MA-250 Probability and Statistics

Nazar Khan

PUCIT

Lecture 2

Descriptive Statistics

- Large amounts of data needs to be summarized
  - Histogram
  - Mean and Standard Deviation
 - The Normal Curve

Histogram

- To represent distribution of data
- Area of a block represents percentage of data lying within that block
- What should be the total area under a histogram?

Figure 1. A histogram. This graph shows the distribution of families by income in the U.S. in 1973.
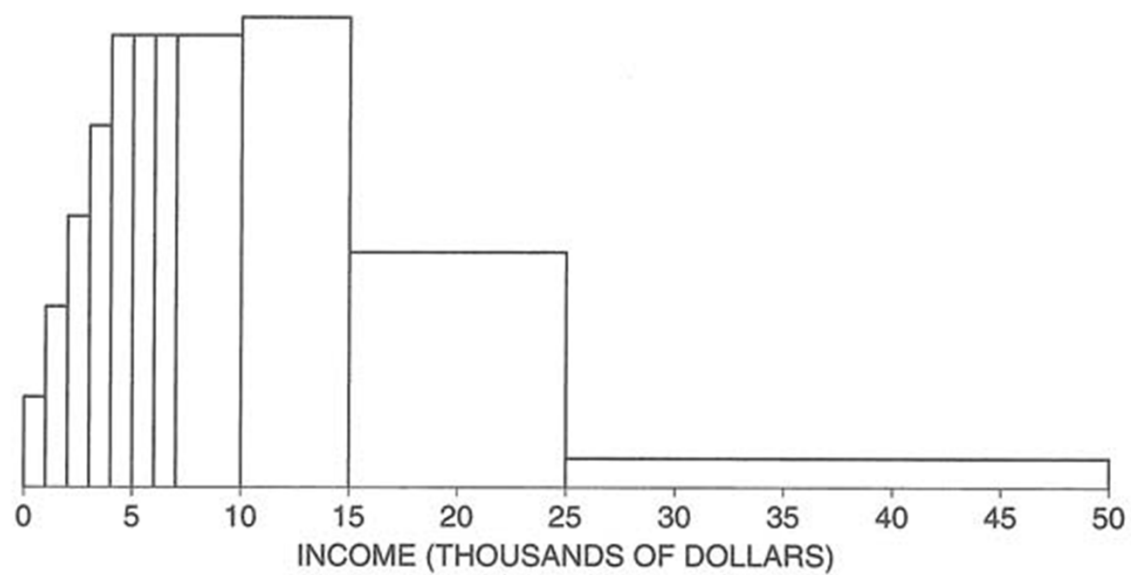
The histogram below shows the distribution of final scores in a certain class.

(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?

