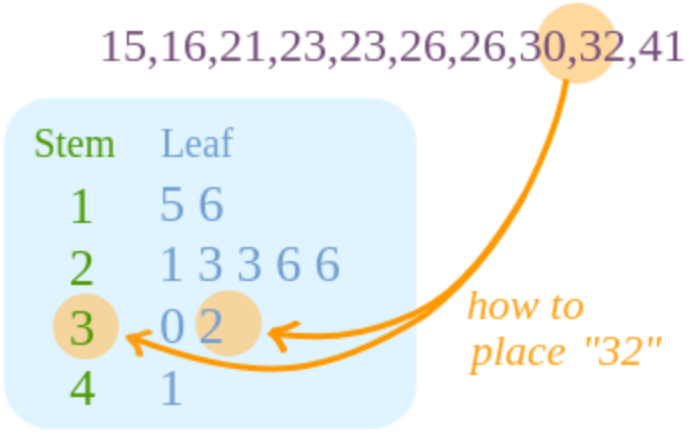
MA-120 Probability and Statistics

Nazar Khan
PUCIT
Lecture 3: Descriptive Statistics

Descriptive Statistics

- Large amounts of data needs to be summarized
  - Stem-and-leaf display
  - Frequency distribution
  - Histogram
  - Bar Chart, Pictogram, Pie Chart
  - Mean
  - Median
  - Mode
  - Standard Deviation
  - The Normal Curve

Stem-and-leaf Display



- To give a general overview of data
- Tens digit on the stem
- Units digit on the leaves
- Sorted leaves

42	37	57	49	39
24	39	51	34	28
23	33	28	25	52
32	58	50	49	27
35	50	23	32	34
44	28	29	25	23
37	46	49	28	44
44	34	55	60	59
55	59	38	25	57
51	42	60	39	27

Frequency Distribution

- To represent distribution of data
- Data is divided into groups or categories
- Usually 6 to 15 categories
- Preferable to have classes beginning or ending in multiples of 5 or 10.

23	24	18	14	20	24	24	26	23	21
16	15	19	20	22	14	13	20	19	27
29	22	38	28	34	32	23	19	21	31
16	28	19	18	12	27	15	21	25	16
30	17	22	29	29	18	25	20	16	11
17	12	15	24	25	21	22	17	18	15
21	20	23	18	17	15	16	26	23	22
11	16	18	20	23	19	17	15	20	10

Amount of time (in hours) spent by 80 college students in leisure activities in a typical school week

Hours	*Tally*	*Frequency*
10–14	𝍸 |||	8
15–19	𝍸 𝍸 𝍸 𝍸 𝍸 |||	28
20–24	𝍸 𝍸 𝍸 𝍸 𝍸 ||	27
25–29	𝍸 𝍸 ||	12
30–34	||||	4
35–39	|	1
Total		80

Frequency distribution

Frequency Distribution

- Class limits
  - lower and upper
- Class boundaries (“real” limits)
  - to handle rounding-off
  - example: values >=9.5 and <14.5 lie in the class 10-14
- Class marks = (lower+upper)/2
- Class intervals
  - Differences between class marks
- Distribution interval
  - Equal to class interval if all intervals are same

Hours	*Tally*	*Frequency*
10–14	𝍸 |||	8
15–19	𝍸 𝍸 𝍸 𝍸 𝍸 |||	28
20–24	𝍸 𝍸 𝍸 𝍸 𝍸 ||	27
25–29	𝍸 𝍸 ||	12
30–34	||||	4
35–39	|	1
Total		80

Frequency distribution

Frequency Distribution Variations

- Percentage distribution
- Cumulative distribution
- Cumulative percentage distribution

Histogram

- Graphical representation of frequency distribution
- Two types
  1. Count
  2. Percentage (area of a block represents percentage of data lying within that block)
- What should be the total area under a percentage histogram?

Mean vs. Median

- Different measures of “central location”.
- Median position = (n+1)/2
- Take any list of numbers and compute the mean and median.
- Recompute mean and median after reversing the digits of one of the numbers.
- The new mean will be affected more than the new median.
- Mean is sensitive to outliers while median is a robust statistic.

