Graphical and Tabular Descriptive Techniques

Introduction & Re-cap…

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

Descriptive statistics make use of graphical techniques and numerical techniques (such as averages) to summarize and present the data.

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

Definitions…

A **variable** **[Typically called a "random" variable since we do not know it's value until we observe it]** is some characteristic of a population or sample.

E.g. student grades, weight of a potato, # heads in 10 flips of a coin, etc.

Typically denoted with a capital letter: X, Y, Z…

The **values** of the variable are the range of possible values for a variable.

E.g. student marks (0..100)

Data are the ***observed values*** of a random variable.

E.g. student marks: {67, 74, 71, 83, 93, 55, 48}

We Deal with "2" Types of Data

Numerical/Quantitative Data [Real Numbers]:

* height of a randomly selected student
* weight of a quarter pounder meat patty
* temperature of coffee bought at McDonkals
* etc.

Qualitative/Categorical Data [Labels rather than numbers]:

* classification of a UTA student [F, S, J, Senior]
* favorite color
* the part of a new automobile that breaks first
* the reason you get mad at your spouse
* etc.

Quantitative/Numerical Data…

Real numbers, i.e. heights, weights, prices, etc.

Arithmetic operations can be performed on Quantitative Data, thus its meaningful to talk about 2*Height, or Price + $1, and so on.

TWO TYPES OF QUANTITATIVE DATA:

Interval Data: has no natural "0" such as temperature. 100 degrees is 50 degrees hotter than 50 degrees BUT not twice as hot.

Ratio Data: has a natural "0" such as weight. 100 pounds is 50 pounds heaver than 50 pounds AND is twice as heavy.

Normally will not be a factor in what you do unless you do something "less than smart" like claiming 100 degrees is twice as hot as 30 degrees.

Quantitative/Numerical Data…

ONE MORE THING…….

Quantitative Data is further broken down into

* Continuous Data – Data can be any real number within a given range. Normally measurement data [weights, times, volumes, etc]

* Discrete Data – Data can only be very specific values which we can list. Normally count data [# of firecrackers in a package of 100 that fail to pop, # of accidents on the UTA campus each week, etc]

What about the amount of money spent by each customer at WalMart?????????????

Qualitative/Categorical Data

Nominal Data [has no natural order to the values].

E.g. responses to questions about marital status:
Single = 1, Married = 2, Divorced = 3, Widowed = 4

Arithmetic operations don't make any sense (e.g. does Widowed ÷ 2 = Married?!)

Ordinal Data [values have a natural ***order***]:

E.g. College course rating system:

poor = 1, fair = 2, good = 3, very good = 4, excellent = 5

Has major impact on way you arrange tabular and graphical summaries.

Describe characteristics of the following random variables.

[Numerical/Categorical],[Continuous/Discrete]

[Ratio/Interval][Nominal/Ordinal]

Number of students who drop this statistics course.

Time student spends studying for their first statistics test.

The amount owed on a credit card (explain you answer)

Kind of pet you first owned.

Length of the longest local telephone call made by customers per month.

Title of employees working for a major corporation

Temperature of patients in hospital with the flu.

Graphical & Tabular Techniques for Nominal Data...

The only allowable calculation on nominal data is to count the frequency of each value of the variable.

We can summarize the data in a table that presents the categories and their counts called a ***frequency distribution.***

A ***relative frequency distribution*** lists the categories and the proportion with which each occurs.

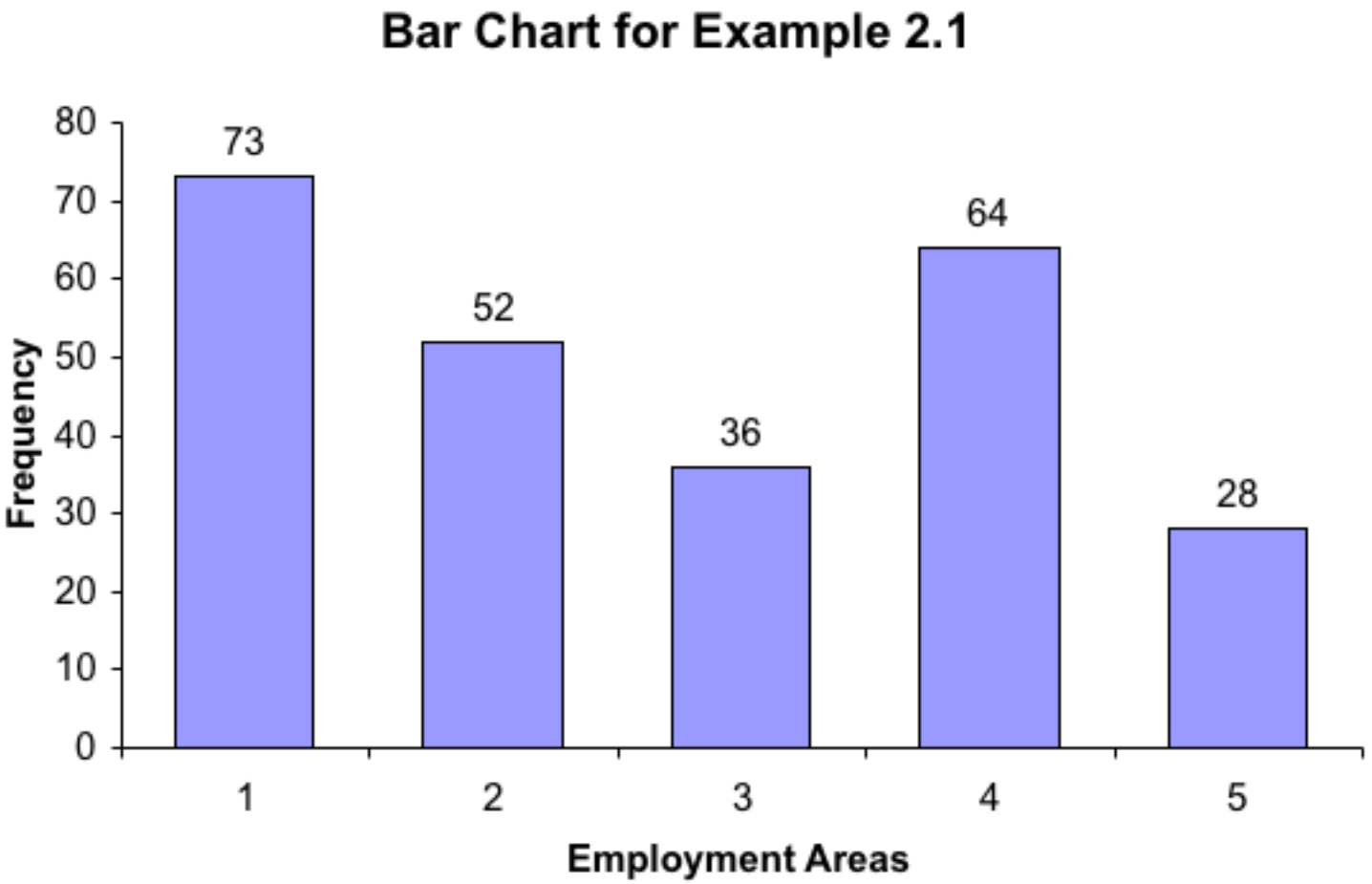
Since Nominal data has no order, if we arrange the outcomes from the most frequently occurring to the least frequently occurring, we call this a "pareto chart"

