

STP 226 ELEMENTARY STATISTICS

NOTES

PART 2 - DESCRIPTIVE STATISTICS

CHAPTER 2

ORGANIZING DATA

Descriptive Statistics - include methods for organizing and summarizing information clearly and effectively.

- classify data by type
- organize into tables
- summarize with graphs

2.1 Variables and Data.

Variable - a characteristic that varies from one person or thing to another

Example - height, weight, # of siblings, sex, marital status, eye color

Qualitative variable - a non-numerically valued variable

Example - sex, marital status, eye color

Quantitative variable - a numerically valued variable

Example - height, weight, # of sibling, class rank

Discrete variable - a quantitative variable whose possible values form a finite (or countably infinite) set of numbers.

- counting variable

Example - # of siblings, class rank

Continuous variable (measuring variable)- a quantitative variable whose possible values form some interval of numbers

Example - height, weight

Data - information obtained by observing values of a variable.

Qualitative data - data obtained by observing values of a qualitative variable.

Quantitative data - data obtained by observing values of a quantitative variable

Discrete data - data obtained by observing values of a discrete variable

Continuous data - data obtained by observing values of a continuous variable

Observation - individual piece of data, eg. 5

Data set - collection of all observations

Examples of variables and data:

Names	Sex	Class Rank	Weight (lb)	Blood type
Mary	F	1 st	130	AB
John	M	2 nd	171	A
Fatima	F	3 rd	127	B
Cindy	F	4 th	136	O
Alex	M	5 th	180	A

Sex and blood type are Qualitative variables, Class rank is quantitative discrete variable (you may also call it ordinal), Weight is quantitative continuous variable.

It is important to classify data correctly since descriptive and inferential statistics uses only certain types of data.

Even statisticians sometimes disagree on type of data.

2.2 Grouping Data

As the quantity of data becomes large, grouping into categories/classes becomes necessary to make data more readable and understandable.

Guidelines

1. # of classes small enough to provide effective summary but large enough to display relevant characteristics of the data (anywhere from 5 -20 classes seem reasonable)
2. each observation must belong to one and only one, class
3. wherever feasible, all classes should have the same width

Use of the **Frequency Distribution** (table with columns for classes and frequencies) to display the information

Use of the **Relative Frequency Distribution** (table with columns for classes and relative frequencies) to display the information

Classes - categories for grouping data

Frequency - # of observations that fall in a class

Frequency distribution - listing of all classes along with their frequencies

Relative Frequency - ratio of the frequency of a class to the total # of observations

- may involve round-off errors

Relative Frequency distribution - listing of all classes along with their relative frequencies

Lower cut point - smallest value that can go into a class

Upper cut point - smallest value that can go into the next higher class

UCP of any class = LCP of the next higher class

Midpoint - middle of a class = average of the cut points = $(LCP + UCP)/2$

Width - $(UCP - LCP)$

Grouped-data table - table that includes the columns; classes, frequencies, relative frequencies, and midpoints.

Single value grouping - one number per class

Example:

Construct **frequency and relative frequency table** for the given data representing a number of school age children in 20 households

1 2 3 4 2 4 5 2 1 3 4 5 1 0 2 3 4 0 1 1

<i>Number of school children</i>	<i>Frequency</i>	<i>Relative frequency</i>
0	2	$2/20=0.10=10\%$
1	5	$5/20=0.25=25\%$
2	4	$4/20=0.20=20\%$
3	3	$3/20=0.15=15\%$
4	4	$4/20=0.20=20\%$
5	2	$2/20=0.10=10\%$
	20	1=100%

Frequency/relative frequency distribution for qualitative data

- information can still be put into classes

- can compute frequencies and relative frequencies

Example:

Construct frequency and relative frequency histogram for the given data representing days to maturity for 40 short term investments:

70 64 99 55 89 87 65 62 38 67 70 60 69 78 39 75 56 71 51 99
68 95 86 57 53 47 50 55 81 80 98 51 36 63 66 85 79 83 70 64

<i>Days to maturity</i>	<i>Frequen cy</i>	<i>Relative frequency</i>
30<40	3	$3/40=7.5\%$
40<50	1	$1/40=2.5\%$
50<60	8	$8/40=20\%$
60<70	10	$10/40=25\%$
70<80	7	$7/40=17.5\%$
80<90	7	$7/40=17.5\%$
90<100	4	$4/40=10\%$
	40	$1=100\%$

