Numerical Descriptions of Data

Measures of Center

$$Mean\ \bar{x} = \frac{\sum x_i}{n}$$

x = data values
n = sample size

Excel: = average ( )

Weighted mean

$$\bar{x} = \frac{\sum(x_i * w_i)}{\sum w_i}$$

x_i= ith data value
w_i= weight of the ith data value

Median = data value in middle of ordered array

Excel: = median ( )

Mode = most frequently occurring value

Excel: =mode ( )

Unimodal – 1 mode
Bimodal – 2 mode
Multimodal – more than 2 modes

Determining the most appropriate measure of center:
1. For qualitative data, the mode should be used
2. For quantitative data the means should be used, unless the data set contains outliers or is skewed
3. For quantitative data sets that are skewed or contain outliers, the median should be used

Properties of the mean
1. Most familiar and widely used
2. Its value is affected by every value in the data set
3. Is not necessarily a value in the data set
4. Appropriate choice for quantitative data with no outliers

Properties of the median
1. Easy to compute by hand
2. The middle number of the ordered data set
3. Only determined by middle values of a data set, and not affected by extreme numbers
4. Useful measure of center for skewed distributions
5. Is not necessarily a value in the data set

Created by: Megan Krou
Spring 2017
STUDENT LEARNING ASSISTANCE CENTER (SLAC)
Texas State University

Properties of the mode

1. A data set does not have to have a mode
2. A data set can have more than one mode
3. If a mode exists for a data set, the mode is a value in the data set
4. Not affected by outliers in the data set
5. Only measure of center appropriate for qualitative data

Measures of Dispersion

Range- difference between the largest data value and the smallest data value

Population Variance

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$$

μ = population mean
N = population size

Excel: =var( )

Sample Variance

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

$\bar{x}$ = sample mean
n = sample size

Excel: =var( )

Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

μ = population mean
N = population size

Excel: =stdev( )

Sample Standard Deviation

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

$\bar{x}$ = sample mean
n = sample size

Excel: =stdev ( )

Created by: Megan Krou
Spring 2017
STUDENT LEARNING ASSISTANCE CENTER (SLAC)
Texas State University

Applying the Standard Deviation

Coefficient of Variation (CV)

Sample CV

$$Sample\ CV = \frac{s}{\bar{x}}$$

$\bar{x}$ = sample mean
s = sample standard deviation

Population CV

$$Population\ CV = \frac{\sigma}{\mu} * 100$$

μ = population mean
$\sigma =$ *populatin standard deviation*

Standard deviation for Grouped Data

$$s = \sqrt{\frac{n[\sum(f*x^2)] - [\sum(f*x)^2]}{n(n-1)}}$$

n = sample size
f = frequency
x = midpoint between classes

Empirical Rule (bell shaped):

Approximately 68% of data lies within 1 standard deviations of the mean
Approximately 95% of data lies within 2 standard deviations of the mean
Approximately 99.7% of the data lies within 3 standard deviations of the mean

Example:

$$number\ of\ standard\ deviations\ of\ the\ mean = \frac{x - \mu}{\sigma}$$

x = data value
μ = mean
σ = standard deviation

Chebyshev's Theorem:

The proportion of data that lies within k standard deviations of the mean is at least $1 - \frac{1}{k^2}$ for k>1

Example: when k = 2 at least $1 - \frac{1}{2^2}$ = ¾ = 75% of the data lies within 2 standard deviations of the mean.

