Chap 1 Notes: Exploring Data by Graphical Methods

J. Harner A. Billings

Department of Statistics
West Virginia University

Stat 211 Fall 2007

Outline

What is Statistics?

Statistics The science of collecting, organizing, and analyzing data for the purpose of estimation and making inferences.

Data Values which arise from observing characteristics on a selected group (sample) of individuals.

Variables Characteristics or attributes observed on each individual.

Population The group from which the individuals are selected.

Types of Data

Variables can be classified into one of two types:

Numerical: Variables whose values represent quantities.

Categorical: Variables whose values are non-numeric.

Examples:

- Age
- Height
- Weight
- Gender

Numeric Variables

Numeric variables can be further classified as:

Discrete: Variables which usually arise by counting.

Continuous: Variables which usually arise by measuring.

Values of discrete variables are generally natural numbers, i.e., non-negative integers represented by $\{0, 1, 2, \ldots\}$.

Values of continuous variables are real numbers (technically represented by decimals) contained in a range or interval.

Numeric Variable Examples

Examples: Discrete variables

- Number of cars in a parking lot
- Student credit hours
- Number of books you own

Examples: Continuous variables

- Height of a person
- Amount of time spent studying
- Weight of an apple

Question: Is the variable “Age” discrete or continuous?

Categorical Variables

Values of categorical variables are non-numeric.

Categorical variables can be classified as:

Categorical: Values are unordered.

Ordered: Values possess a natural ordering.

An ordered categorical variable is also said to be ranked.

Categorical Variable Examples

Examples: Categorical variables

- Gender
- Blood Type
- Zip code

Examples: Ordered variables

- Class rank
- Course grade
- Taste test (bad $\cdots$ good)

Question: Are “phone number” and “student ID” categorical?

Branches of Statistics

The two major branches of statistics are:

Descriptive Statistics Use graphical displays and numeric summaries to represent data.

Inferential Statistics Use analytic methods and the theory of probability to draw conclusions or make decisions.

Displaying Small Sets of Numbers

The following elementary plot types are suitable for small data sets:

1. Dotplots: a dot represents each value of a numerical variable.
2. Stem-and-leaf plot (stemplot): each value is represented by a stem and a leaf.

Dotplots

The dotplot shows how the numerical variable values are distributed for small data sets.

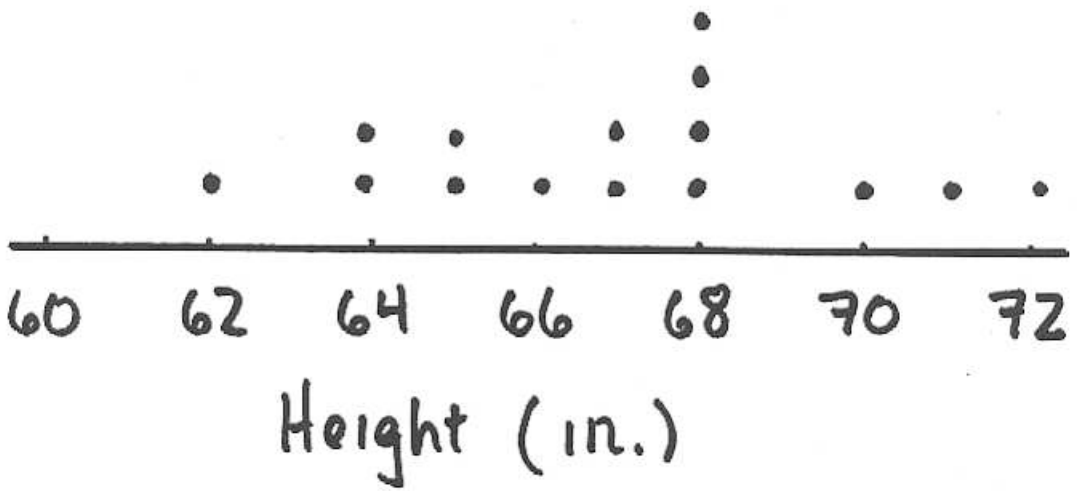
For this plot to be useful, the values should:

- have repeated values;
- be concentrated within a relatively small interval.

Dotplots: Example

Height (in.) of students

62	71	65	68	64
72	66	68	70	67
67	68	64	65	68



Stem-and-Leaf Plots (Stemplots)

Represents the data using the actual digits that make up the data.

- The leading digit(s) becomes the stem.
- The trailing digit(s) comprise the leaf.