Weighted Mean

-

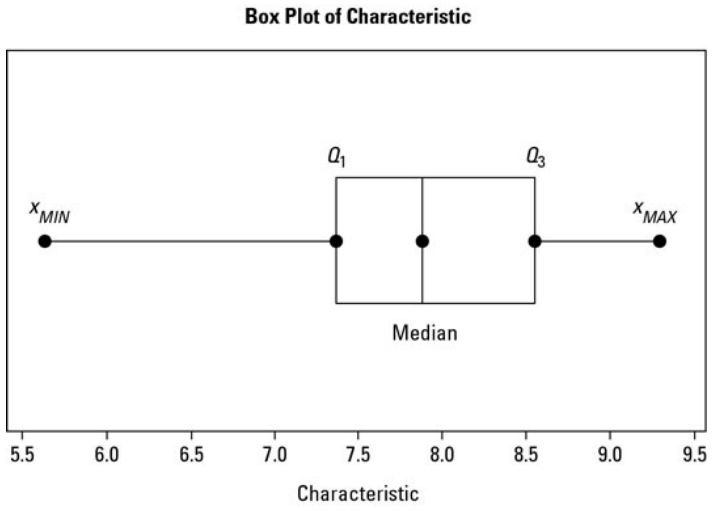
Computing median using stem-and-leaf plot

Quartiles and Percentiles

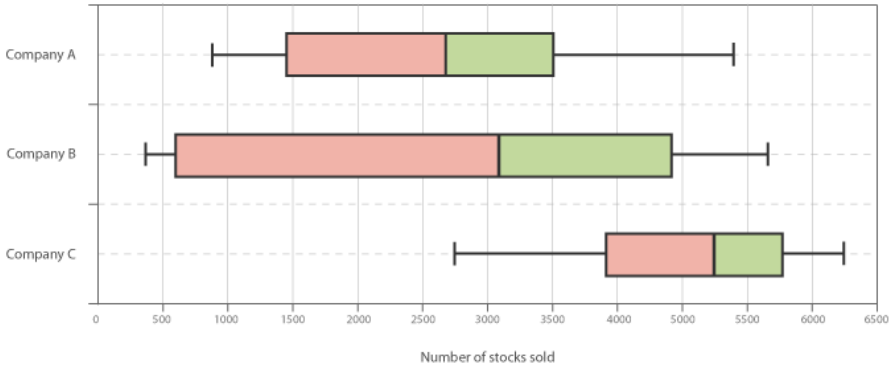
- First, second and third quartiles Q1, Q2, Q3
- Median is the second quartile Q2.
- 70^{th} percentile is the value for which 70% of the numbers in the list are smaller.

Box-and-whisker plot

- Graphical plot summarizing the
  - Min
  - Max
  - Q1
  - Q2
  - Q3



Comparison of three different companies using box-and-whisker plots summarizing their stock sales.



