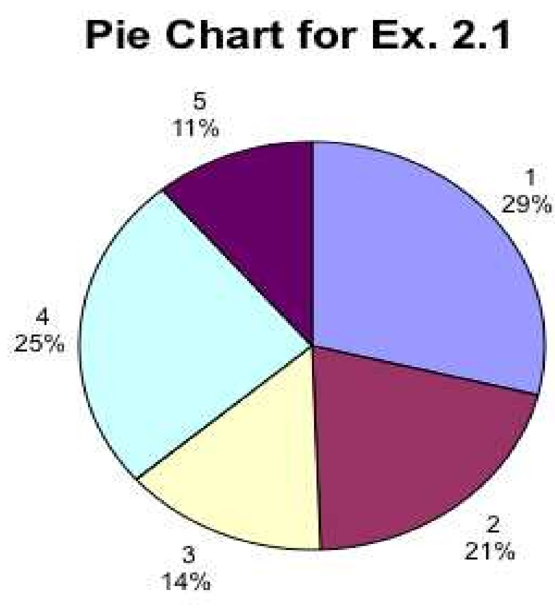
One can also use a **relative frequency distribution**, which lists the categories and the *proportion* with which each occurs.

Example: Student Placement

Area	Frequency	Relative Frequency
Accounting	73	28.9%
Finance	52	20.6%
General Management	36	14.2%
Marketing/Sales	64	25.3%
Other	28	11.1%
Total	253	100

Frequency distributions and relative frequency distributions can also be summarized as **bar charts** and **pie charts**, respectively.





Interval Data — Tables and Graphs...

Interval data are typically summarized in a **histogram**. Steps for constructing a histogram is as follows.

Step 1: Partition the data range into *classes* or *bins*. General guidelines are:

- Use between 6 and 15 bins. One suggested formula (Sturges) is:

  $$\text{Number of Classes} = 1 + 3.3\log(n)$$

  where n is the total number of observations.
- All bins should have the same width.
- Use "natural" values for the bin width (e.g., 10–20, 20–30, etc.).

Step 2: Count the number of observations that fall in each class.

Step 3: Summarize the resulting frequency distribution as a table or as a bar chart.

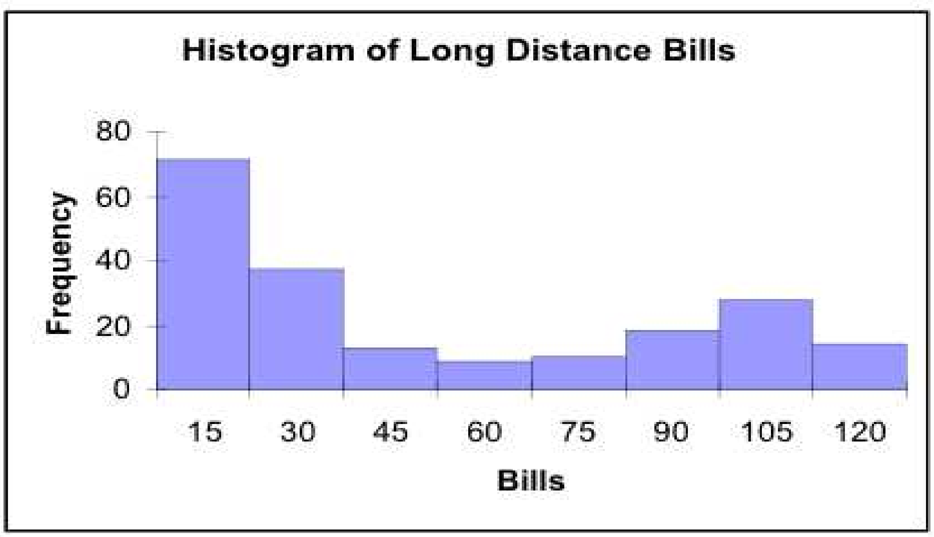
Example: Monthly Long-Distance Telephone Bills

We have (Xm02-04.xls):