Stemplot: Example

Math 126 Exam Grades

76	74	82	96	66	76
93	86	84	62	82	75
58	71	73	79	65	80

Stem (tens)	Leaf (ones)
5	8
6	6 2 5
7	6 4 6 5 1 3 9
8	2 6 4 2 0
9	6 3

Stemplot: Example

Data

1.3 2.4 1.7 3.2 5.6

Stem (ones)	Leaf (tenths)
1	3 7
2	4
3	2
4	
5	6

Back-to-Back Stem and Leaf

Tables 1.6–1.8 (p. 21 of the text)

Outlier Example

Outlier: an observation whose value is unusual or extreme.

Speed of cars on High St.

15	21	13	18	23	19
25	16	71	22	14	21

Stem (tens)	Leaf (ones)
1*	3 4
1·	5 8 9 6
2*	1 3 2 1
2·	5
HI	71

71 is an outlier.

Grouped Frequency Table

Frequencies are tabulated for each value of the categorical variable.

Example: Pet Ownership

Value	Frequency
Pet owner	9
Non-pet owner	7

Grouped Frequency Table: Example

Stat 211 Spring 2006 grade distribution:

Grade	Frequency
A	23
B	29
C	22
D	7
F	12
W	23

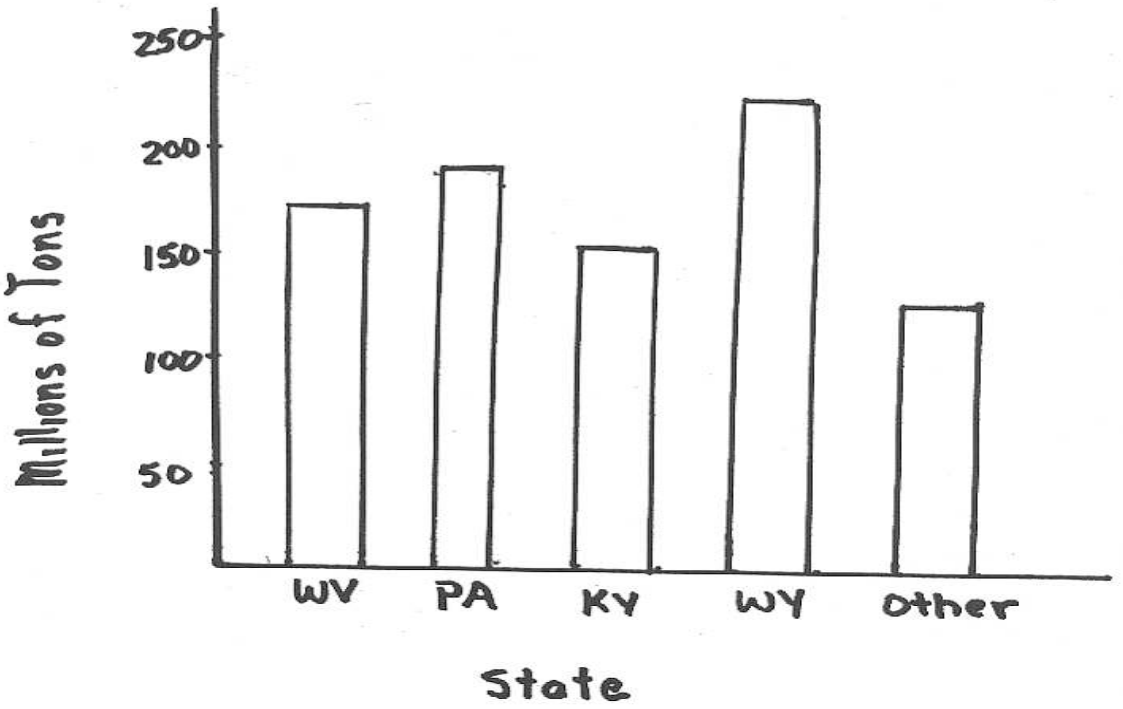
Bar Chart

A bar chart (graph) represents a categorical variable by showing the frequency of each category as proportionally-sized rectangles.