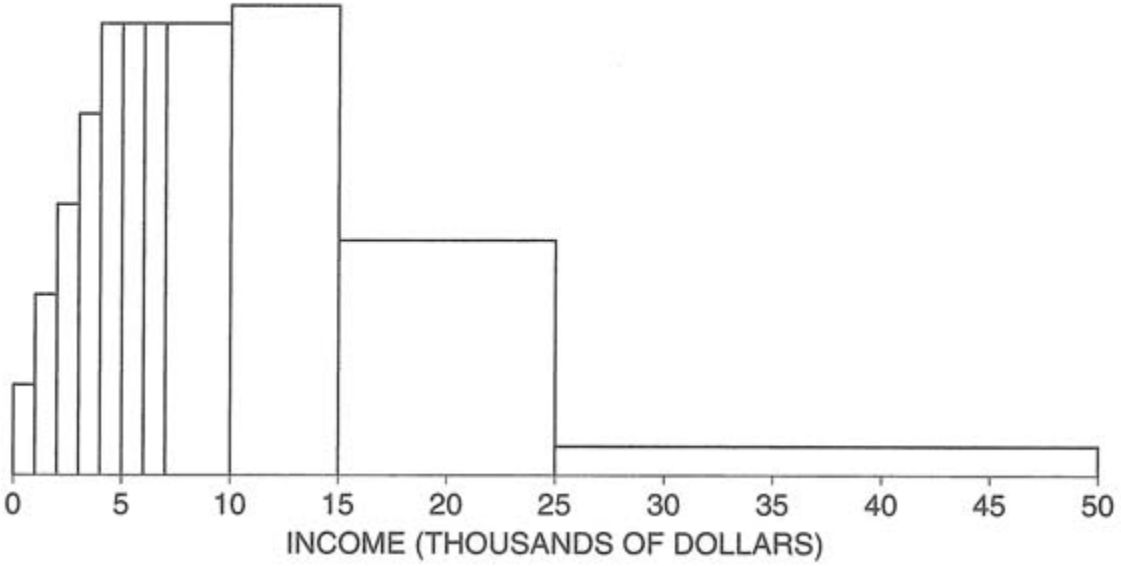
Mode

- Most frequent value.
- Represents most typical behaviour.

Standard deviation

- Measure of the spread of the data.

Figure 1. A histogram. This graph shows the distribution of families by income in the U.S. in 1973.

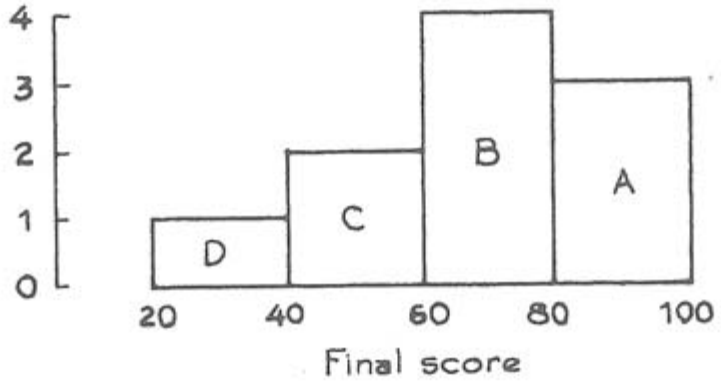


The histogram below shows the distribution of final scores in a certain class.

(a) Which block represents the people who scored between 60 and 80?

(b) Ten percent scored between 20 and 40. About what percentage scored between 40 and 60?

(c) About what percentage scored over 60?



Drawing a histogram

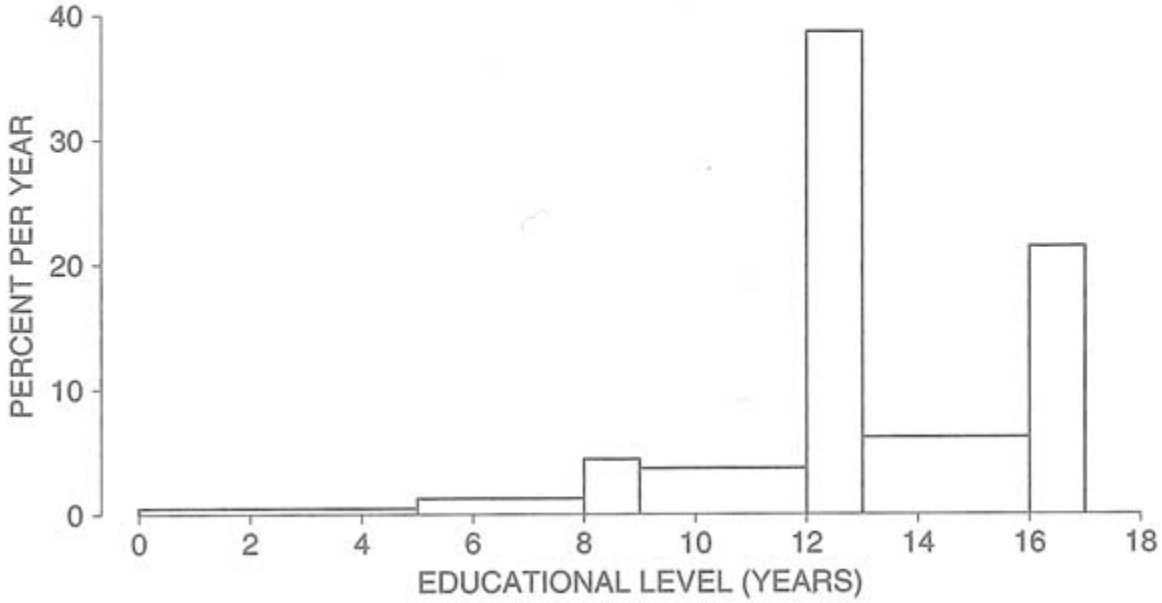
- Horizontal scale
- Vertical scale
- End-point notation

Table 1. Distribution of families by income in the U.S. in 1973. Class intervals include the left endpoint, but not the right endpoint.