2.3 Graphs and Charts

Frequency histogram - frequency of each class is the height of the bar which is equal the frequency

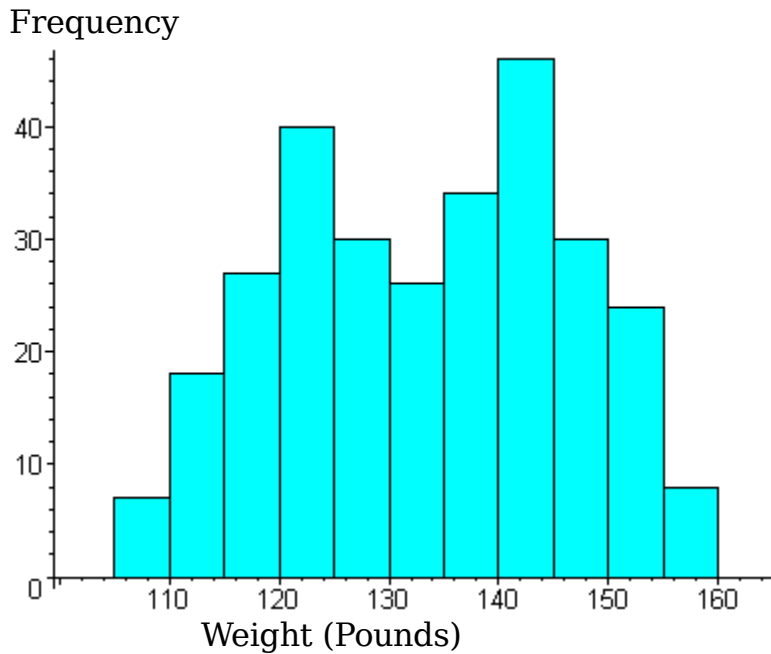
- displays the classes on the horizontal axis and the frequencies
- on the vertical axis.

Relative frequency histogram - relative frequency of each class is the height of the bar which is equal the relative frequency

- displays the classes on the horizontal axis and the relative frequencies on the vertical axis.

Histograms for single value grouping - the single values are placed below the middle of the bars

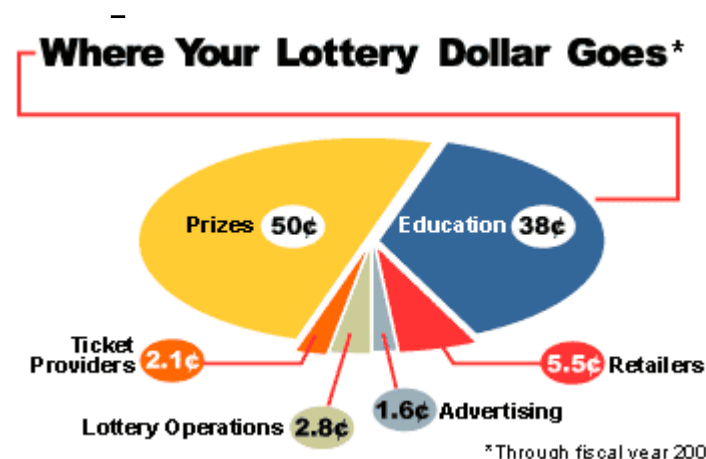
Example: A frequency histogram is given below for the weights of a sample of college students.