Bar Chart: Example

U.S. Coal Production by State

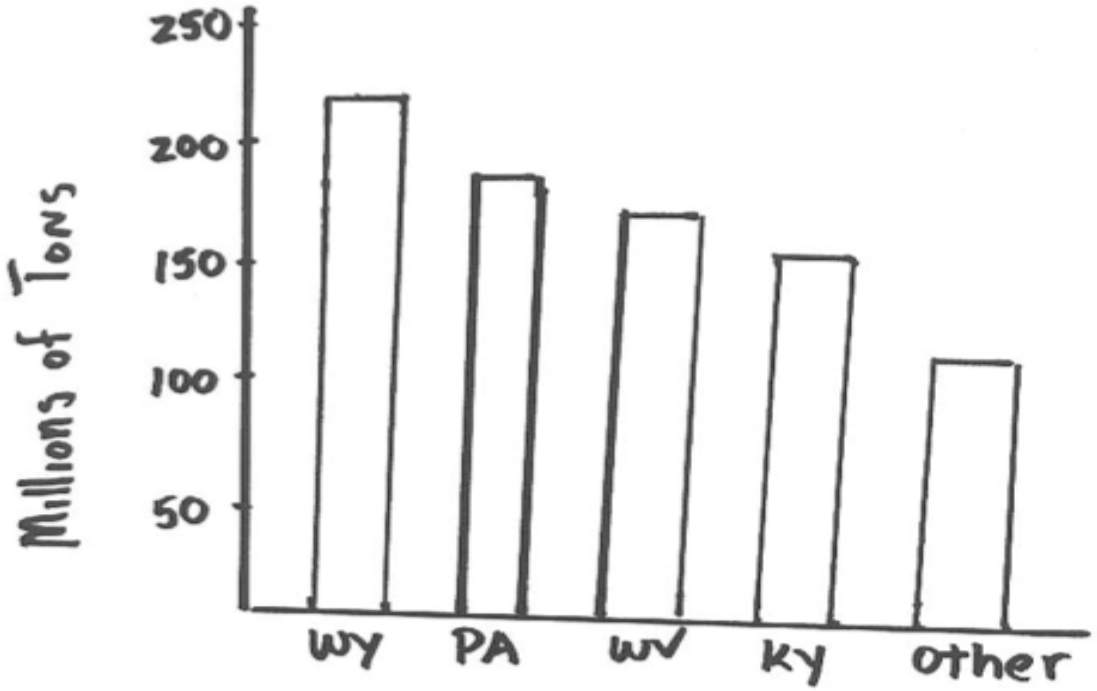
State	Millions of Tons
WV	172.0
PA	189.2
KY	154.8
WY	223.6
Other	120.4



Pareto Chart

A Pareto chart is a bar chart with the bars arranged from the tallest to shortest.

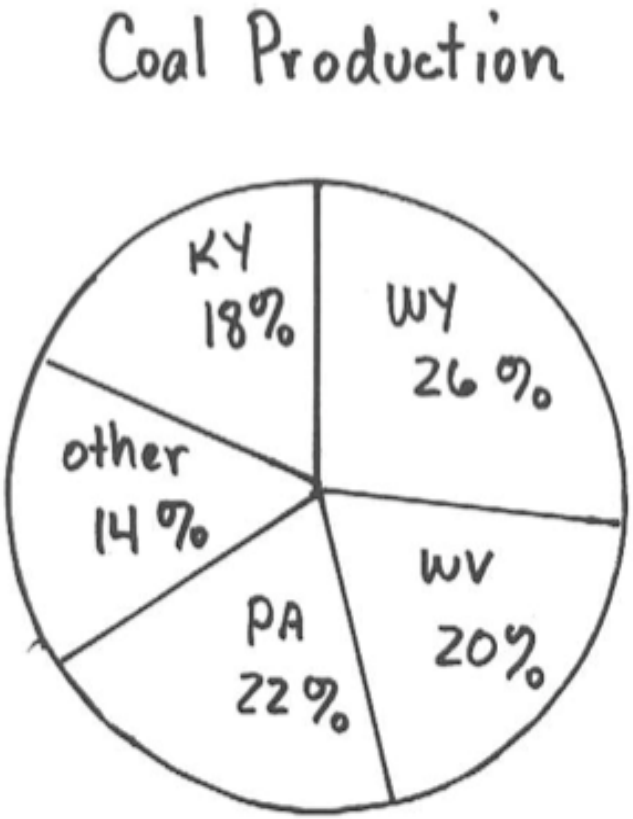
Example: US Coal Production



Pie Chart

A pie chart shows the amount of data that belongs to each category as a proportional part of a circle.

Example: US Coal Production



Grouped Frequency Distribution

A grouped frequency distribution is a list (or table) which pairs ranges of values of a numerical variable with their frequencies (counts). Each range is called a class.

Rules for constructing a grouped frequency distribution:

1. Each class should be the same width, unless there is a good reason for groups of unequal width.
2. Classes should not overlap.
3. Each observation should fall into one, and only one, class.
4. Between 3 and 15 classes should be used.

