Chapter 2

Describing Distributions with Numbers

Numerical Summaries

- Center of the data
  - mean
  - median
- Variation
  - range
  - quartiles (interquartile range)
  - variance
  - standard deviation

Mean or Average

- Traditional measure of center
- Sum the values and divide by the number of values

$$\bar{x} = \frac{1}{n}\left(x_1 + x_2 + \cdots + x_n\right) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Median (M)

- A <u>*resistant measure*</u> of the data's center
- At least half of the ***ordered*** values are less than or equal to the median value
- At least half of the ***ordered*** values are greater than or equal to the median value
- If n is <u>odd</u>, the median is the middle ordered value
- If n is <u>even</u>, the median is the average of the two middle ordered values

Median (M)

Location of the median: $L(M) = (n+1)/2$, where n = sample size.

Example: If 25 data values are recorded, the Median would be the $(25+1)/2 = 13^{th}$ ordered value.

Median

- Example 1 data: 2 4 6
  Median (M) = 4
- Example 2 data: 2 4 6 8
  Median = 5 (ave. of 4 and 6)
- Example 3 data: 6 2 4
  Median $\neq$ 2
  (**order** the values: 2 4 6 , so Median = 4)

Comparing the Mean & Median

- The mean and median of data from a symmetric distribution should be close together. The actual (true) mean and median of a symmetric distribution are exactly the same.
- In a skewed distribution, the mean is farther out in the long tail than is the median [the mean is 'pulled' in the direction of the possible outlier(s)].

Question

A recent newspaper article in California said that the ***median*** price of single-family homes sold in the past year in the local area was \$136,000 and the ***mean*** price was \$149,160. Which do you think is more useful to someone considering the purchase of a home, the median or the mean?

Answer

Both! Average is affected by outliers while median is not. For example, if one house is extremely expensive, then the average will rise. The median would ignore that outlier.

Case Study

Airline fares

appeared in the *New York Times* on November 5, 1995

"...about 60% of airline passengers 'pay less than the average fare' for their specific flight."

- How can this be?

  13% of passengers pay more than 1.5 times the average fare for their flight

Spread, or Variability

- If all values are the same, then they all equal the mean. There is no variability.
- Variability exists when some values are different from (above or below) the mean.
- We will discuss the following measures of spread: range, quartiles, variance, and standard deviation

Range

- One way to measure spread is to give the smallest (*minimum*) and largest (*maximum*) values in the data set;

  Range = max - min
- The range is strongly affected by outliers

  (e.g. one house is extremely expensive and the rest all have the same price. The range is large while there is little variability!)

Quartiles

- Three numbers which divide the <u>ordered</u> data into four equal sized groups.
- Q_1 has 25% of the data below it.
- Q_2 has 50% of the data below it. (Median)
- Q_3 has 75% of the data below it.

Obtaining the Quartiles

- ***Order*** the data.
- For $\mathbf{Q_2}$, just find the median.
- For $\mathbf{Q_1}$, look at the lower half of the data values, those to the left of the median location; find the *median* of this lower half.
- For $\mathbf{Q_3}$, look at the upper half of the data values, those to the right of the median location; find the *median* of this upper half.

Weight Data: Sorted

100	124	148	170	185	215
101	125	150	170	185	220
106	127	150	172	186	260
106	128	152	175	187	
110	130	155	175	192	
110	130	157	180	194	
119	133	~~165~~	180	195	
120	135	165	180	203	
120	139	165	180	210	
123	140	170	185	212	

L(M)=(53+1)/2=27 L(Q1)=(26+1)/2=13.5

Weight Data: Quartiles

- Q_1= 127.5
- Q_2= 165 (Median)
- Q_3= 185

Five-Number Summary

- minimum = 100
- Q_1 = 127.5
- M = 165
- Q_3 = 185
- maximum = 260

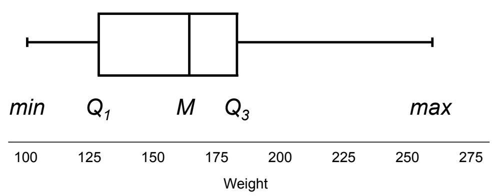
Interquartile Range (IQR) = $\boldsymbol{Q_3 - Q_1}$ = 57.5

IQR gives spread of middle 50% of the data

Boxplot

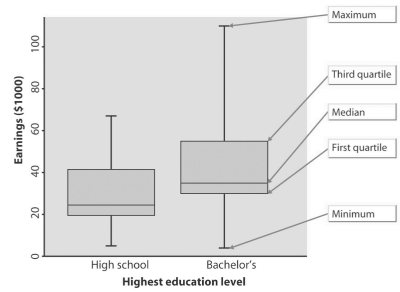
- Central box spans Q_1 and Q_3.
- A line in the box marks the median *M*.
- Lines extend from the box out to the minimum and maximum.

