Applied Statistics

Domain	**Exploring Data**	
Cluster	**Select appropriate graphical and numerical methods to explore data.**	
Standard(s)	M.ASHS.2	Calculate appropriate measures of center, variability, and position for data. Instructional note: Include comparisons of mean vs. median, standard deviation vs. IQR.

Content Examples

- Finding mean, median, and mode – measures of central tendency: *https://www.youtube.com/watch?v=eMPspP_Rkio*
- Calculating mean, standard deviation and variance for a sample: *https://www.youtube.com/watch?v=qqOyy_NjflU*
- How to find the interquartile range and any outliers: *https://www.youtube.com/watch?v=STSP8gTSdT8*
- 1 Variable Stats on TI-84: *https://www.youtube.com/watch?v=hf8ziCMsLvs*

Relevant Content

Vocabulary

- Mean: Sum of the values divided by the total number of values
- Median: Midpoint of the data array, symbolized as MD
- Mode: The value that occurs most often in the data set
- Unimodal: A data set with only one value that occurs with the greatest frequency
- Bimodal: A data set with only two values that occurs with the same greatest frequency
- Multimodal: A data set with more than two values that occurs with the same greatest frequency
- Midrange: The sum of the lowest and highest values in the data set divided by two, symbolized as MR
- Weighted mean: Type of mean that considers additional factors when the values are not all equally represented
- Positively or right-skewed shape: The majority of the data values are left of the mean and cluster at the lower end of the distribution, with the tail to the right
- Negatively or left-skewed shape: The majority of the data values are right of the mean and cluster at the upper end of the distribution, with the tail to the left

- Symmetrical distribution: Data values are evenly distributed on both sides of the mean
- Range: The difference between the maximum and minimum values
- Variance: The average of the squares of the distance each value is from the mean, symbolized as σ^2
- Standard deviation: The square root of the variance, symbolized as σ
- Coefficient of variation: A statistic used to compare standard deviations when the units are different, symbolized as $CVar$
- Percentiles: Divides the data set into 100 equal groups
- Quartiles: Division of the distribution into four groups, separated by Q_1, Q_2, & Q_3, where Q_1 is the 25th percentile, Q_2 is the 50th percentile or the median, and Q_3 is the 75th percentile
- Interquartile range (IQR): The range of the middle 50% of data (difference between Q_1 and Q_3)
- Deciles: Division of the distribution into 10 groups, separated by D_1, D_2, D_3, . . . , D_9, where D_1, D_2, D_3, . . . , D_9 corresponds to P_{10}, P_{20}, P_{30}, . . . , P_{90}
- Outliers: An extremely high or low data value when compared to the rest of the data values

Formulas

- Sample mean: $\bar{x} = \frac{\sum x_i}{n}$, where n is the total number of values in the sample
- Population mean: $\mu = \frac{\sum x_i}{N}$, where N is the total number of values in the population
- Mean for grouped data: $\bar{x} = \frac{\sum f \cdot x_m}{n}$, where $\sum f \cdot x_m$ represents the sum of the product of the frequency (f) and the midpoint (x_m) for each class
- Midrange: $MR = \frac{x_{\min} + x_{\max}}{2}$
- Weighted mean: $\bar{x} = \frac{\sum wx_i}{\sum w}$, where $w_1, w_2, ..., w_n$ are the corresponding weights of the values $x_1, x_2, ..., x_n$.
- Range: $R = x_{\max} - x_{\min}$
- Variance: $\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$, where x_i is an individual value, μ is the population mean, and N is the population size
- Standard deviation: $\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$, where x_i is an individual value μ is the population mean, and N is the population size
- Sample variance: $s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$, or alternately as $s^2 = \frac{n\left(\sum x_i^2\right) - \left(\sum x_i\right)^2}{n(n-1)}$, where x_i is an individual value, $\bar{x}$ is the sample mean, and n is the sample size
- Sample standard deviation: $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$, or alternately as $s = \sqrt{\frac{n\left(\sum x_i^2\right) - \left(\sum x_i\right)^2}{n(n-1)}}$, where x_i is an individual value, $\bar{x}$ is the sample mean, and n is the sample size
- Coefficient of variation for populations: $CVar = \frac{\sigma}{\mu} \cdot 100\%$
- Coefficient of variation for samples: $CVar = \frac{s}{\bar{x}} \cdot 100\%$
- Range Rule of Thumb: $s \approx \frac{x_{\max} - x_{\min}}{4}$
- Percentile Formula: $Percentile = \frac{(number\ of\ values\ below\ x_i) + 0.5}{total\ number\ of\ values} \cdot 100\%$

Measures of Central Tendency

» The Mean
- Is found by using all the values of the data
- Is more consistent than the median or mode when samples are taken from the same population
- Is used in computing other statistics, such as the variance
- Is unique and not necessarily one of the data values
- Is not used for the data in a frequency distribution that has an open-ended class
- Is affected by extremely high or low outliers, and may not be the appropriate average to use in these situations

» The Median
- Is used to find the center or middle value of a data set
- Is used when it is necessary to find out whether the data values fall into the upper half or lower half of the distribution
- Is used for an open-ended distribution
- Is affected less than the mean by extremely high or extremely low values

» The Mode
- Is used when the most typical case is desired
- Is the easiest average to compute
- Is an appropriate measure when the data are nominal, such as religious preference, gender, or political affiliation
- Is not always unique

» The Midrange
- Is easy to compute
- Is the midpoint
- Is affected by extremely high or low values in a data set

Variance and Standard Deviation

» Are used to determine the spread of the data

» Are used to determine the consistency of a variable

» Are used to determine the number of data values that fall within a specified interval

» Are used to compare two or more data sets to determine which is more variable

Identifying Outliers

» Arrange the data in order and find Q_1 & Q_3

» Find IQR

» Check the data set for any data value that is smaller than $Q_1 - 1.5(IQR)$ or larger than $Q_3 + 1.5(IQR)$

Measures of Center and Standard Deviation:
https://www.learner.org/series/against-all-odds-inside-statistics/

Assessment Links or Tasks

Analyzing Box Plots: *https://www.cpalms.org/Public/PreviewResourceLesson/Preview/128715*