Grouped Frequency Distribution: Example

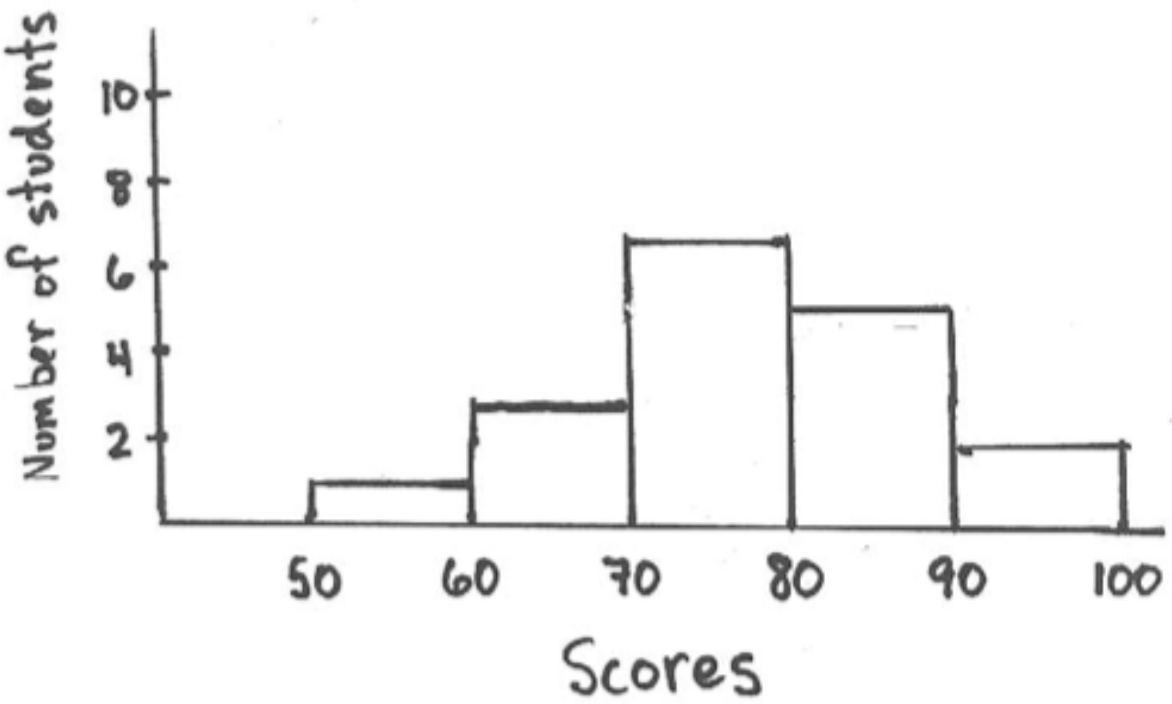
Math 126 Exam Scores

Class	Frequency
50–59	1
60–69	3
70–79	7
80–89	5
90–100	2

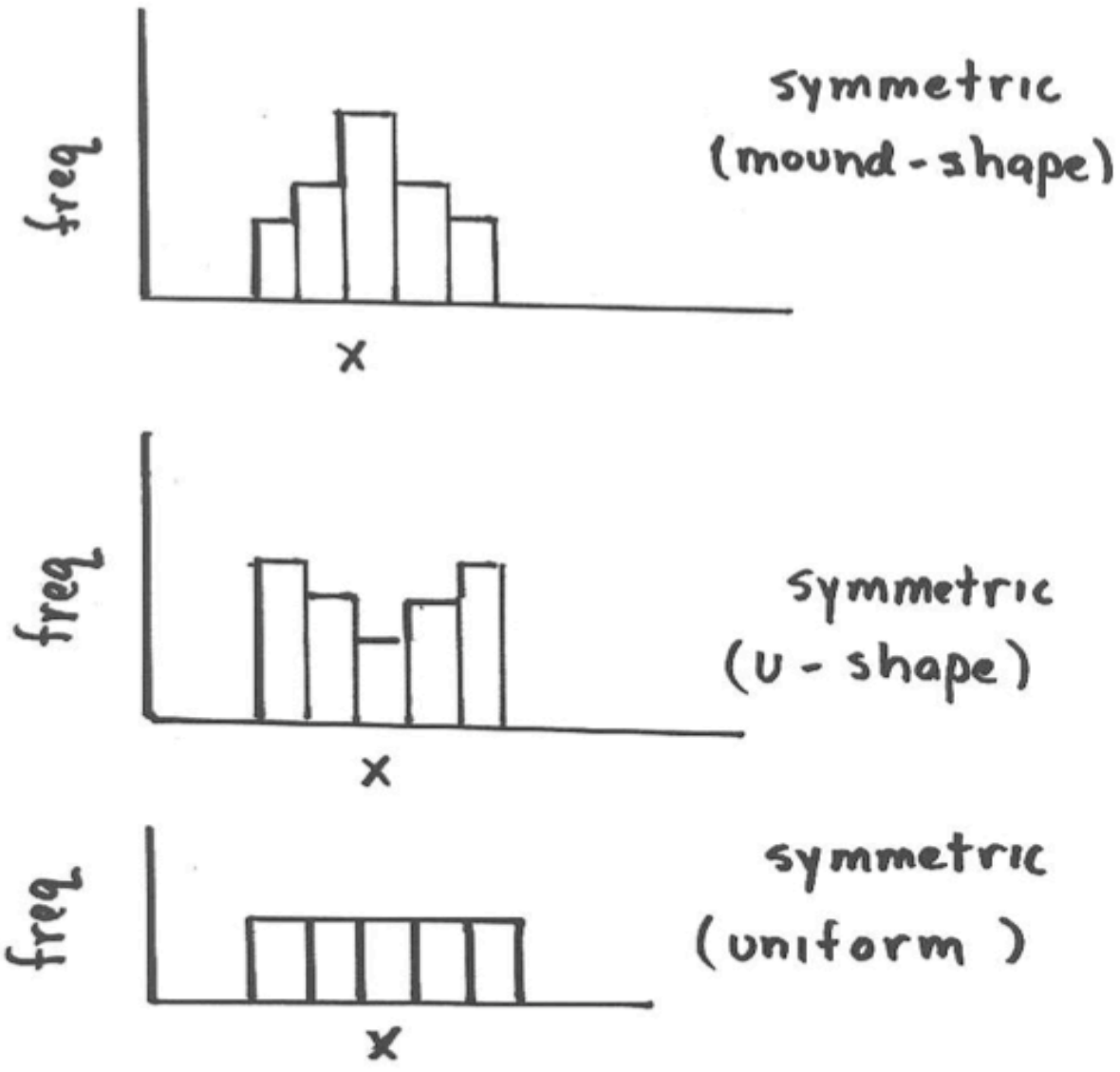
