Chapter 3

DISPLAYING AND SUMMARIZING QUANTITATIVE DATA

1

A DISTRIBUTION

- The distribution describes the overall layout of the data
- One way we can visualize the distribution is by first partitioning our variable into small groupings, or bins

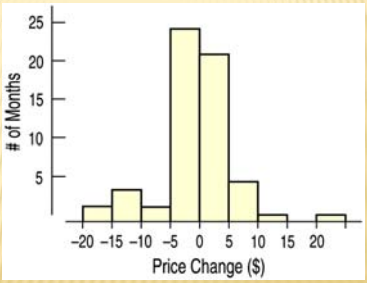
2

HISTOGRAMS

- A histogram is a bar chart, where the bars are adjacent, used to give a visual image of the distribution of a quantitative variable
- The counts, or frequencies, are on the vertical axis. The quantitative variable is plotted along the horizontal axis, divided by its bins
- We could also have a relative frequency histogram, where the percentages (instead of frequencies) are along the vertical axis

3

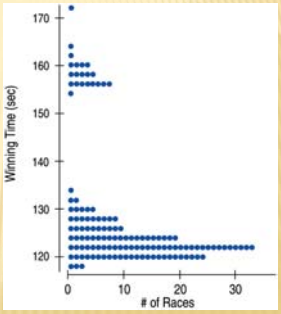
HISTOGRAM



4

DOTPLOT

- Each data value is plotted as a point (or dot) along a scale of values. Dots representing the same value are stacked.



From *Stats Modeling the World* by Bock, Velleman, & De Veaux, 2010, p. 49.

5

STEM-AND-LEAF PLOT

- The stem contains all but the last digit of a number, and the leaf is the last digit of the number

Example: Suppose we have the following test scores:
67, 72, 85, 75, 89, 89, 88, 90, 99, 100

Stem	Leaves
6	7
7	2 5
8	5 8 9 9
9	0 9
10	0

6

STEM-AND-LEAF PLOT

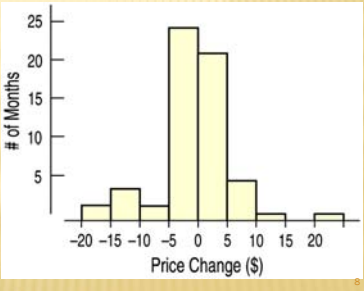
- If we have a large data set, we could expand the plot subdividing rows into those with digits 0 through 4 and those with digits 5 through 9.
- For the previous example:

Stem	Leaves
6	7
7	2
7	5
8	
8	5 8 9 9
9	0
9	9
10	0

7

DESCRIBING A DISTRIBUTION

- Shape
- Center
- Spread



8

SHAPE OF A DISTRIBUTION

- The mode is the value that occurs most frequently
- Examples
  - For the data set: 5, 5, 5, 3, 1, 5, 1, 4, 3, 5
    - The mode is 5
  - For the data set: 1, 2, 2, 2, 3, 4, 5, 6, 6, 6, 7, 9
    - There are two modes: 2 and 6
    - This is called bimodal
  - For the data set: 1, 2, 3, 6, 7, 8, 9, 10
    - There is no mode

9

SHAPE OF A DISTRIBUTION

- A distribution of data is skewed if it is not symmetric and if it extends more to one side than the other
- An outlier is a data point that is not consistent with the bulk of the data
  - It is unusually high or low

10

CENTER OF THE DISTRIBUTION

- The median is the middle value when the original data values are arranged in order of increasing (or decreasing) magnitude
  - If the number of values is odd, the median is the number located in the exact middle of the list
    - In the (n+1)/2 position
  - If the number of values is even, the median is found by computing the average of the two middle numbers
    - Average of the n/2 and (n/2 + 1) positions

Here, n is the number of data points

11

MEDIAN EXAMPLES

- Find the median of the following salaries (in millions) paid to executives: 6.72, 3.46, 3.60, 6.44, 26.70
  - The median is 6.44
- Find the median of the following salaries (in millions) paid to executives: 6.72, 3.46, 3.60, 6.44
  - The median is 5.02

12

THE SPREAD OF THE DISTRIBUTION

- The range of a distribution is the calculated as range = (high value) – (low value)
- The lower quartile is the median of the lower half of the ordered data values
  - It's the median of the data values below the median of the data set
- The upper quartile is the median of the upper half of the ordered data values
  - It's the median of the data values that are above the median of the data set

13

SPREAD CONTINUED

- The interquartile range (IQR) is calculated as IQR = (upper quartile) – (lower quartile)
- The five-number summary displays the lowest value; the cutoff points for the lower quartile, median, and upper quartile of the data; and the highest value

14

SPREAD CONTINUED

- The k^{th} percentile is the number that has k% of the data values at or below it
  - The lower quartile = 25th percentile
  - The median = 50th percentile
  - The upper quartile = 75th percentile

15

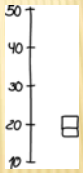
BOXPLOT

- The box covers the middle 50% of the data (the top of the box is the upper quartile and the base of the box is the lower quartile) and a line within the box marks the median value
- Lines extend from the box marking the extreme values, except possible outliers (further than 1.5 × IQR from the quartile) which are marked as separate data points

16

CONSTRUCTING BOXPLOTS

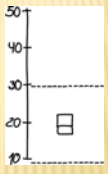
1. Draw a single vertical axis spanning the range of the data. Draw short horizontal lines at the lower and upper quartiles and at the median. Then connect them with vertical lines to form a box.



17

CONSTRUCTING BOXPLOTS (CONT.)

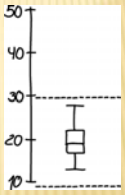
2. Erect "fences" around the main part of the data.
   - The upper fence is 1.5 IQRs above the upper quartile.
   - The lower fence is 1.5 IQRs below the lower quartile.
   - Note: the fences only help with constructing the boxplot and should not appear in the final display.



18

CONSTRUCTING BOXPLOTS (CONT.)

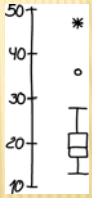
3. Use the fences to grow "whiskers."
   - Draw lines from the ends of the box up and down to the *most extreme data values found within the fences*.
   - If a data value falls outside one of the fences, we do *not* connect it with a whisker.



19

CONSTRUCTING BOXPLOTS (CONT.)

4. Add the outliers by displaying any data values beyond the fences with special symbols.
   - We often use a different symbol for "far outliers" that are farther than 3 IQRs from the quartiles.



20

THE MEAN

- The mean is the number obtained by adding the values of the data points and dividing the total by the number of values

$$\bar{y} = \frac{\sum y}{n}$$

where $y_1, y_2, \ldots, y_n$ represent the individual raw data values and n is the number of individuals in the sample

21

STANDARD DEVIATION

- The standard deviation of a sample is a measure of variation of values about the mean

$$s = \sqrt{\frac{\sum (y - \bar{y})^2}{n-1}}$$

- s is the standard deviation and s^2 is the variance

22

STANDARD DEVIATION AND IQR

- The standard deviation describes more than the IQR because it takes into account how far from the mean each data point lies

23

DESCRIBING CENTER AND SPREAD

- For a skewed distribution, it's better to report the median and the IQR
- For a symmetric distribution, mean and median are both good descriptors but reporting the mean and standard deviation has several uses we'll learn

24