Nominal Data (Tabular Summary) -

Table 2.1 Frequency and Relative Frequency Distributions for Example 2.1

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

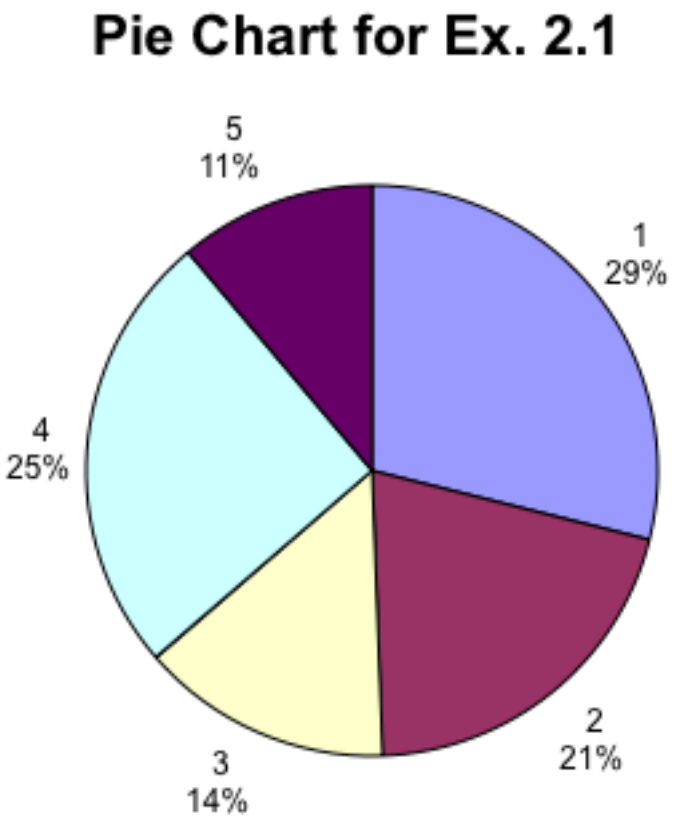
Nominal Data (Frequency)



Bar Charts are often used to display ***frequencies…***

Is there a better way to order these? Would Bar Chart look different if we plotted "relative frequency" rather than "frequency"?

Nominal Data (Relative Frequency)

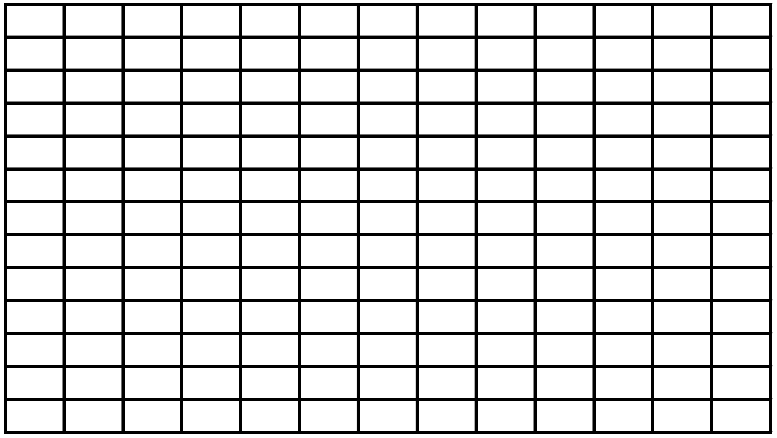


Pie Charts show ***relative frequencies…***

Students Work

6. Historically, grades for this course are shown below. When answering the questions, be neat and show "all" relevant information on each figure. Construct a bar chart showing both frequency and relative frequency on the same chart. Show all relevant information.

GRADE	FREQUENCY
A	210
B	130
C	100
D	14
F	2
W	44
TOTAL =	500



Graphical Techniques for Quantitative Data

There are several graphical methods that are used when the data are ***quantitative*** (i.e. numeric, non-categorical).

The most important of these graphical methods is the ***histogram***.

The histogram is not only a powerful graphical technique used to ***summarize*** interval data, but it is also used to help ***explain*** probabilities.

Building a Histogram…

1) Collect the Data ✓ (Example 2.4) 200 Long distance telephone bills.
2) Create a frequency distribution for the data…

   How?

   a) Determine the number of ***classes*** to use…

      How? [rule of thumb, not cast in concrete]

      Refer to Table 2.6

With 200 observations, we should have between 7 & 10 classes…

Table 2.6 Approximate Number of Classes in Frequency Distributions

Number of Observations	Number of Classes
Less than 50	5 - 7
50 - 200	7 - 9
200 - 500	9 - 10
500 - 1,000	10 - 11
1,000 - 5,000	11 - 13
5,000 - 50,000	13 - 17
More than 50,000	17 - 20

Alternative, we could use Sturges' formula:
Number of class intervals = **1 + 3.3 log (n)**

Building a Histogram…

1) Collect the Data ✓
2) Create a frequency distribution for the data…
   a) Determine the number of ***classes*** to use. [8]
   b) Determine how large to make each class…

Look at the ***range*** of the data, that is,

Range = Largest Observation – Smallest Observation

Range = $119.63 – $0 = $119.63

Then each class width becomes:

Range ÷ (# classes) = 119.63 ÷ 8 ≈ 15 [round up]

Building a Histogram…

1) Collect the Data ✓
2) Create a frequency distribution for the data…

   How?

   a) Determine the number of ***classes*** to use. [8] ✓

   b) Determine how large to make each class. [15] ✓

   c) Place the data into each class…

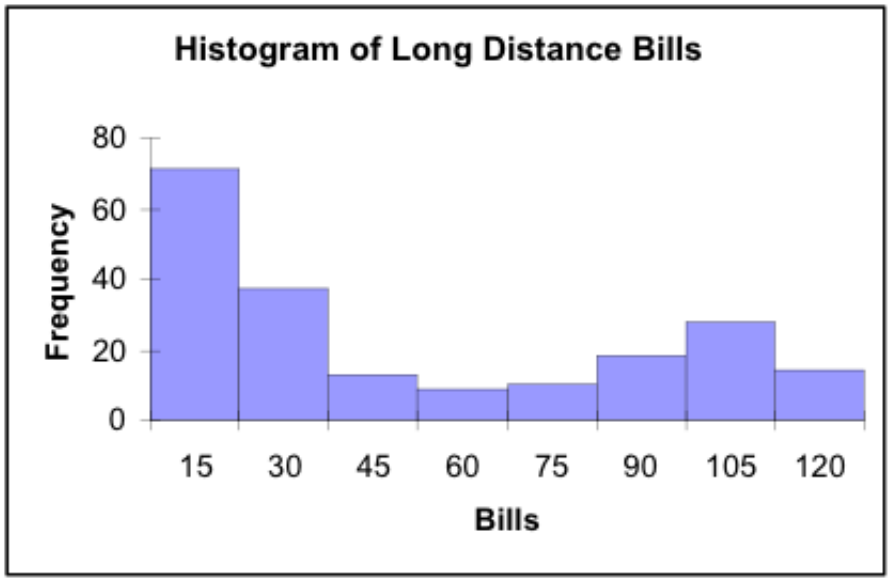
   - ***each item can only belong to one class;***
   - ***classes contain observations greater than their lower limits and less than or equal to their upper limits.***

Building a Histogram…

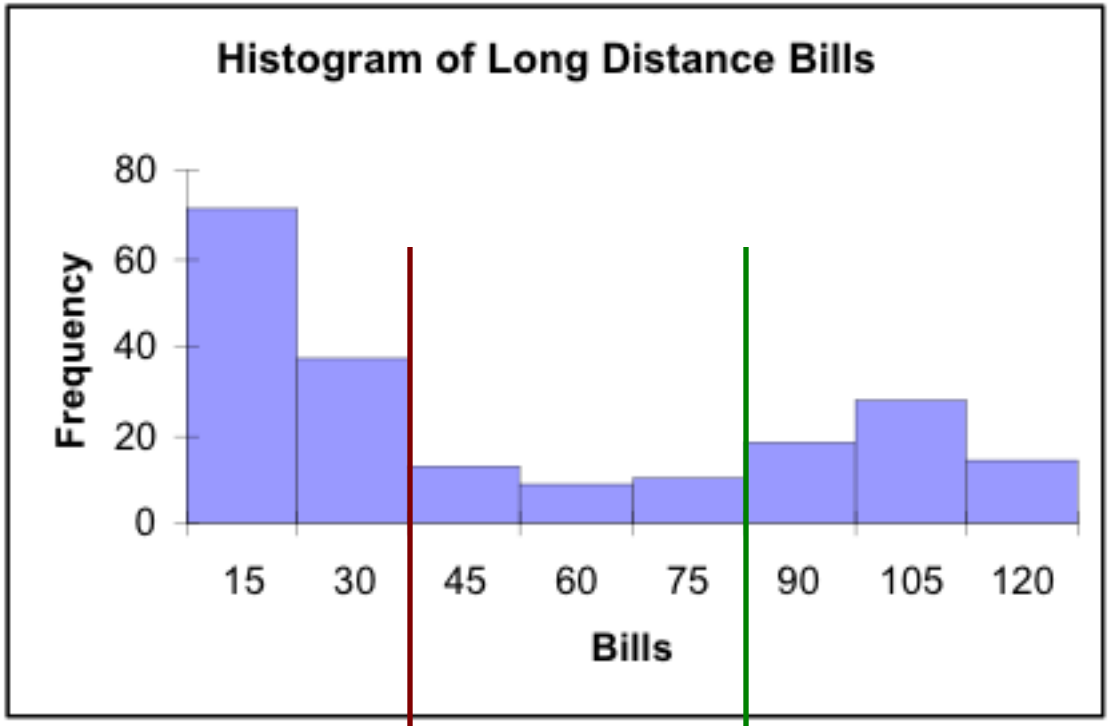
1) Collect the Data ✓ - [200 telephone bills]
2) Create a frequency [relative?] distribution for the data. ✓
3) Draw the Histogram. ✓

Table 2.5 Frequency Distribution of the Long-Distance Bills in Example 2.4

Class Limits	Frequency
0 to 15	71
15 to 30	37
30 to 45	13
45 to 60	9
60 to 75	10
75 to 90	18
90 to 105	28
105 to 120	14
Total	200



Interpret…



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

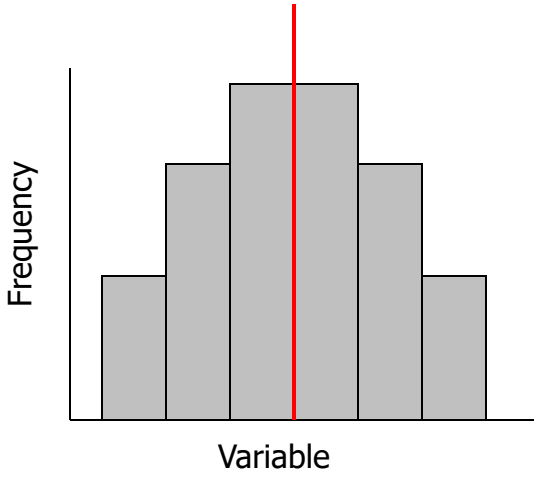
There are only a few telephone bills in the middle range.

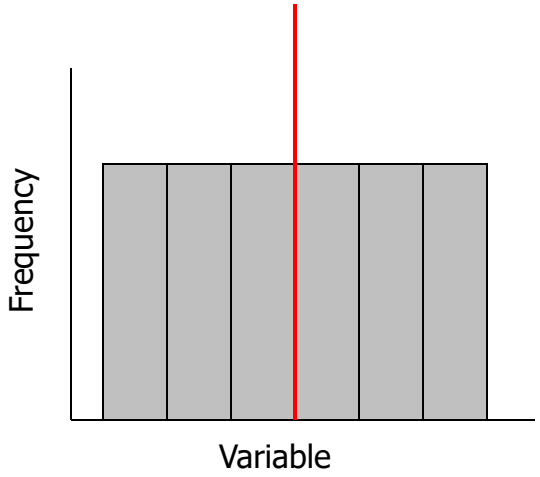
(18+28+14=60)÷200 = 30%
i.e. nearly a third of the phone bills are greater than $75

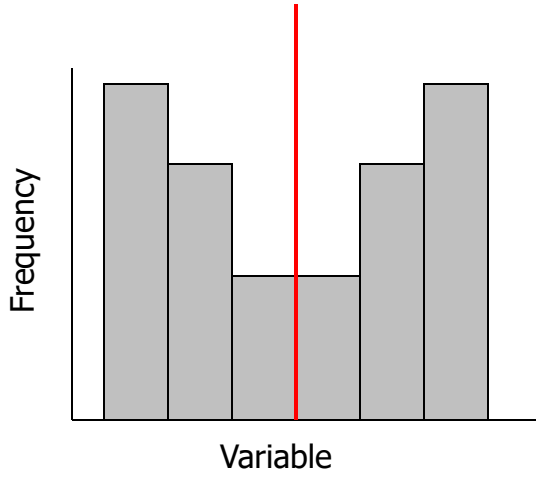
Shapes of Histograms...

