CHEBYSHEV'S THEOREM

The __proportion__(or fraction) of any data set lying within K standard deviations of the mean is always __at__ __least__ $1-\frac{1}{K^2},\ K \geq 1$. For $K = 2$ or $K = 3$, we get the following statements:

- At least ¾ or 75% of all values lie within 2 standard deviations of the mean.
- At least 8/9 or 89% of all values lie within 3 standard deviations of the mean.

COMPARING VARIATION IN DIFFERENT POPULATIONS

When comparing __variation__ in __2__ different sets of __data__, the __standard__ deviations should be compared only if the two sets of data use the same __units__ and __scale__ and they have approximately the same __mean__.

DEFINITION

The **coefficient of variation (aka CV)** for a set of nonnegative sample or population data, expressed as a percent, describes the standard deviation __relative__ to the __mean__, and is given by the following:

Sample : $$CV = \frac{s}{\bar{x}} \cdot 100\%$$

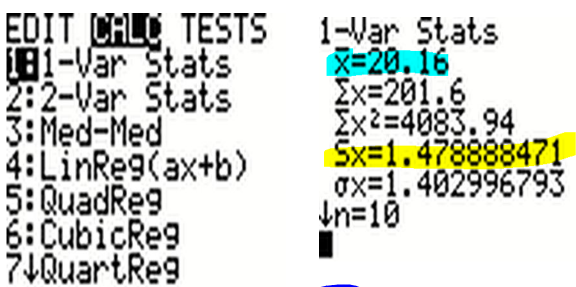
Population: $$CV = \frac{\sigma}{\mu} \cdot 100\%$$

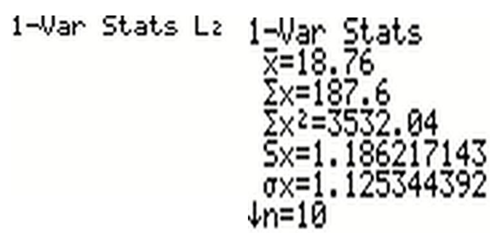
Example 3: Find the coefficient of variation for each of the two sets of data, then compare the variation.

The trend of thinner Miss America winners has generated charges that the contest encourages unhealthy diet habits among young women. Listed below are body mass indexes (BMI) for Miss America winners from two different time periods.

BMI (from the 1920s and 1930s): 20.4 21.9 22.1 22.3 20.3 18.8 18.9 19.4 18.4 19.1

BMI (from recent winners): 19.5 20.3 19.6 20.2 17.8 17.9 19.1 18.8 17.6 16.8





1920s and 1930s

$CV = \frac{s}{\bar{x}} = \frac{1.48}{20.16} \cdot 100\% = 7.3\%$

recent winners

$CV = \frac{s}{\bar{x}} = \frac{1.19}{18.76} \cdot 100\% = 6.3\%$

The CVs are only 1% apart, so one data set does not seem to vary more than the other.

3.4 MEASURES OF RELATIVE STANDING AND BOXPLOTS

BASICS OF *Z*-SCORES, PERCENTILES, QUARTILES, AND BOXPLOTS

A __z-score__ (aka standard value) is found by converting a value to a __standardize__ scale.

DEFINITION

The ***z* score (aka standard value)** is the number of __Standard__ deviations a given value x is above or below the __mean__. The z score is calculated by using one of the following:

$$\text{Sample: } z = \frac{x - \bar{x}}{s} \qquad \text{Population: } z = \frac{x - \mu}{\sigma}$$

ROUND-OFF RULE FOR Z SCORES

Round z scores to ___2___ decimal places. This rule is due to the fact that the standard table of z scores (Table A-2 in Appendix A) has z scores with two decimal places.

Z SCORES, UNUSUAL VALUES, AND OUTLIERS

In Section 3.3 we used the ___range___ ___rule___ of ___thumb___ to conclude that a value is ___unusual___ if it is more than 2 standard deviations away from the ___mean___. It follows that unusual values have z scores less than ___-2___ or greater than ___2___.