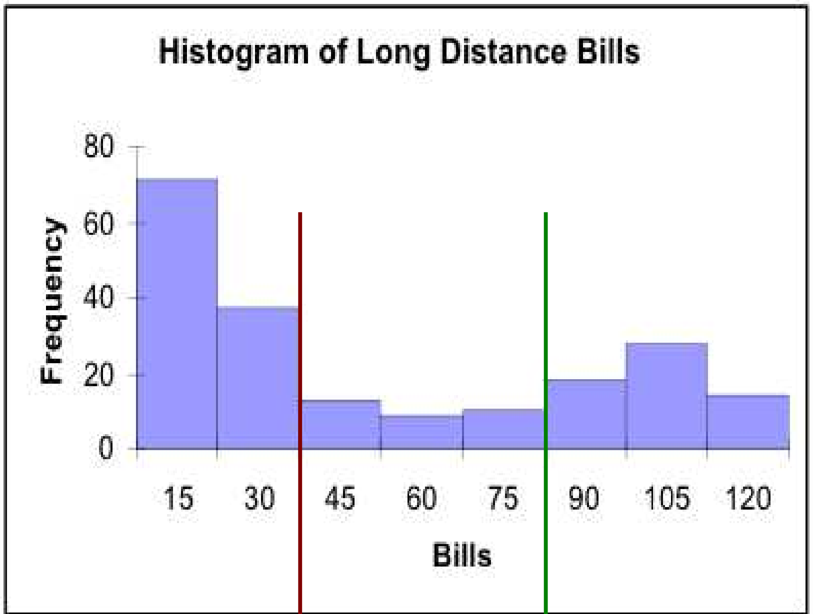
- $n = 200$ (number of subscribers surveyed)
- $$\begin{aligned} \text{Range} &= \text{Largest Observation - Smallest Observation} \\ &= \$119.63 - \$0 \\ &= \$119.63 \end{aligned}$$
- Suggested Number of Classes $= 1 + 3.3 \log(n) = 8.59$
- Since $120/8.59 = 13.97$, Width $= 15$ seems to be a "natural" choice
- Number of Classes $= 120/15 = 8$

The results are:

Lower Limit	Upper Limit	Frequency
0	15	71
15	30	37
30	45	13
45	60	9
60	75	10
75	90	18
90	105	28
105	120	14
	Total	200



Observations. . .



about half (71+37=108) of the bills are “small”, i.e. less than \$30

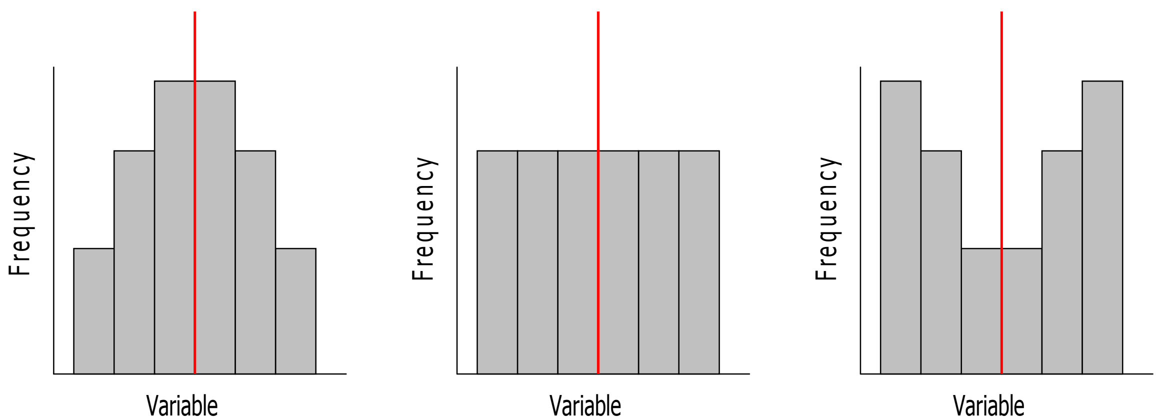
(18+28+14=60)÷200 = 30% i.e. nearly a third of the phone bills are \$90 or more.

There are only a few telephone bills in the middle range.

Shapes of Histograms...