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:



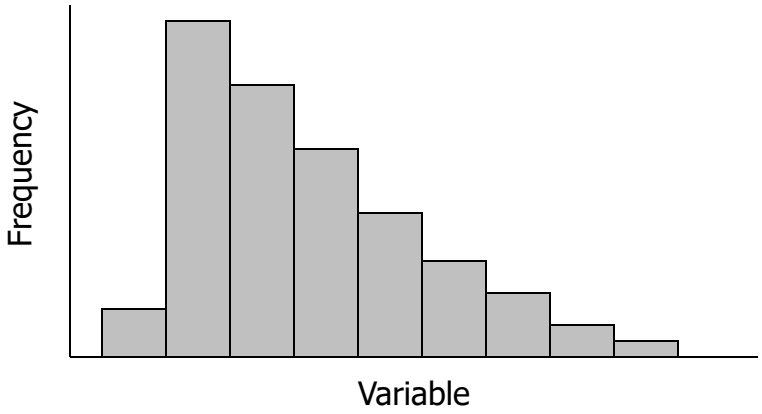




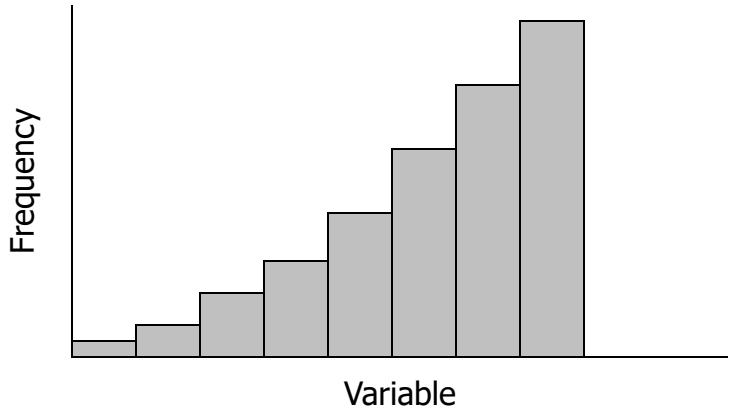
Shapes of Histograms…

Skewness

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



Positively Skewed

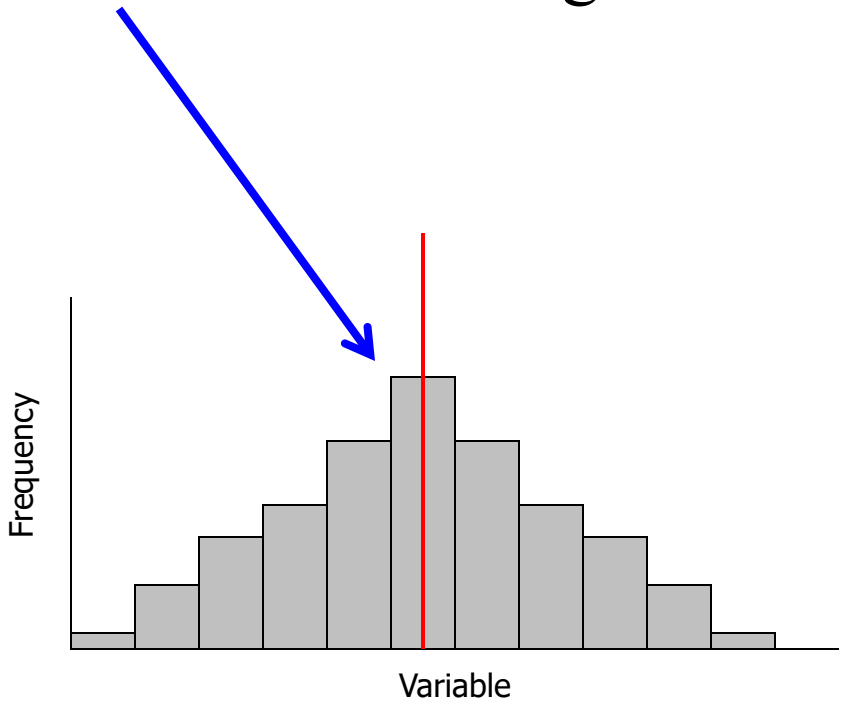


Negatively Skewed

Shapes of Histograms...

Bell Shape

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



Bell Shaped

Many statistical techniques require that the population be bell shaped.

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

Relative Frequencies…

For example, we had 71 observations in our first class (telephone bills from \$0.00 to \$15.00). Thus, the relative frequency for this class is 71 ÷ 200 (the total # of phone bills) = 0.355 (or 35.5%)

Table 2.7 Relative Frequency Distribution for Example 2.4

Class Limits	Relative Frequency
0 to 15	71/200 = .355
15 to 30	37/200 = .185
30 to 45	13/200 = .065
45 to 60	9/200 = .045
60 to 75	10/200 = .050
75 to 90	18/200 = .090
90 to 105	28/200 = .140
105 to 120	14/200 = .070
Total	200/200 = 1.0

Cumulative Relative Frequencies…

Table 2.8 Cumulative Relative Frequency Distribution for Example 2.4

Class Limits	Relative Frequency	Cumulative Relative Frequency
0 to 15	71/200 = .355	71/200 = .355
15 to 30	37/200 = .185	108/200 = .540
30 to 45	13/200 = .065	121/200 = .605
45 to 60	9/200 = .045	130/200 = .650
60 to 75	10/200 = .05	140/200 = .700
75 to 90	18/200 = .09	158/200 = .790
90 to 105	28/200 = .14	186/200 = .930
105 to 120	14/200 = .07	200/200 = 1.00