Example 1: The U.S. Army requires women's heights to be between 58 inches and 80 inches. Women have heights with a mean of 63.6 inches and a standard deviation of 2.5 inches. Find the z score corresponding to the minimum height requirement and find the z score corresponding to the maximum height requirement. Determine whether the minimum and maximum heights are unusual.

minimum height requirement

$$z = \frac{58 - 63.6}{2.5}$$

$$z \approx -2.24$$

unusual ✓

maximum height requirement

$$z = \frac{80 - 63.6}{2.5}$$

$$z \approx 6.56$$

unusual ✓

DEFINITION

Percentiles are measures of ___location___, denoted ___$P_1, P_2, P_3, \ldots, P_{99}$___, which divide a set of data into ___100___ groups with about ___1%___ of the values in each group. The process of finding the percentile that corresponds to a particular data value x is given by the following:

$$\text{Percentile of } x = \frac{\text{\# of values less than } x}{\text{total \# of values}} \cdot 100$$

NOTATION

n total # of values in a sample

k percentile being used $L = \frac{k}{100} \cdot n$

L locator that gives us the position of a value
- If L is a whole number, use the position found and the next one up and then average the 2 numbers
- If L is a decimal, round up and use the # in that position.

P_k Kth percentile

Example 2: Use the given sorted values, which are the number of points scored in the Super Bowl for a recent period of 24 years.

36 37 37 39 39 41 43 44 44 47 50 53 54 55 56 56 57 59 61 61 65 69 69 75

a. Find the percentile corresponding to the given number of points.

i. 65 Percentile of $65 = \frac{20}{24} \cdot 100 = 83 \rightarrow \boxed{P_{83} = 65}$

ii. 41 Percentile of $41 = \frac{5}{24} \cdot 100 = 21 \rightarrow \boxed{P_{21} = 41}$

b. Find the indicated percentile or quartile.

i. $Q_1 = P_{25} \rightarrow L = \frac{25}{100} \cdot 24 \rightarrow L = 6$, so we average the scores in positions 6 and 7

$\frac{41+43}{2} = 42$ $\boxed{Q_1 = 42}$

ii. $P_{80} \rightarrow L = \frac{80}{100} \cdot 24 \rightarrow L \approx 19.2$

so $L = 20 \rightarrow \boxed{P_{80} = 61}$

iii. $P_{95} \rightarrow L = \frac{95}{100} \cdot 24$

$L = 22.8 \rightarrow L = 23 \rightarrow \boxed{P_{95} = 69}$

DEFINITION

median

Quartiles are measures of location, denoted Q_1, Q_2, Q_3, which divide a set of data into 4 groups with about 25% of the values in each group.

FIRST QUARTILE:

separates the bottom 25% from the top 75%

SECOND QUARTILE:

separates the bottom 50% from the top 50%

THIRD QUARTILE:

separates the bottom 75% from the top 25%

DEFINITION

For a set of data, the **5-number summary** consists of the minimum value, the first quartile, the median (aka second quartile), the third quartile, and the maximum value.

A **boxplot (aka box-and-whisker diagram)** is a graph of a data set that consists of a line extending from the minimum value to the maximum value, and a box with lines drawn at the first quartile, the median, and the third quartile.

OUTLIERS

When ___analyzing___ data, it is important to ___identify___ and ___consider___ outliers because they can strongly affect values of some important statistics, such as the __mean__ and __standard__ __deviation__.

In __modified__ __boxplot__, a data value is an __outlier__ if it is...

above quartile 3 by an amount greater than 1.5 x inner quartile range or **below quartile 1 by an amount greater than 1.5 x inner quartile range**

________________________ ________________________ are called __________________________ or __________________________ boxplots, which represent _________________________________ as special points. A **modified boxplot** is a boxplot constructed with these modifications: (1) A special symbol, such as an ________________________ or point is used to identify _______________________ and (2) the solid horizontal line extends only as far as the minimum and maximum values which are not outliers.

Example 3: Use the given sorted values, which are the number of points scored in the Super Bowl for a recent period of 24 years to construct a boxplot. Are there any outliers?

36 37 37 39 39 41 43 44 44 47 50 53 54 55 56 56 57 59 61 61 65 69 69 75

Outlier check: