Chapter 3: Histograms

Context

■ So far we looked at data collection:
- ◆ Study design:
  - ▪ Controlled experiments (ch 1)
  - ▪ Observational studies (ch 2)
- ◆ Sampling (ch 19)

■ We'll now look at summarizing data with *histograms*, tables, bar charts and pie charts.

Histograms

Histogram

■ Main lesson: areas of the blocks represent percentages

■ See Figure 2 on page 33

■ Discuss horizontal axis: *class intervals*

■ Discuss *endpoint convention*: rule about what we do with data points that fall on the border of two class intervals

■ Discuss problem 1 on page 33

■ See overhead

Drawing a histogram

■ Start with table that gives you the percentage of cases in each class interval (like Table 1 on page 35)

■ Draw the horizontal axis

■ Determine the height of the blocks

■ Draw the blocks

■ Put the right units along the axes

■ See overhead

Histogram: main points

■ A histogram is used to summarize data

■ Histogram is a graph consisting of blocks

■ The horizontal axis consists of *class-intervals*

■ *Areas* of the blocks represent *percentages*

■ *Height* of the block represents the *percentage of cases per horizontal unit* or *crowding*

■ The unit of the vertical axis is *percent per horizontal unit*

■ Total area of histogram is 100%

■ In articles you will sometimes see histograms with a different vertical scale. The scale that we use is called the *density scale*.

Example: income data

■ The unit of the horizontal axis is: thousand dollar

■ The unit of the vertical axis is: percent per thousand dollar

■ The height of a block represents the percentage of people in a thousand dollar interval

■ What is the height of the block for class interval $\$7,000 - \$10,000$ which contains 15% of the people?

  ◆ The class interval contains 3 intervals of thousand dollar

  ◆ Hence, the percent of people in a thousand dollar interval is

$$\frac{15\%}{3} = 5\%$$

  ◆ The height of the block is 5 percent per thousand dollar

Example: income data

■ What percentage of people made $\$15,000 - \$20,000$ per year?

  ◆ We need to compute the corresponding area of the histogram - see overhead

  ◆ The height of the histogram is (the units are in black):
  2.6 percent per thousand dollars

  ◆ The length of the interval of interest is:
  5 thousands of dollars

  ◆ The area of a block is given by length times height

  ◆ So the area is:
  5 thousands of dollars $\times$ 2.6 percent per thousand dollars $= 13$ percent

---

Variables

- See overhead

Examples of variables

- Examples of qualitative variables:
  - ◆ marital status (single, married, widowed, divorced)
  - ◆ employment status (unemployed, employed, not in the labor force
  - ◆ gender (male, female)
  - ◆ type of pie that was sold (blueberry, blackberry, etc)
- Examples of quantitative variables:
  - ◆ age
  - ◆ family size
  - ◆ income

Quantitative variables

- Quantitative variables can be:
  - ◆ Discrete - values can differ only by fixed amounts. Example:
    - ▪ family size
  - ◆ Continuous - values can differ by any amount. Example:
    - ▪ income

Histogram for discrete variable

- Center the class intervals at the possible values
- We don't need to make an endpoint convention
- See Figure 6 on page 44

Qualitative data

- How to summarize qualitative data? (not in book)
  - ◆ Table
  - ◆ Pie charts
  - ◆ Bar charts
- See examples

Controlling for a variable

Controlling for a variable

- Make histograms or tables for groups that are comparable
- Example:
  - ◆ Does the contraceptive pill increase blood pressure?
  - ◆ Compare women who use the pill to women who don't use the pill
  - ◆ Women who don't use the pill tend to be older, and blood pressure increases with age. So age is a confounding factor
  - ◆ Solution: make histograms, boxplots, tables for women in various age groups: 17-24, 25-34, 35-44, 45-58.

Examples

- Figure 7 on page 45
- Table 2 on page 47