Frequency Histogram

A frequency histogram is a graphical representation of a frequency distribution.

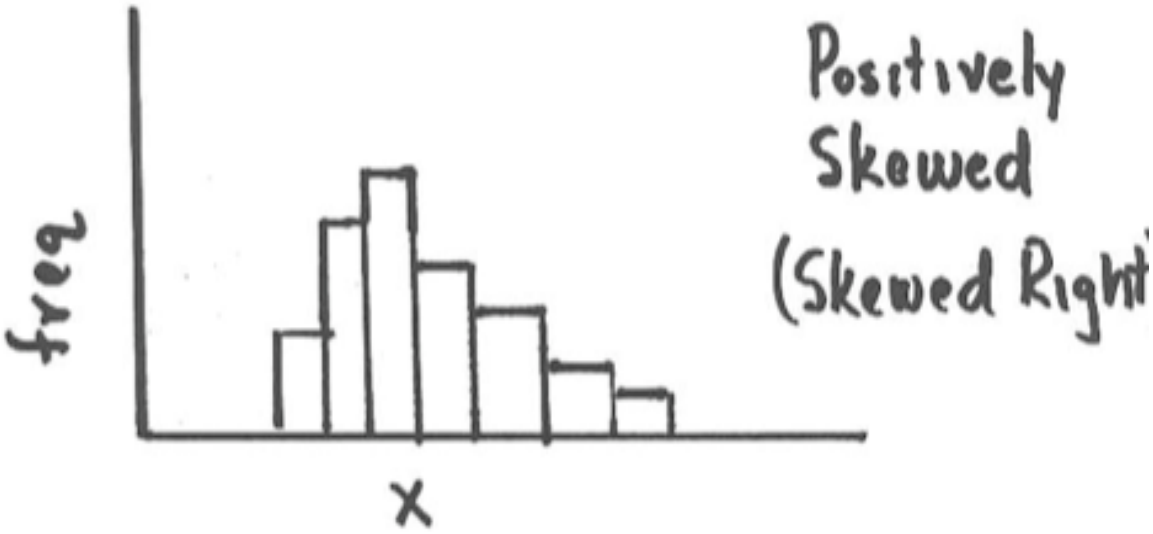
Example: Math 126 Exam Scores (continued)