Weight Data: Boxplot



BPS - 5th Ed. Chapter 2 19

Example from Text: Boxplots



BPS - 5th Ed. Chapter 2 20

Identifying Outliers

- The central box of a boxplot spans Q_1 and Q_3; recall that this distance is the Interquartile Range (*IQR*).
- We call an observation a suspected **outlier** if it falls more than 1.5 × *IQR* above the third quartile or below the first quartile.

BPS - 5th Ed. Chapter 2 21

Variance and Standard Deviation

- Recall that variability exists when some values are different from (above or below) the mean.
- Each data value has an associated *deviation from the mean*:

$$x_i - \bar{x}$$

BPS - 5th Ed. Chapter 2 22

Deviations

- what is a *typical* deviation from the mean? *(standard deviation)*
- small values of this typical deviation indicate small variability in the data
- large values of this typical deviation indicate large variability in the data

BPS - 5th Ed. Chapter 2 23

Variance

- Find the mean
- Find the deviation of each value from the mean
- Square the deviations
- Sum the squared deviations
- Divide the sum by *n-1*

(gives typical squared deviation from mean)

BPS - 5th Ed. Chapter 2 24

Variance Formula

$$s^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

BPS - 5th Ed. Chapter 2 25

Standard Deviation Formula

typical deviation from the mean

$$s = \sqrt{\frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \overline{x})^2}$$

[standard deviation = square root of the variance]

BPS - 5th Ed. Chapter 2 26

Variance and Standard Deviation

Example from Text

Metabolic rates of 7 men (cal./24hr.) :

1792 1666 1362 1614 1460 1867 1439

$$\overline{x} = \frac{1792 + 1666 + 1362 + 1614 + 1460 + 1867 + 1439}{7} = \frac{11{,}200}{7} = 1600$$

BPS - 5th Ed. Chapter 2 27

Variance and Standard Deviation

Example from Text

Observations x_i	Deviations $x_i - \overline{x}$	Squared deviations $(x_i - \overline{x})^2$
1792	1792-1600 = 192	$(192)^2$ = 36,864
1666	1666 -1600 = 66	$(66)^2$ = 4,356
1362	1362 -1600 = -238	$(-238)^2$ = 56,644
1614	1614 -1600 = 14	$(14)^2$ = 196
1460	1460 -1600 = -140	$(-140)^2$ = 19,600
1867	1867 -1600 = 267	$(267)^2$ = 71,289
1439	1439 -1600 = -161	$(-161)^2$ = 25,921
	sum = 0	sum = 214,870

BPS - 5th Ed. Chapter 2 28

Variance and Standard Deviation

Example from Text

$$s^2 = \frac{214{,}870}{7-1} = 35{,}811.67$$

$$s = \sqrt{35{,}811.67} = 189.24 \text{ calories}$$

BPS - 5th Ed. Chapter 2 29

Choosing a Summary

- Outliers affect the values of the mean and standard deviation.
- The five-number summary should be used to describe center and spread for skewed distributions, or when outliers are present.
- Use the mean and standard deviation for reasonably symmetric distributions that are free of outliers.
- Best to use both!

BPS - 5th Ed. Chapter 2 30

Number of Books Read for Pleasure: Sorted

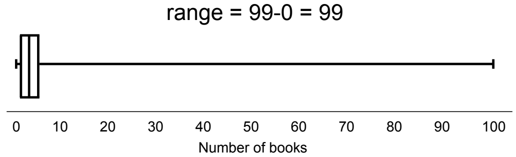
0	1	2	4	10	30
0	1	2	4	10	99
0	1	2	4	12	
0	1	3	5	13	
0	2	3	5	14	
0	2	3	5	14	
0	2	3	5	15	
0	2	4	5	15	
0	2	4	5	20	
1	2	4	6	20	

5.5+(5.5-1)x1.5=12.25

BPS - 5th Ed. Chapter 2 31

Five-Number Summary: Boxplot

Median = 3
interquartile range (iqr) = 5.5-1.0 = 4.5
range = 99-0 = 99



Mean = 7.06 s.d. = 14.43

BPS - 5th Ed. Chapter 2 32