first class…

next class: .355+.185=.540

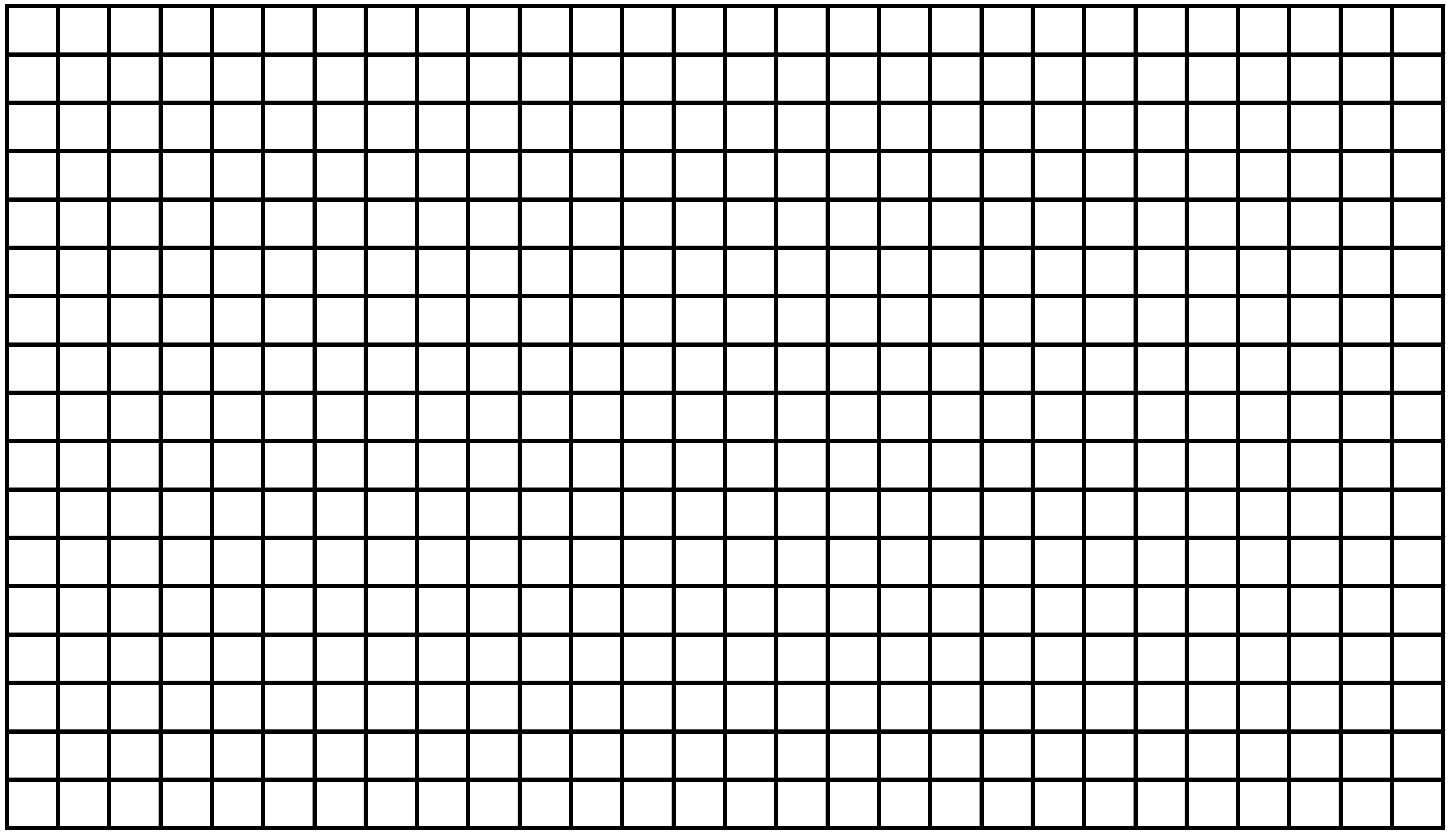
last class: .930+.070=1.00

STUDENTS WORK. For the following data,

a. construct a frequency and relative frequency histogram on one graph (use 5 cells).

b. does this data appear to be normally distributed (MOUNDED)? Explain.

18.13	18.17	18.19	18.21	18.23	18.30	18.51	18.63	18.66	18.73
18.80	18.84	18.92	19.24	19.27	19.30	19.31	19.32	19.50	19.63
19.63	19.93	20.31	20.32	20.37	20.50	20.69	20.97	21.10	21.12
21.40	21.44	21.50	21.57	21.71	21.72	21.76	21.92	21.93	21.98



Contingency Table…

In Example 2.8, 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).

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

Contingency Table... Look at Column Relative Frequencies

Interpretation: The relative frequencies in the columns 2 & 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

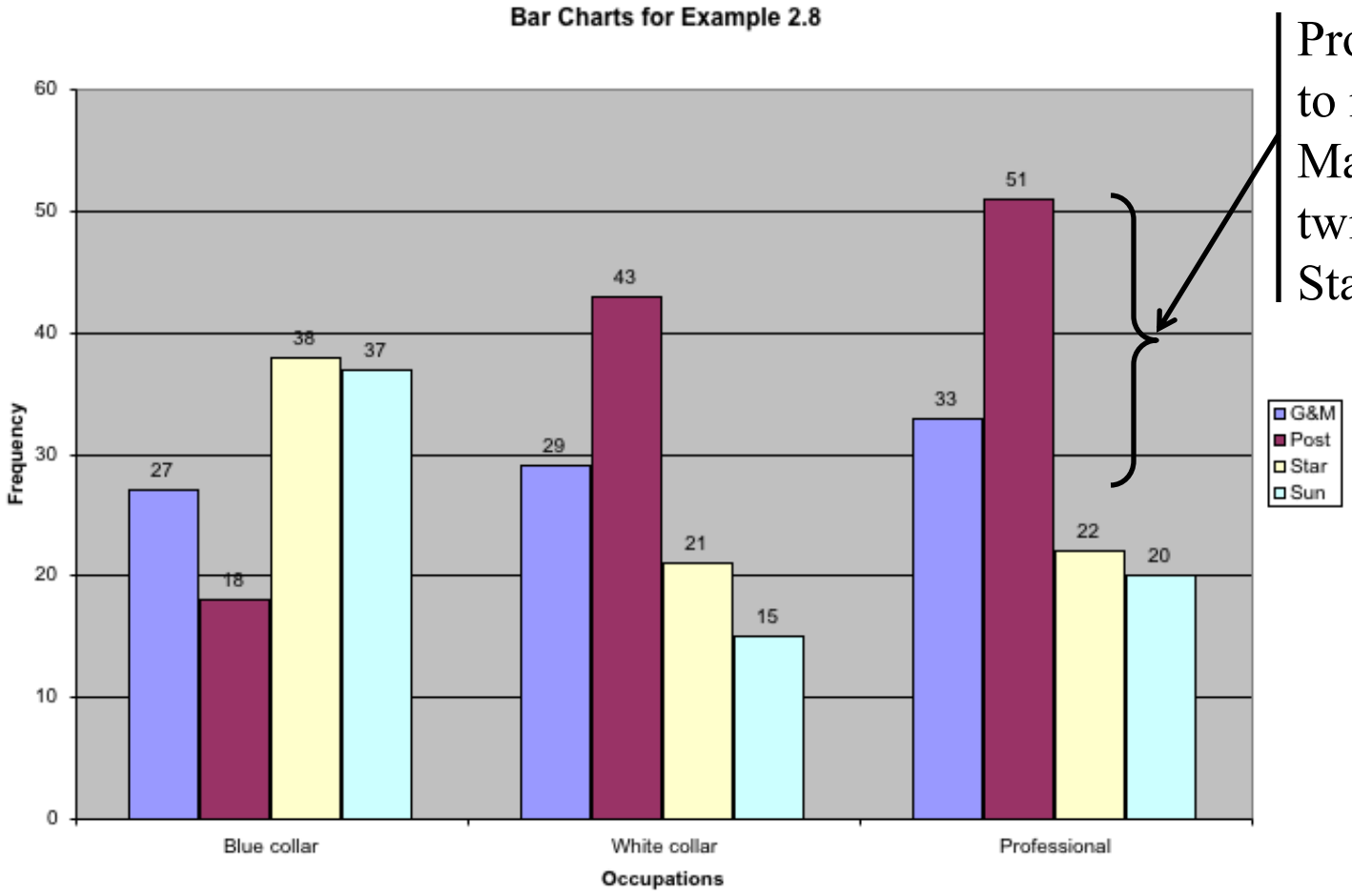
similar

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. How do you use this knowledge for marketing strategies.

