Statistics

- Measures of the Center
- Measures of the Spread
- Measures of Relative Standing

Mean

- The mean (often thought of the average) is defined by

Population Mean: $\mu = \dfrac{1}{n}\sum x$

Sample Mean: $\bar{x} = \dfrac{1}{n}\sum x$

Median, Mode, and Midrange

- Median: The middle number from the data or the mean of the middle two.
- Mode: The data value(s) that occurs the most frequently.
- Midrange: The average of the minimum and maximum values.

Weighted Mean

- Weighted Mean: $x = \dfrac{\sum wx}{\sum w}$

Example: Your grade is based on : Homework = 100, Quizzes = 100

Exams = 600, Projects = 200

Suppose your homework average is 95%, your quiz average is 84%, your exam average is 78%, and your project average is 98%. What is your course grade (based on the weighted average)?

Variance and Standard Deviation

- The standard deviation is a measure of the spread of the data. It is defined by

$$\text{Sample Standard Deviation: } s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

$$\text{Population Standard Deviation: } \sigma = \sqrt{\frac{\sum(x - \bar{x})^2}{n}}$$

- The variance is the square of the standard deviation.

Rule of Thumb

- If we take the data and throw away the outliers, the "usual values" are what's left. The maximum and minimum of the usual values are approximately

$$\text{Minumum usual value} \approx \overline{x} - 2s$$

$$\text{Maximum usual value} \approx \overline{x} + 2s$$

Estimating the Standard Deviation

- To estimate the standard deviation, remove the outliers and divide the range by 4.

The Empirical Rule

- About 68% of the data lie within 1 standard deviation of the mean.

- About 95% of the data lie within 2 standard deviations of the mean.

- About 99.7% of the data lie within 3 standard deviations of the mean.

- Note: The empirical rule works best when the data is approximately normal.

Chebyshev's Theorem

The proportion of any set of data lying within K standard deviations of the mean is always at least $1 - 1/K^2$.

- At least 75% of all values lie within 2 standard deviations of the mean.

- At least 89% of all values lie within 3 standard deviations of the mean.

Coefficient of Variation

- Coefficient of Variation

$$CV = \frac{s}{\overline{x}} \cdot 100\%$$

z scores

- The z score is defined by

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

It tell us the number of standard deviations from the mean.

Percentiles

- The percentiles divide the data into 100 groups with about 1% of the data in each group.

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

- The first quartile (Q1) is the 25th percentile and the third quartile is the 75th percentile.
- The Interquartile Range (IQR) is the difference between the first and third quartiles

Boxplots

We construct a boxplot as follows

- Find the five point summary: Minimum, Q1, Median, Q3, and Maximum.

- Put horizontal bars at each of these values.

- Connect Q1 and Q3 to make a box.

- Draw lines from the Min to Q1 bars and from the Max to Q3 bars to make the whiskers.