Final score
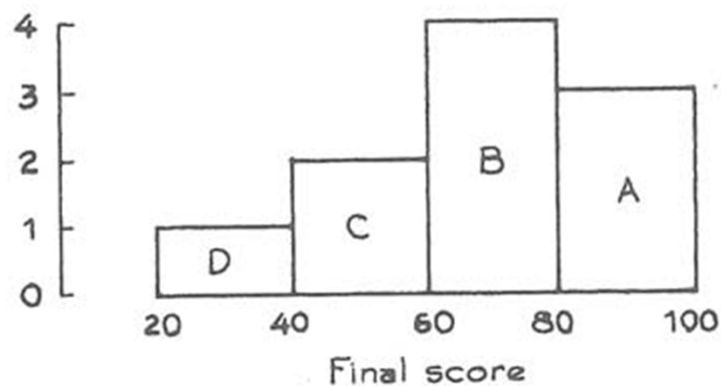
Drawing a histogram

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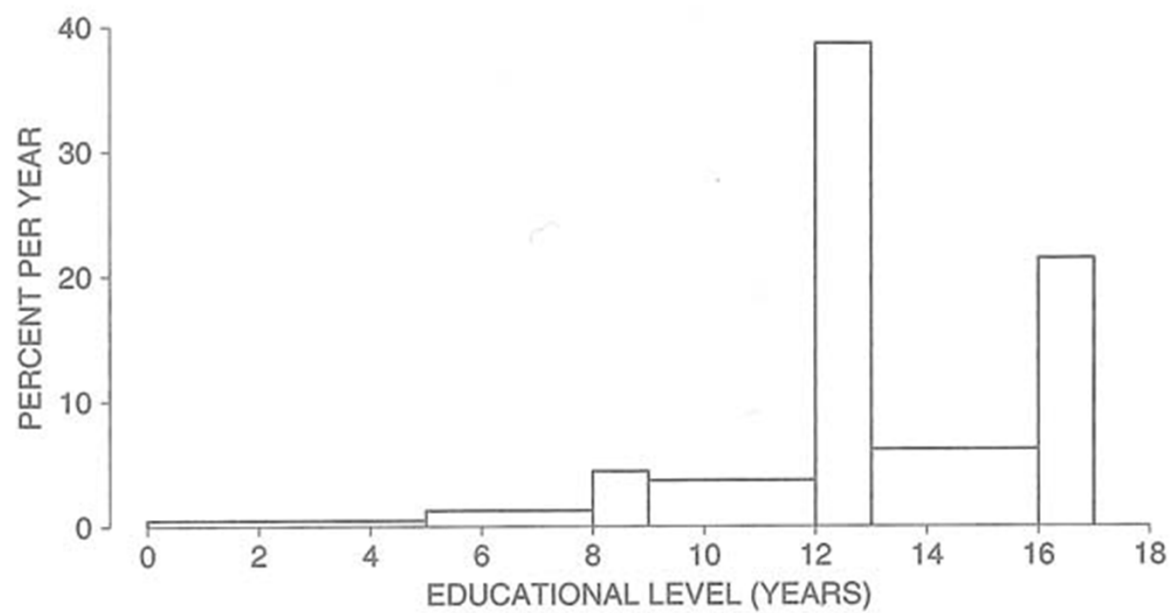
Example 2.   Someone has sketched a histogram for the weights of some people, using the density scale. What's wrong?
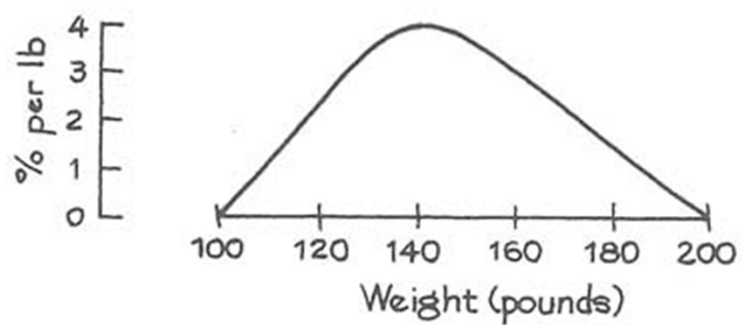
Variables

- 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
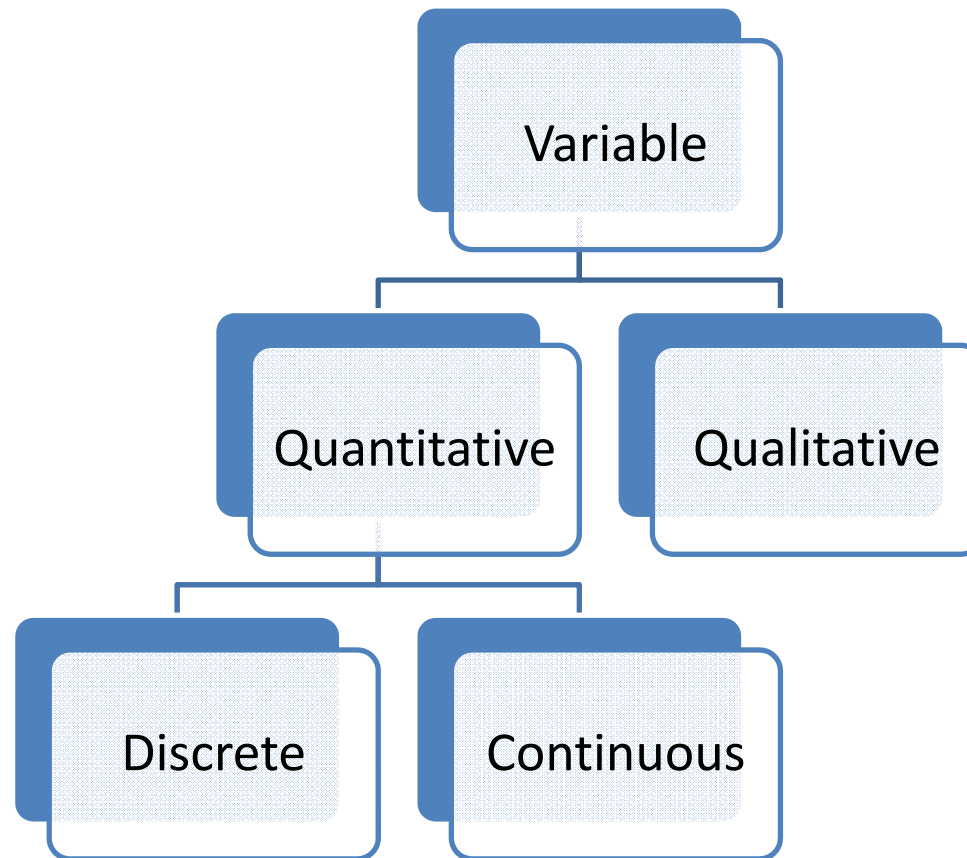   - Age
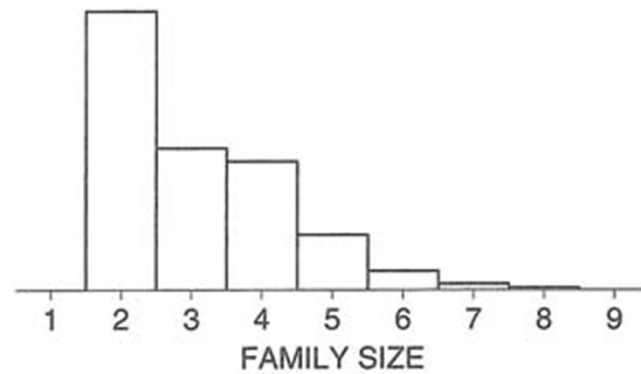 - Continuous
    - Income

Variables

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

- How can we handle confounding factors in observational studies?
  - Remember, we have no control over the formation of treatment and control groups.

- What if both groups are different? For example, in terms of age.

Smokers (all > 30 years)	Non-smokers (kids)

- We can control for age in observational studies by comparing across specific age-groups in both treatment and control groups.

Smokers (25-30 years)	Non-smokers (25-30 years)

Cross-tabulation

- To handle confounding factors

Is intelligence genetic?

- A theory[1] in Psychology states that scores on intellectual tests depend on
  - general intelligence, and
  - ability specific to the test
- Robert Tyron checked this on rats.
  - 142 rats
  - 19 runs through a maze
  - Dead-ends were counted for each run of each mouse.
- Selective breeding
  - Maze-bright rats were bred with each other.
  - Maze-dull rats were bred with each other.

1 Charles Spearman, *Abilities of Man*, 1927

Selective Breeding

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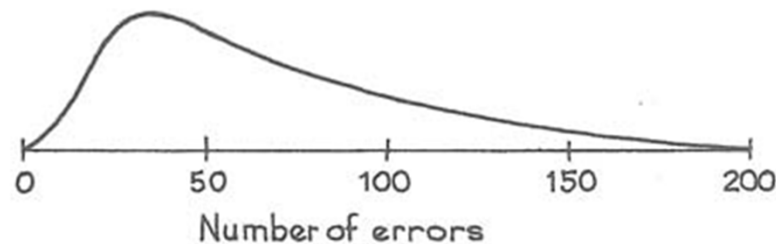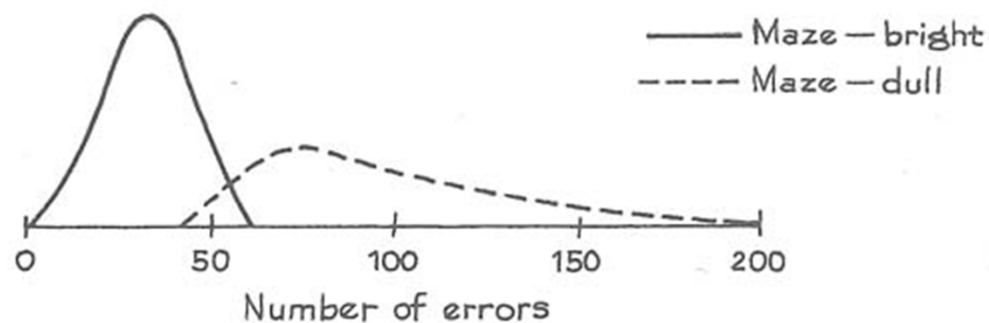

Number of errors

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

——— Maze — bright
----- Maze — dull

Number of errors

Does this support Spearman's theory?
Is intelligence genetic?

Is intelligence genetic?

- Interestingly, Tyron found that maze-bright rats were no better than maze-dull rats on other tasks.
- This actually goes against Spearman's theory.
  - A certain mental ability might have been genetically transferred, but not overall intelligence.