Proportions:

Measure the fraction of a group that possesses some characteristics

Population Proportion

$$p = \frac{x}{N}$$

x = number that possesses characteristics
N = population size

Created by: Megan Krou
Spring 2017
STUDENT LEARNING ASSISTANCE CENTER (SLAC)
Texas State University

Sample Proportion

$$\hat{p} = \frac{x}{n}$$

x = number that possesses characteristics
n = sample size

Measures of Relative Position

Percentile

$$\ell = n * \frac{P}{100}$$

$\ell = location\ of\ the\ Pth\ percentile$
n = sample size

If ℓ is decimal → round to next larger integer
If ℓ is whole number → the percentile's value is the mean of the value in that location and the one in the next largest location

Step 1: form ordered array from smallest to largest
Step 2: solve equation
Step 3: ℓ rules above

Percentile of data value x

$$percentile\ of\ x = \frac{number\ of\ data\ values < x}{total\ number\ of\ data\ vales} * 100$$

Always round up

Quartiles

1. Order the data set
2. Find the median, Q_2first
3. Use the median to divide data set into 2 parts. If data set is odd, include the median in each half, if data set is even, do not include median in each half
4. Q_1 is the median of lower half
5. Q_3 is the median of upper half

Five-Number Summary

Min, Q_1, Q_2, Q_3, Max

Box plot:

Is a graphical summary of the central tendency, the spread, the skewness, and the potential existence of outliers. Constructed from the five-number summary above

Box and Whiskers Plot:

Box refers to a box that is between Q_1 and Q_3 , whiskers extend to reach the min and max

Interquartile Range

$Q_3 - Q_1$

Created by: Megan Krou
Spring 2017
STUDENT LEARNING ASSISTANCE CENTER (SLAC)
Texas State University

Outlier

If it is at least 1.5 times the interquartile range above the 75th percentile or 1.5 times the interquartile range below the 25th percentile.

Z – Score

Transforms a data value into the number of standard deviations that value is from the mean, measure of relative position, with respect to the mean and variability

$$z = \frac{x-\mu}{\sigma}$$

x = data value
μ = mean
σ = standard deviations

Excel to find area to the left of the z-value:
= normdist(x, μ,σ, 1 or 0)
1= if want everything less than and equal to x
0 = if want exactly x

Excel to find the z value given the area to left of the z value:
= normsinv(area to left of z)

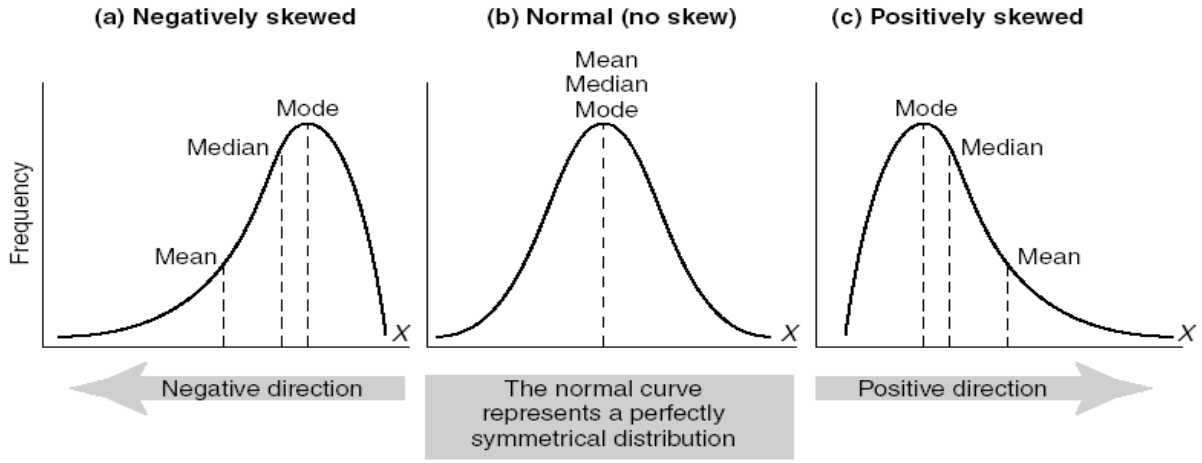

FIGURE 15.6 Examples of normal and skewed distributions

Created by: Megan Krou
Spring 2017
STUDENT LEARNING ASSISTANCE CENTER (SLAC)
Texas State University