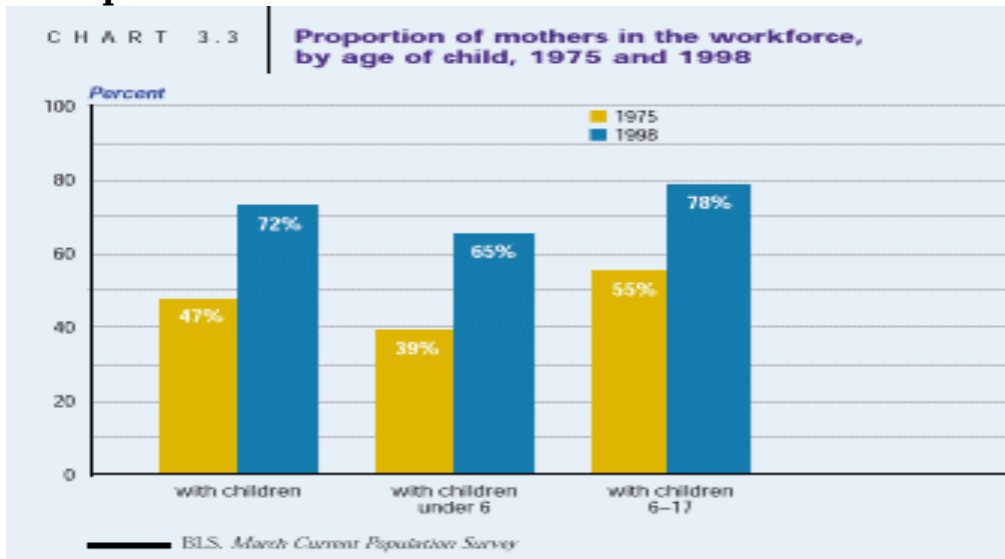


Graphical displays for qualitative data

- **Pie charts**
- **Bar charts** - similar to histograms, except that the bars do not touch each other since the categories are not on some numeric scale. They represent distinct categories.

Example: Pie chart



Example: bar chart:**Stem-and-leaf Diagrams (Stemplot)-fast way to make a histogram:****How to make a stemplot:**

1. Separate each observation into a **stem** (has all but the last digit, can be 1, 2, or more digits) consisting of all but the final (rightmost) digit and a **leaf** (has only one digit), the final digit.
2. Write the stems in a vertical column with the smallest at the top, and draw a vertical line at the right of this column.
3. Write each leaf in the row to the right of its stem, in increasing order out from the stem.

Example: stem plot for Days to maturity for 40 short term investments:

3 869	stems: tens
4 7	
5 71635105	leaves: ones
6 2473640985	
7 0510980	
8 5917036	
9 9985	

Back-to-back stemplot – uses one stem and two sets of leaves, one on either side of the stem helps to make comparison between two data sets.

Stemplot with two leaves per stem:

The number of stems can be doubled by **splitting the stem in two** ; one with leaves from 0 to 4 and the other with leaves 5 to 9.

For example our previous data can be represented as below, leaves are also ordered

3	stems: tens
3 689	leaves: ones
4	
4 7	
5 0113	
5 5567	
6 02344	
6 56789	
7 0001	
7 589	
8 013	
8 5679	
9	
9 5899	

Good idea is to round off numbers to only a few digits before trying to make a stemplot (lose some accuracy in measurements)

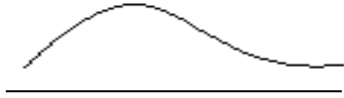
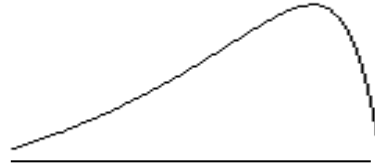
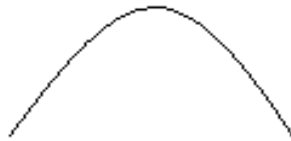
2.4 Distribution Shapes

Distribution: Shapes, Symmetry, and Skewness.

- a distribution of a data set is a table, graph, or formula that tells us the values of the observation and how often they occur

Distribution Shapes

Bell-shaped	J -shaped	Reverse J-shaped
Right skewed	Multimodal	
Left skewed	Triangular	
Uniform	Bimodal	

Right skewed**Left skewed****Symmetric****Modality:**

Unimodal - has one peak

eg. Bell-shaped, Triangular, Reverse J-shaped, J-shaped, Right skewed, Left skewed

Bimodal - has two peaks (technically, all peaks should be same height, not so in practice)

Multimodal - has 3 or more peaks

Symmetry and Skewness

Symmetry - property of a distribution to be divided into 2 parts that are mirror images of each other.

Do not have to be exact in identifying symmetry.

Eg. bell-shaped, triangular, uniform.

Non-symmetric Distribution - Reverse J-shaped, J-shaped, Right skewed, Left skewed

The **distribution of population data** is called population distribution, or the distribution of the variable.

The **distribution of sample data** is a sample distribution.

The distribution of a random sample from a population approximates the population distribution, hence, larger samples give better approximation.

2.5 Misleading Graphs

Scales on the vertical axis should begin at 0 to compare the heights of the bars correctly.

When the vertical scale does not begin at 0, it may be easier to see an up/down trend as you move from left to right, but one may easily make the error of comparing the heights of the bars.

- not recommended even though they still appear in some reputable publications.

Need to be careful to use and observe proper scaling so that one can get precise and consistent interpretation.

If a truncated graph is to be used, use the symbol // to show that a portion of the graph is truncated. This will alert the audience to be careful in interpreting from the graph.