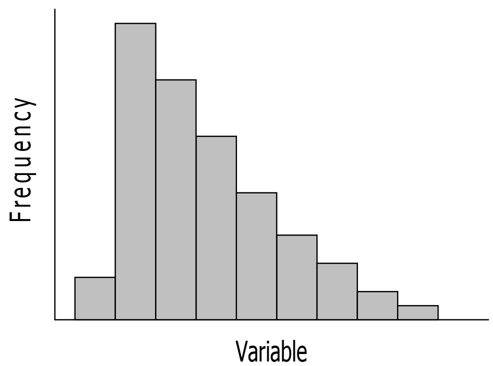
Symmetry

A histogram is said to be **symmetric** if, when we draw a *vertical line* down the center of the histogram, the two sides are identical in shape and size:

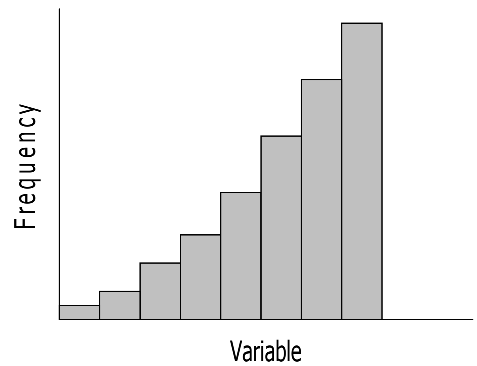


Skewness

A skewed histogram is one with a long tail extending to either the right or the left:



Positively Skewed



Negatively Skewed

Modality

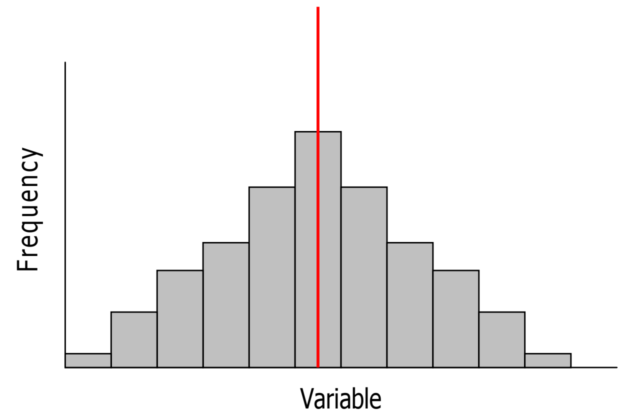
A **unimodal** histogram is one with a *single peak*, while a **bimodal** histogram is one with *two peaks*:

Bell Shape (or Mound Shape)

A special type of *symmetric unimodal* histogram is one that is bell shaped:

Many statistical techniques require that the population be bell shaped.

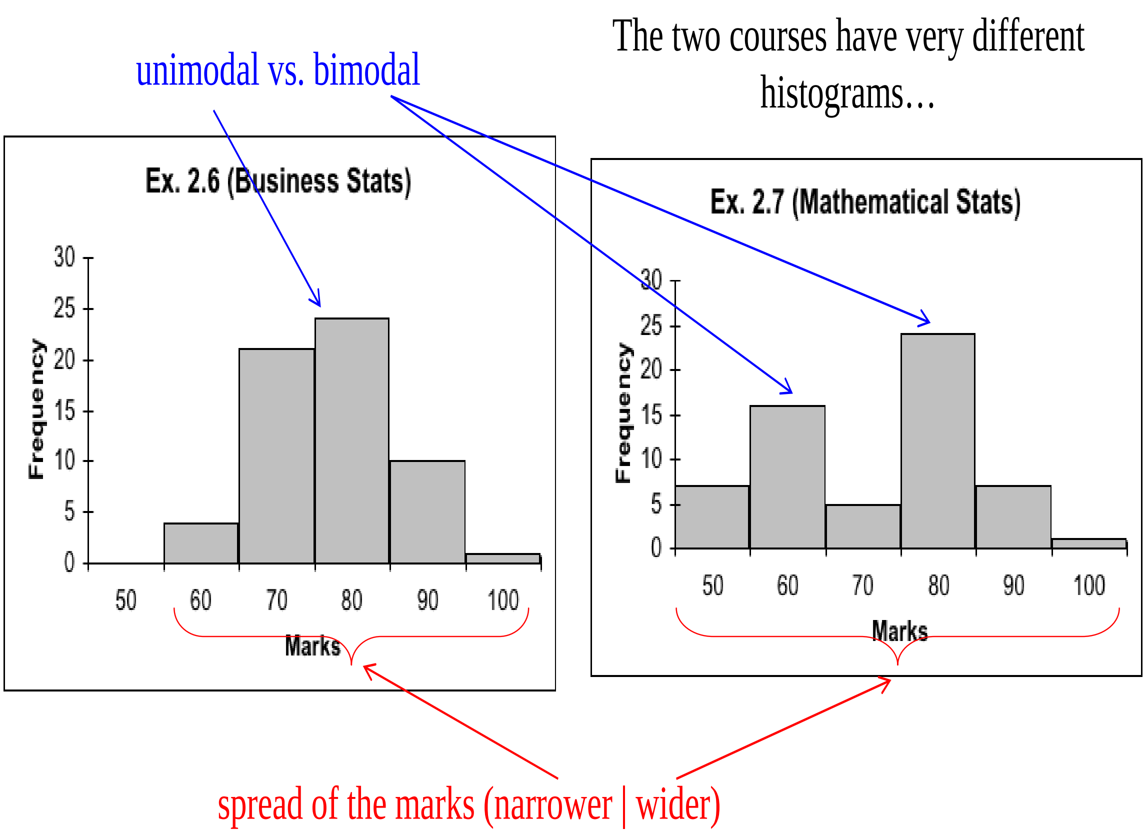
Drawing the histogram helps verify the shape of the population in question.



Bell Shaped

Comparison of Histograms. . .

Comparing histograms often yields useful information. As an example, contrast the following two histograms:



Other Graphical Approaches...

Stem and Leaf Display

...attempts to retain information about individual observations that would normally be lost in the creation of a histogram.

Idea: Split each observation into two parts, a **stem** and a **leaf**.

Suppose the observed value is **42.19**
There are several ways to split it up...

We could split it at the decimal point.
Or split it at the "tens" position (while rounding to the nearest integer in the "ones" position)

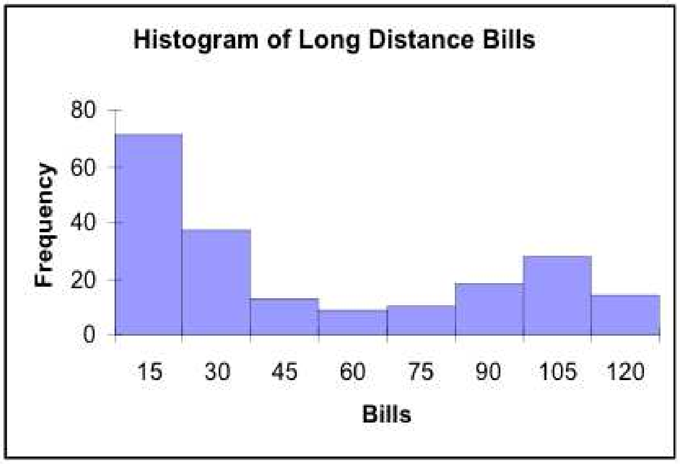
Stem	Leaf
42	19
4	2

Continue this process for all the observations in the long-distance-bills data. Let each possible stem be a "class" and list all observed leafs for each stem, resulting in...

Stem	Leaf
0	0000000000111112222223333345555556666666778888999999
1	000001111233333334455555667889999
2	0000111112344666778999
3	001335589
4	1**2**4445589
5	33566
6	3458
7	022224556789
8	334457889999
9	00112222233344555999
10	001344446699
11	124557889

Thus, we still have access to our original data point's **value**!

Histogram and stem-and-leaf display are similar. . .



Stem	Leaf
0	000000001111122222223333334555555566666667788889999999
1	000001111122333333344555556678899999
2	0000111112344666778999
3	001335589
4	124445589
5	33566
6	3458
7	022224556789
8	334457889999
9	00112222233344555999
10	0013444669
11	124557889

Ogive

…(pronounced "Oh-jive") is a graph of a **cumulative frequency distribution**.

We create an ogive in three steps…

Step 1: Calculate **relative frequencies**, defined as

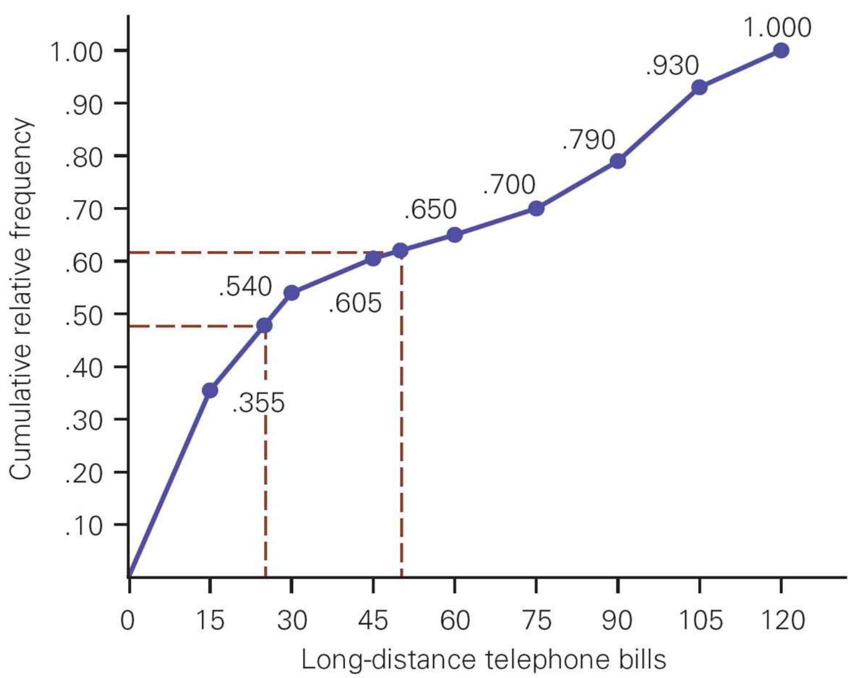
$$\text{Relative Frequency} = \frac{\text{Number of Observations in a Class}}{\text{Total Number of Observations}}$$

Step 2: Calculate the *cumulative* relative frequencies by adding the current class' relative frequency to the previous class' cumulative relative frequency. That is, we accumulate relative frequencies.

Step 3: Graph the cumulative relative frequencies.

For the long-distance-bills data, we have. . .

Lower Limit	Upper Limit	Relative Frequency	Cumulative Relative Frequency
0	15	$71/200 = .355$	.355
15	30	$37/200 = .185$	.540
30	45	$13/200 = .065$	.605
45	60	$9/200 = .045$	.650
60	75	$10/200 = .050$	.700
75	90	$18/200 = .090$	.790
90	105	$28/200 = .140$	.930
105	120	$14/200 = .070$	1.00
	Total	$200/200 = 1$	



What telephone bill value is at the 50th percentile?

Two Nominal Variables...

So far we havve looked at tabular and graphical techniques for one variable (either nominal or interval data).

A **contingency table** (also called a cross-classification table or cross-tabulation table) is used to describe the relationship between *two* nominal variables.

A contingency table lists the *frequency* of *each combination* of the values of the two variables.

Example: Newspaper Preference

A sample of newspaper readers was asked to report which newspaper they read: Globe and Mail (1), Post (2), Star (3), or Sun (4), and to indicate whether they were blue-collar worker (1), white-collar worker (2), or professional (3).

A contingency table is constructed as follows:

Reader	Newspaper	Occupation
1	2	2
2	4	1
3	1	2
.	.	.
.	.	.
352	2	3
353	3	1
354	3	2

This reader's response is captured as part of the total number on the contingency table…

Table 2.9 Contingency Table of Frequencies for Example 2.8

	Occupation			
Newspaper	Blue Collar	White Collar	Professional	Total
G&M	27	29	33	89
Post	18	43	51	112
Star	38	21	22	81
Sun	37	15	20	72
Total	120	108	126	354

Interpretation

The relative frequencies in columns 2 and 3 are similar, but there are large differences between columns 1 and 2 and between columns 1 and 3.

Table 2.10 Column Relative Frequencies for Example 2.8

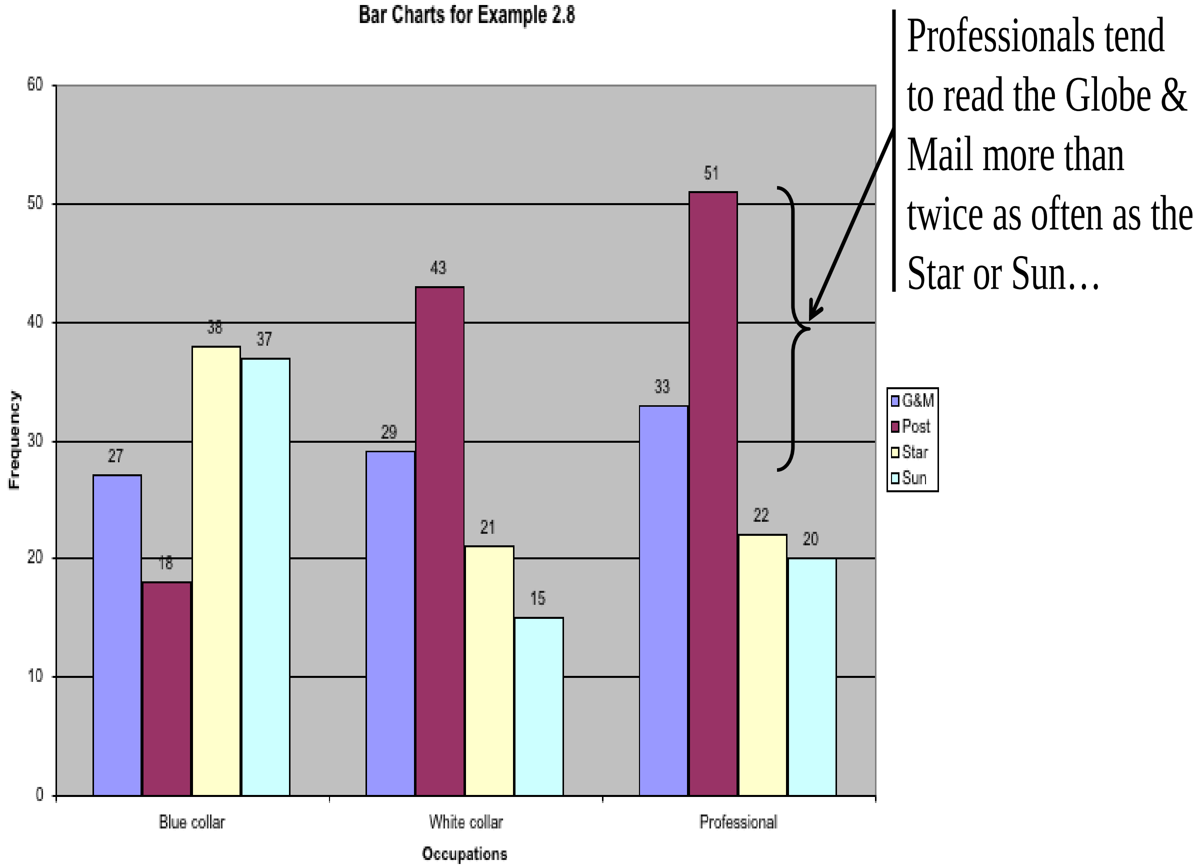
	Occupation		
Newspaper	Blue Collar	White Collar	Professional
G&M	27/120 =.23	29/108 = .27	33/126 = .26
Post	18/120 = .15	43/108 = .40	51/126 = .40
Star	38/120 = .32	21/108 = .19	22/126 = .17
Sun	37/120 = .31	15/108 = .14	20/126 = .16

similar

dissimilar

This tells us that blue collar workers tend to read different newspapers from both white collar workers and professionals, and that white collar and professionals are quite similar in their newspaper choice.

Using the data from the contingency table, we can also create a bar chart:



Two Interval Variables...