Graphing the Relationship Between Two Nominal Variables…

Use the data from the contingency table to create bar charts…



Professionals tend to read the Globe & Mail more than twice as often as the Star or Sun…

Students Work – Grade Distributions

	Fresh	Soph	Jr	Senior	
A	17	28	28	30	103
B	21	23	29	12	85
C	35	26	39	11	111
D	31	36	25	10	102
F	40	26	29	4	99
	144	139	150	67	500

	Fresh	Soph	Jr	Senior	
A					
B					
C					
D					
F					

Scatter Diagram…

Example 2.9 A real estate agent wanted to know to what extent the selling price of a home is related to its size…

Collect the data ✓

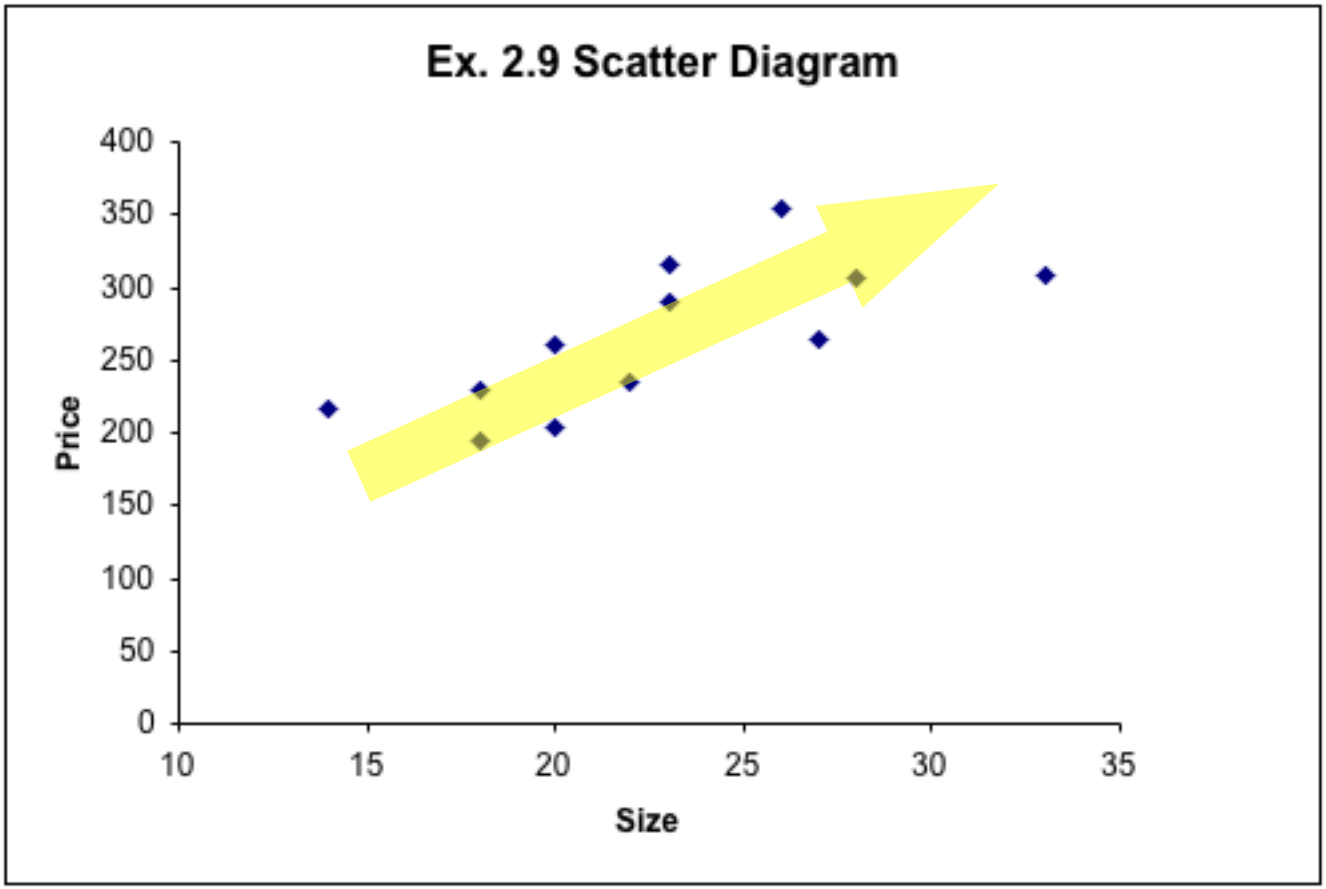
1) Determine the independent variable (X – house size) and the dependent variable (Y – selling price) ✓
2) Use Excel to create a "scatter diagram"…

Use any time you are showing the relationship between 2 variables.

* Your weight 1st of month last 10 years
* Relationship between people's weight and height
* Relationship between # of calories eaten and weight gain/loss

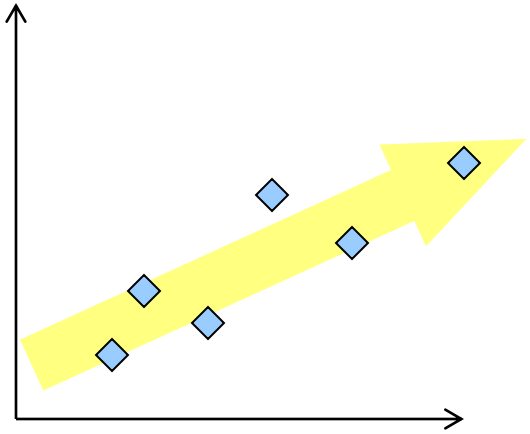
Scatter Diagram…

It appears that in fact there is a relationship, that is, the greater the house size the greater the selling price…

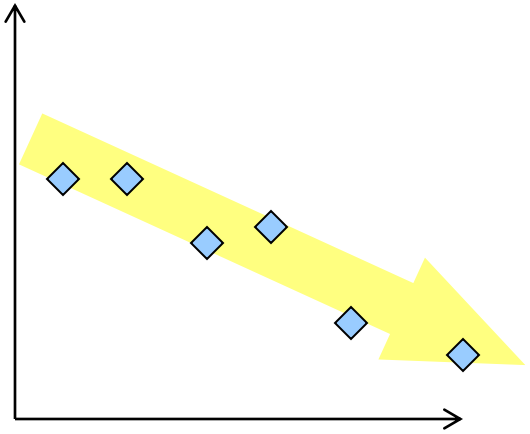


Patterns of Scatter Diagrams...

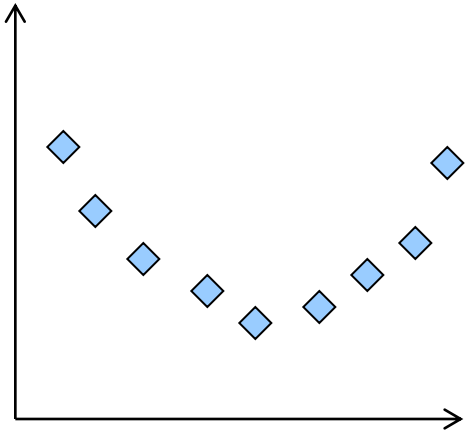
Linearity and Direction are two concepts we are interested in



Positive Linear Relationship



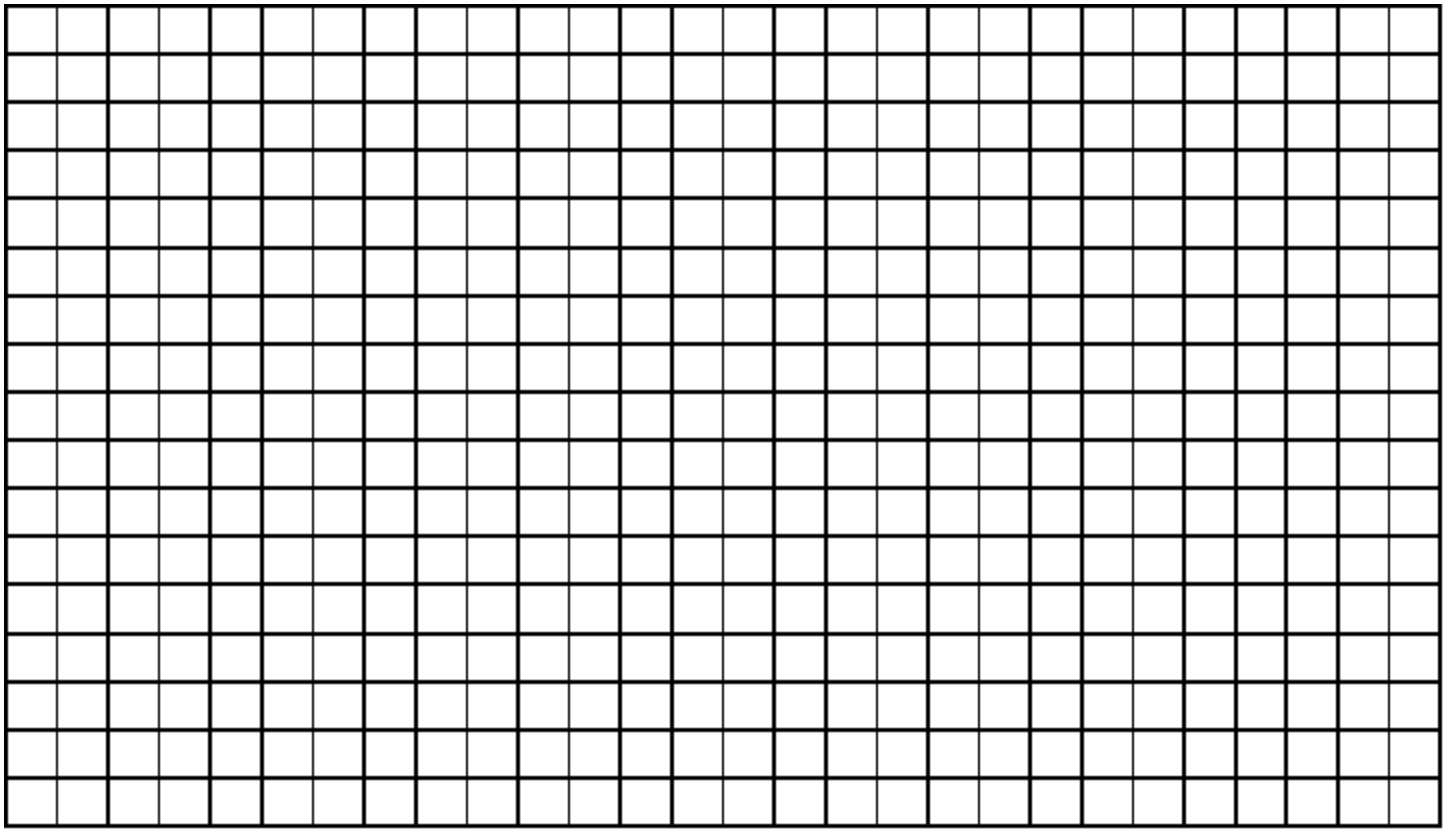
Negative Linear Relationship



Weak or Non-Linear Relationship

Students work: Data from the last statistics course was collected on the number of hours each student studied and their grade on the final is shown below. Plot a scatter diagram for this data.

Number Hours Studied	30	22	40
Grade on Final	85	70	65



Time Series Data…

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

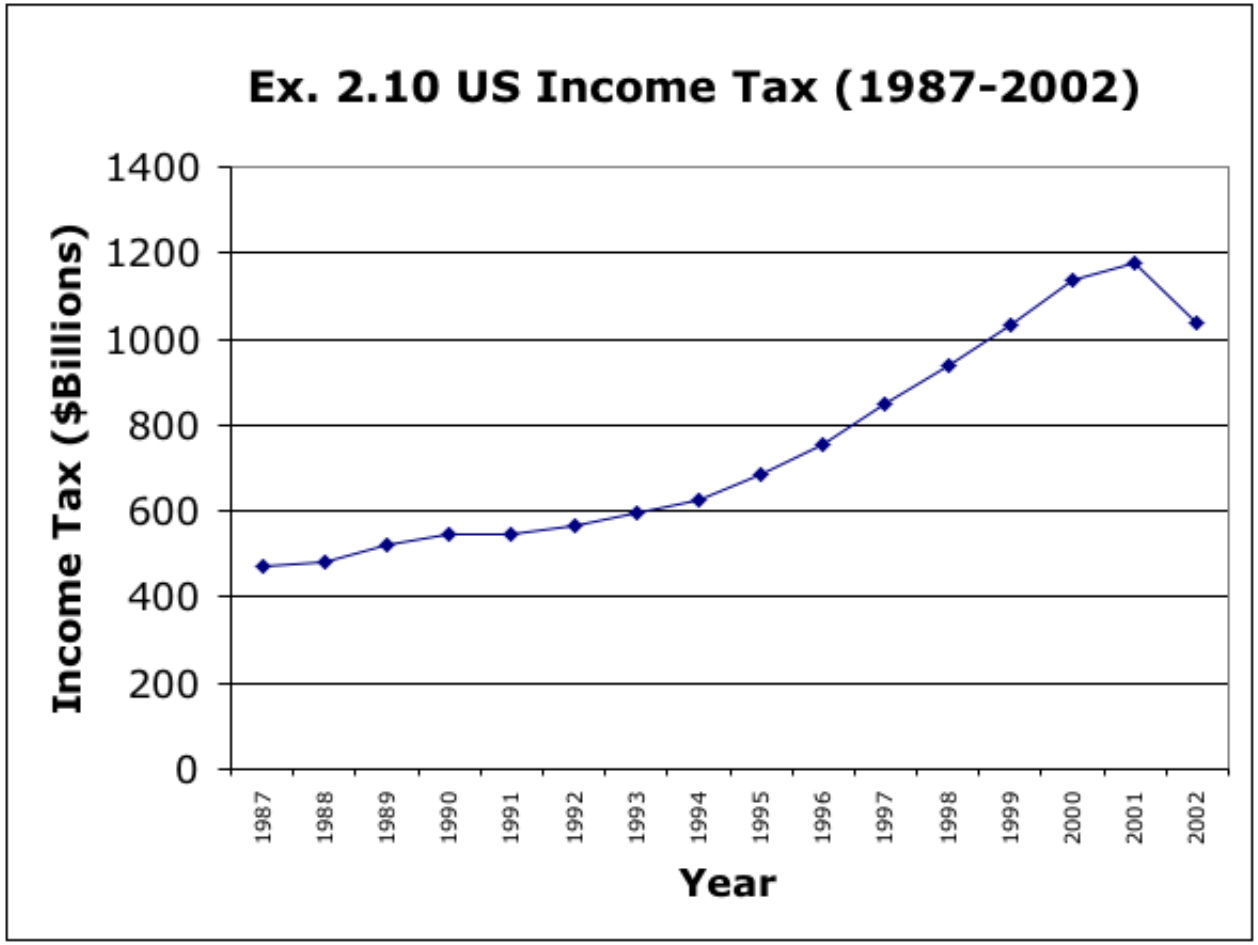
Observations measured at successive points in time are called ***time-series*** data.

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.

In general, you would not be plotting a histogram from time-series data.

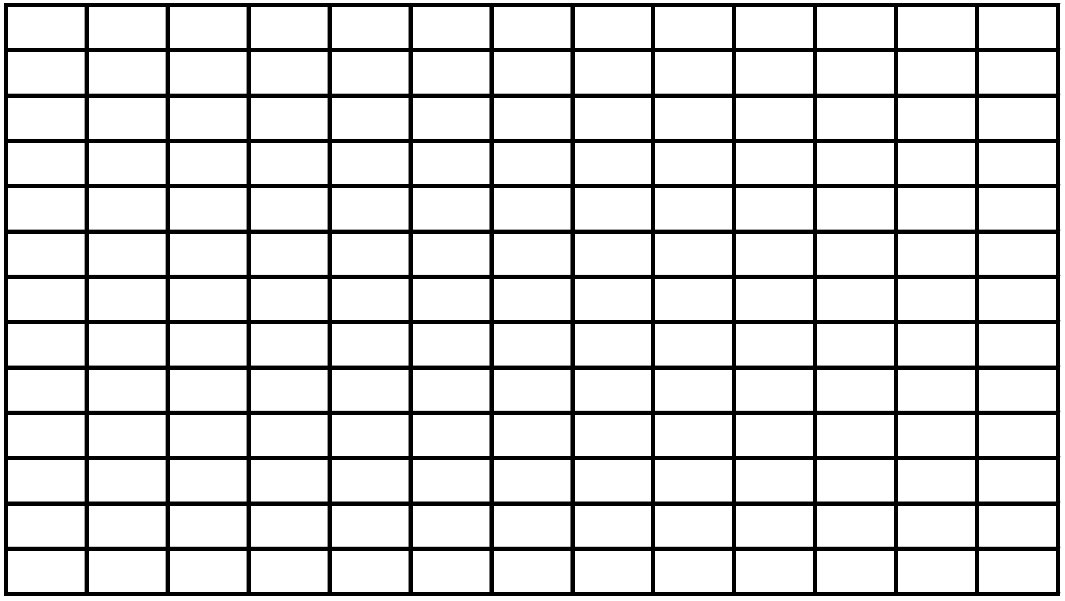
Line Chart…

From Example 2.10, plot the total amounts of U.S. income tax for the years 1987 to 2002…



Students work: Weight of Grandson's pig since purchased 4 weeks ago. Describe pig's weight graphically.

Bought	0	1	2	3	4
Weight	180	184	187	192	196



Problems: Types of Data

For the following data, identify whether or not they are 1. **categorical** [**nominal** or **ordinal**], or 2. **numerical [interval/ratio]** [**discrete** or **continuous**] Give examples of possible values for each random variable. [Example: number of children living in a given home – "interval data [discrete], (0, 1, 2, 3, …)"

*marital status

*number of students who drop this statistics course.

*time student spends studying for their first statistics test.

*the weight loss over the first week of a "fad" diet

*the amount owed on a credit card (explain you answer)

*the part on a new automobile that breaks during the first year of ownership

*the rank of a military officer

Problem: Bar Charts [Categorical Data]

Over the last year data was collected on the type of infections detected in the hospital. For convenience these were coded A, B, C, and D. The raw data is shown below.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
A	A	A	A	A	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B
B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B
C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
C	C	C	C	C	C	C	C	C	C	D	D	D	D	D	D	D	D	D	D

- construct a frequency and relative frequency table.
- construct a bar chart

Problems: Histogram

A random sample of grade point averages for 20 students resulted in the following data.

[1.84, 1.94, 2.15, 2.22, 2.37, 2.76, 2.82, 2.89, 2.94, 2.96. 2.99, 3.05, 3.11, 3.24, 3.26, 3,29, 3.33, 3.54, 3.58, 3.79]

*use 5 class intervals (cells) to describe this data in a frequency table, and relative frequency table.

*draw a frequency histogram. (show all relevant information on graph including relative frequency)

Problem: Scatter Diagram

Data from the last statistics course was collected on the number of hours each student studied and their grade on the final resulting in

Number Hours Studied	30	22	40
Grade on Final	85	70	65

Plot a scatter diagram for this data showing the relationship between the two variables. Show all relevant information on the graph (this means it should look just like one you would put in a report prepared for President O.).