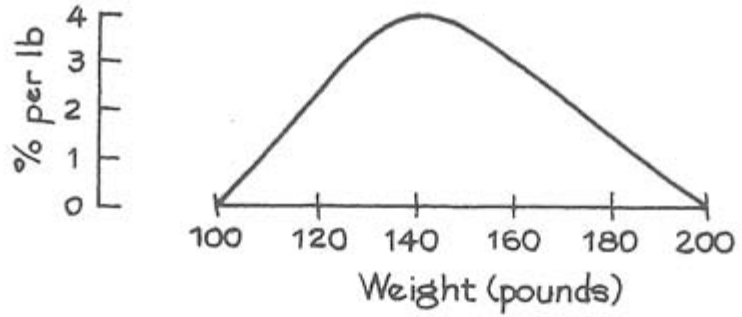
Income level	*Percent*
\$0–\$1,000	1
\$1,000–\$2,000	2
\$2,000–\$3,000	3
\$3,000–\$4,000	4
\$4,000–\$5,000	5
\$5,000–\$6,000	5
\$6,000–\$7,000	5
\$7,000–\$10,000	15
\$10,000–\$15,000	26
\$15,000–\$25,000	26
\$25,000–\$50,000	8
\$50,000 and over	1

Note: Percents do not add to 100%, due to rounding.

Figure 5. Distribution of persons age 25 and over in the U.S. in 1991 by educational level.



Example 2. Someone has sketched a histogram for the weights of some people, using the density scale. What's wrong?

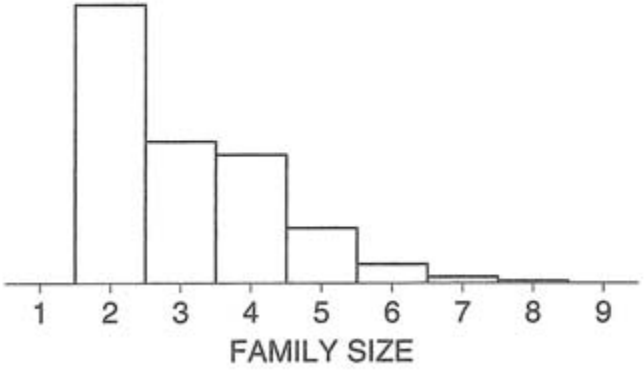


Variable

- A characteristic that changes from sample to sample in the study
  - Age, Income, Smoker, Marital Status, Political affiliation, etc
- Qualitative
  - Marital status, Political affiliation
- Quantitative
  - Discrete
    - Age
  - Continuous
    - Income

Variable	Qualitative	Discrete

Figure 6. Histogram showing distribution of families by size in 2005. With a discrete variable, the class intervals are centered at the possible values.



Controlling for a variable

Cross-tabulation

- To handle confounding factors

Selective Breeding

Figure 8. Tryon's experiment. Distribution of intelligence in the original population.

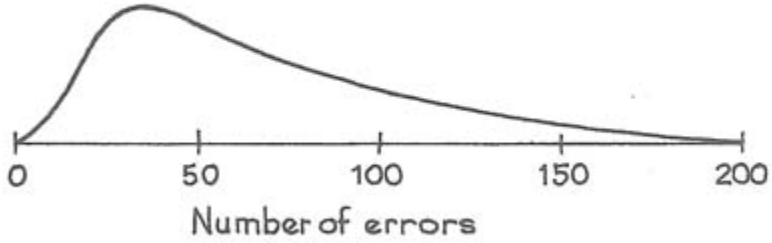


Figure 9. Tryon's experiment. After seven generations of selective breeding, there is a clear separation into "maze-bright" and "maze-dull" strains.