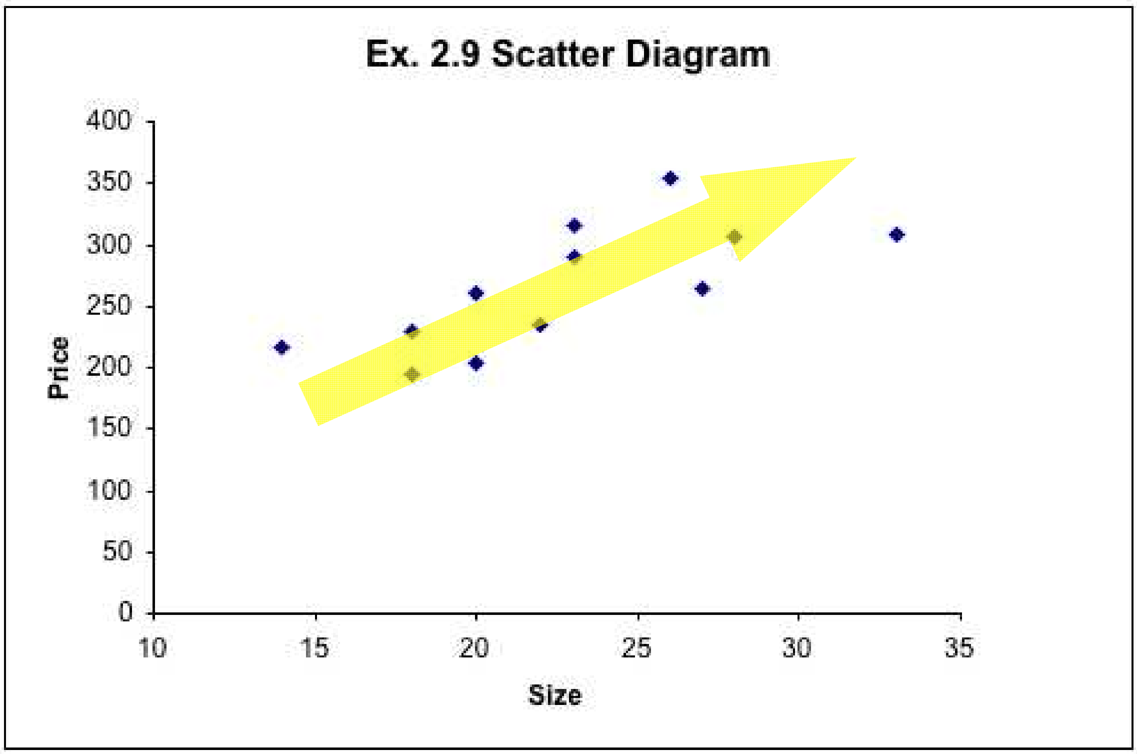
Moving from nominal data to interval data, we are frequently interested in how *two* interval variables are related.

To explore this relationship, we employ a **scatter diagram**, which plots two variables against one another.

The **independent variable** is labeled X and is usually placed on the horizontal axis, while the other, **dependent variable**, Y, is mapped to the vertical axis.

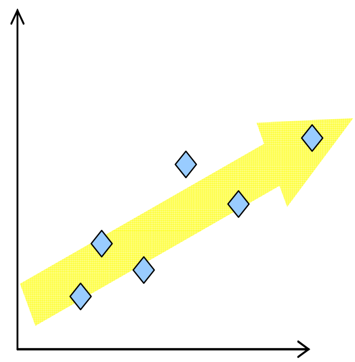
Example: Selling Price of a House

A real estate agent wanted to know to what extent the selling price of a house is related to its size...

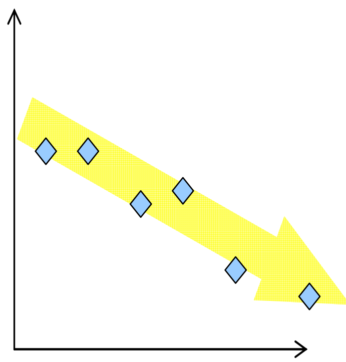


It appears that in fact there is a relationship: the greater the house size the greater the selling price.

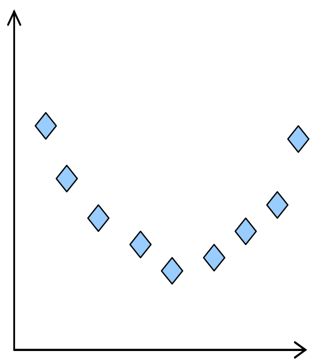
Some possible patterns are...



Positive Linear Relationship



Negative Linear Relationship



Weak or Non-Linear Relationship

Linearity and Direction are two concepts we are often interested in.

Time Series Data...

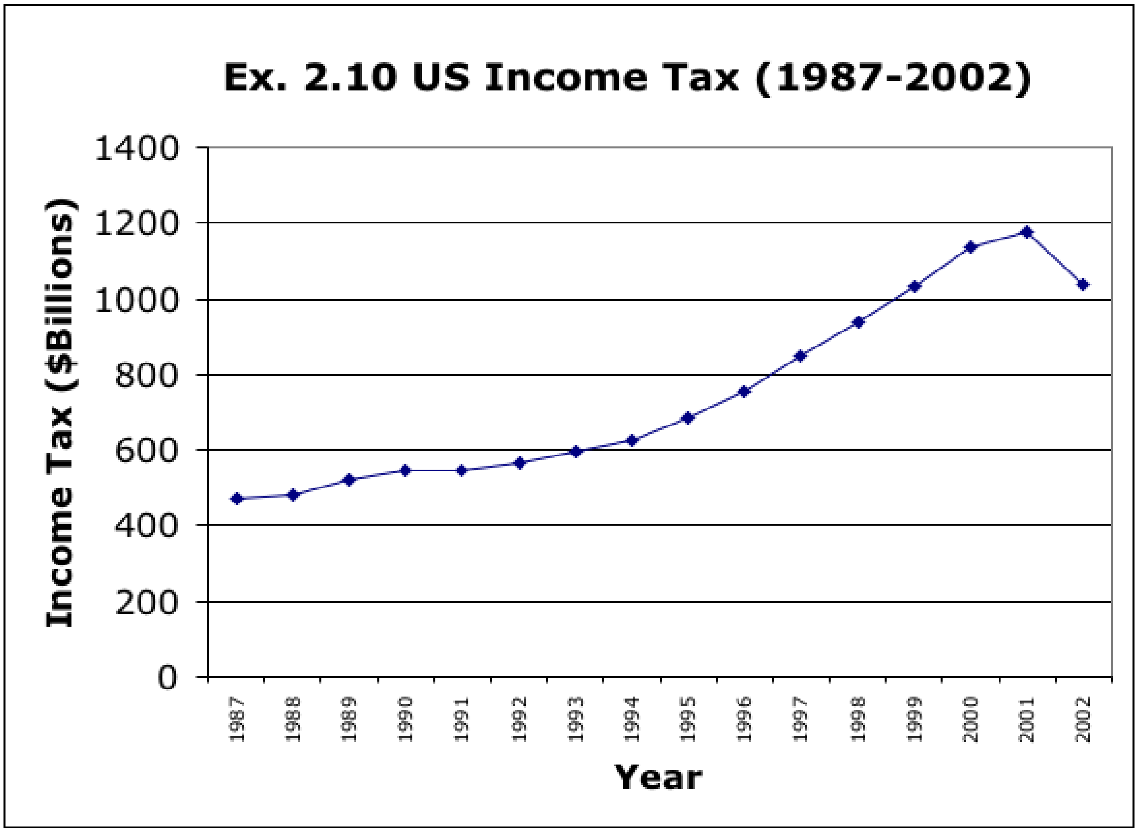
Observations measured at the *same* point in time are called **cross-sectional** data.

Observations measured at *successive* points in time are called **time-series** data.

An example is the closing price of a stock for a particular day versus over a number of days.

Time-series data graphed on a **line chart**, which plots the value of the variable on the vertical axis against the time periods on the horizontal axis.

Example: U.S. Income Tax



From 1987 to 1992, the tax was fairly flat. Starting 1993, there was a rapid increase in taxes until 2001. Finally, there was a downturn in 2002.

Summary...

	Interval Data	Nominal Data
Single Set of Data	Histogram, Ogive, or Stem-and-Leaf Display	Frequency and Relative Frequency Tables, Bar and Pie Charts
Relationship Between Two Variables	Scatter Diagram	Contingency Table, Bar Charts