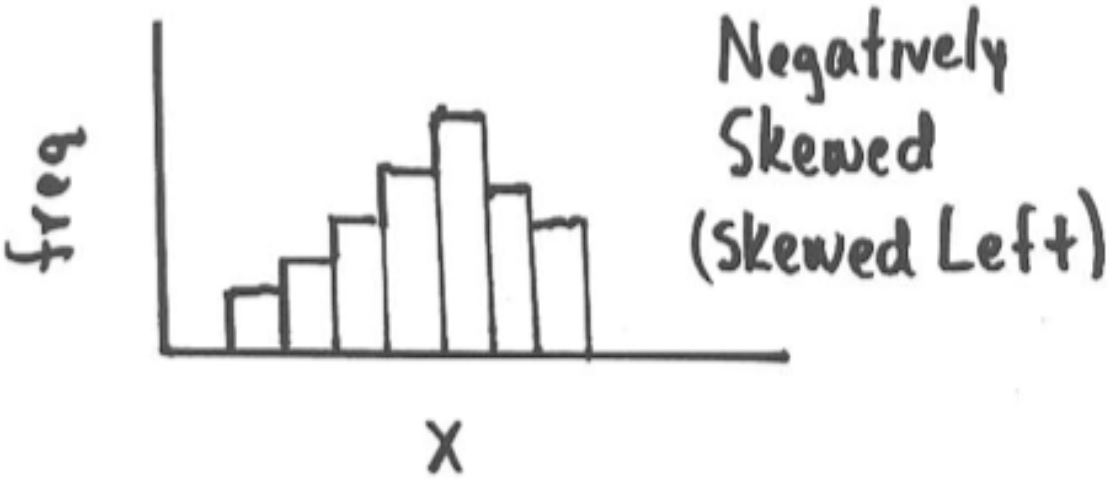


Shape of a Distribution: Symmetry

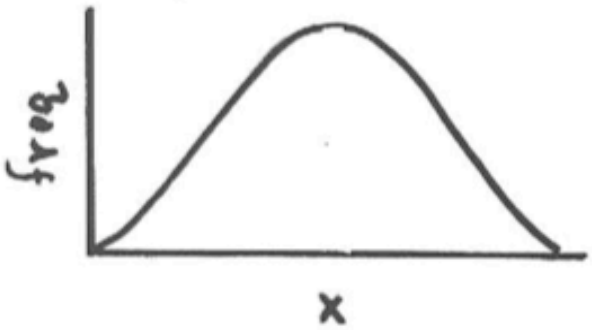


Shape of a Distribution: Skewness

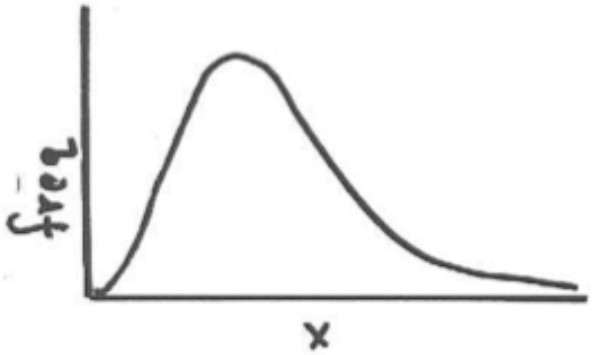




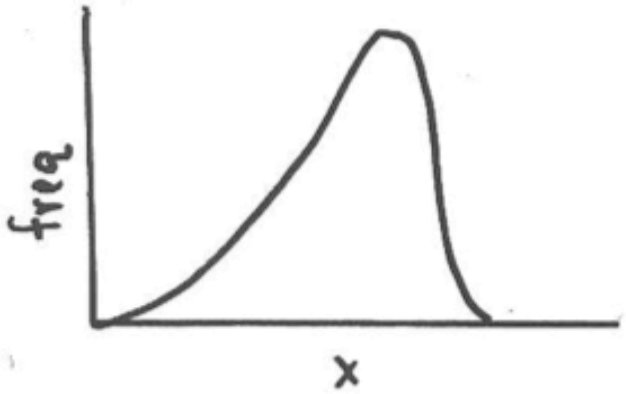
Shape of a Distribution: Continuous



Symmetric



Positively Skewed



Negatively Skewed

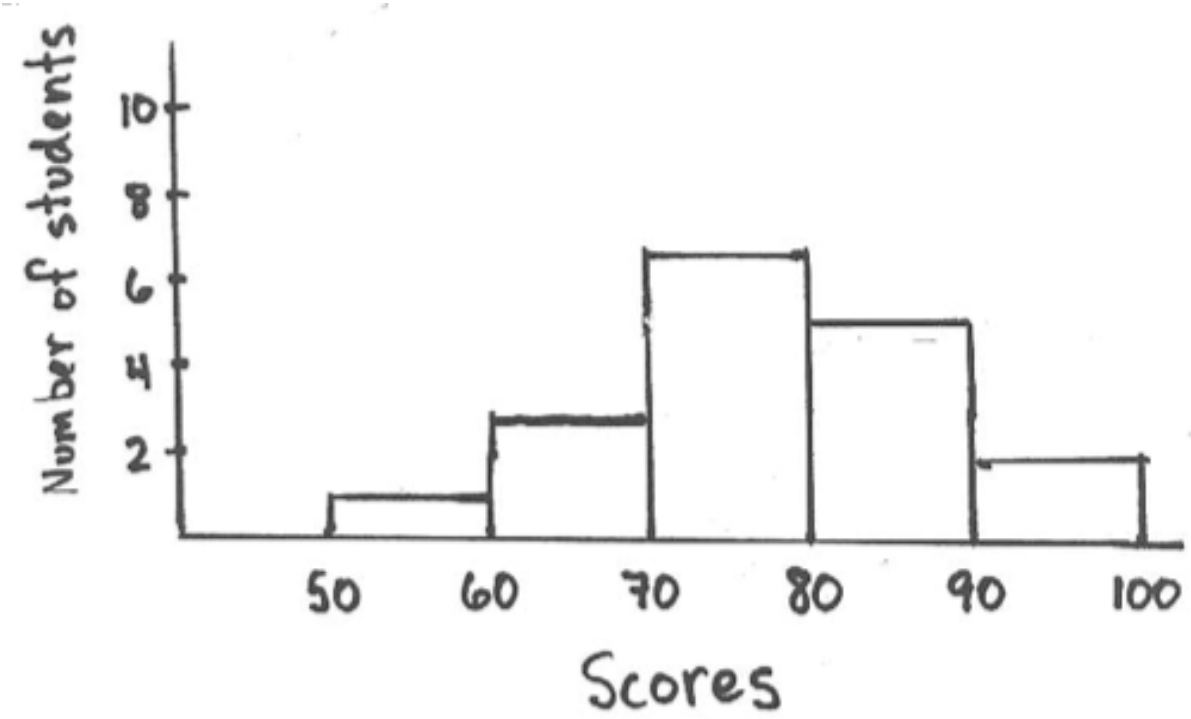
Relative Frequency Histogram

A relative frequency histogram is a histogram in which the vertical axis represents percentages or proportions.

Relative Frequency Histogram: Example

Math 126 Exam Scores

Class	Freq	Rel Freq	Percent
50–59	1	0.056	5.6
60–69	3	0.167	16.7
70–79	7	0.389	38.9
80–89	5	0.278	27.8
90–100	2	0.111	11.1



Density Histogram

A density histogram is a histogram whose vertical axis is scaled so that the sum of the areas of its rectangles is 1.

Note: If the sample size (n) is large, a density histogram can be used to estimate the distribution of the population from which the data was obtained.

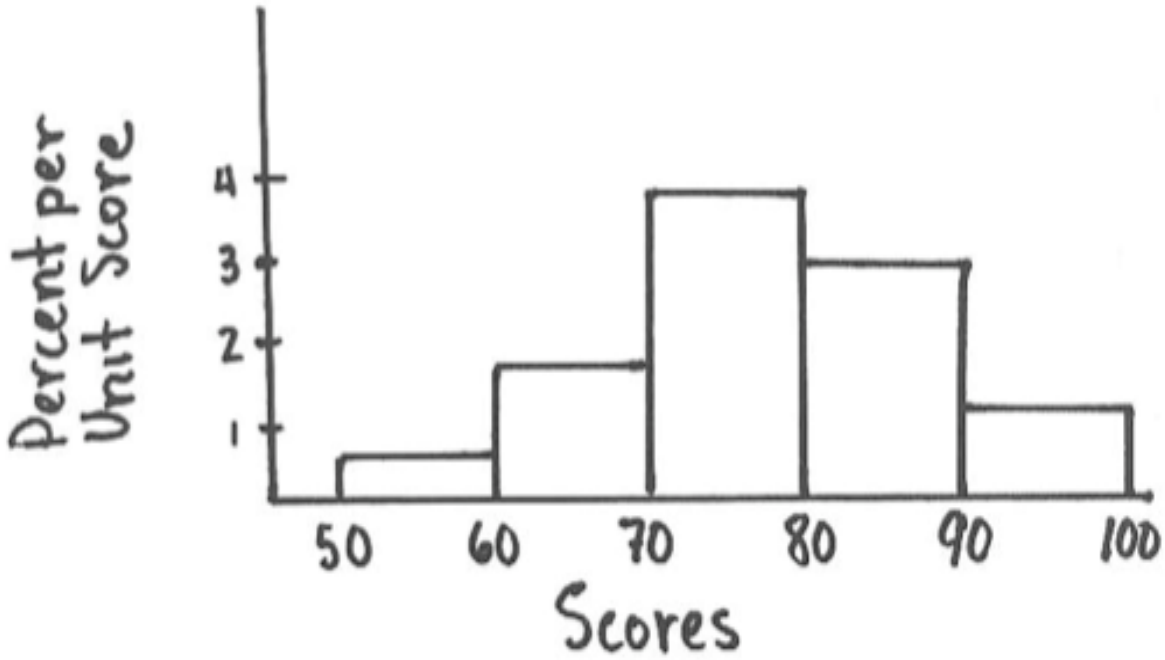
Constructing a density histogram:

1. Compute the proportion (percent) of observations in each class.
2. Divide the proportion (percent) associated with each class by the width of that class (this yields the proportion (percent) of the observations associated with each unit of the measurement scale).
3. Draw the histogram using the values computed in step 2.

Density Histogram: Example

Math 126 Exam Scores

Class	Freq	Percent	Percent/Unit Score
50–59	1	5.6	0.56%
60–69	3	16.7	1.67%
70–79	7	38.9	3.89%
80–89	5	27.8	2.78%
90–100	2	11.1	1.01%



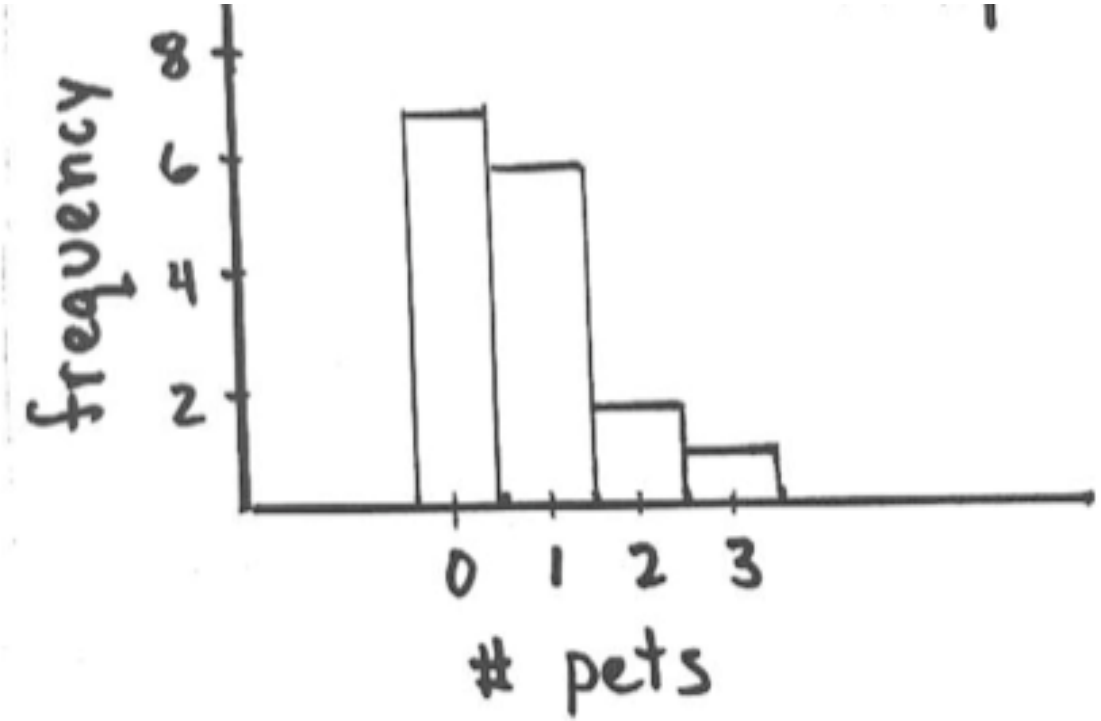
Ungrouped Frequency Distribution

An ungrouped frequency distribution is a list (or table) which pairs each value of a discrete numerical variable with its frequencies (counts).

Ungrouped Frequency Histogram

Number of Pets

Number of Pets	Freq
0	7
1	6
2	2
3	1



Graphical Distortions

- The area fallacy
- The missing baseline
- The Combination Graph