Cheat Sheet: Describing Data Numerically

Essential Concepts

- The median of a data set can be computed by ordering the data values and identifying the value in the middle.
- The mean of a data set can be computed by finding the sum of the data values and dividing it by the number of data values in the data set.
- The mean represents the balance point of the data, and the median represents the 50th percentile, or the value that splits the data in half.
- When a distribution is symmetric, the mean and median occupy the same value. Under a skew, the mean is “pulled” in the direction of the outliers:
    - Right-skewed: the mean is greater than the median
    - Left-skewed: the mean is less than the median
- The median stays relatively fixed in a data set if one value changes by a large amount, the mean does not. This is an indication that the mean is sensitive to the presence of extreme values in the data set and can be a misleading indicator of a “typical” observation value
- A boxplot is a data display specifically designed to show something called the "five-number summary" which divides a data set into four equal sections. The boxplot has distinct points at the median, quartiles, and minimum and maximum of the data set.
- $Q1$, the lower quartile, represents the boundary of the first quarter of the data and $Q3$, the upper quartile, represents the boundary for the last quarter of the data. The IQR is calculated as $Q3 - Q1$ and describes the spread of a boxplot.
- Interquartile range (IQR) is the best method for determining if an observation is an outlier in the distribution. This fences, or boundaries, for the upper and lower outliers value equals either the distance $1.5(IQR)$ less than $Q1$ or greater than $Q3$

- Variability refers to a measure of how spread out the data in a data set is. Variability is measured using standard deviation, variance, IQR, and range.
- In a histogram, variability can be judged by examining the distance of the bars from the statistical center (mean or median) of the graph. If the variability is high, equally sized or taller bars will appear away from the center of the graph. If the variability is low, the data will appear clustered around the center
- Larger values of range indicate more variability in the data, but the range value only utilizes two observations in the entire data set to measure variability. This is not an ideal measure of spread, but when used in combination with other measures of spread, it can help you gain a clearer understanding of the spread of a distribution
- The standard deviation is the "typical" or average distance of each data point to the mean of the data set.
- Standard deviation is generally calculated with technology, but the following steps can be applied to calculate a standard deviation by hand:
    - Calculate the mean of the population or sample
    - Take the difference between each data value and the mean, then square each difference
    - Add up all the squared differences
    - Divide by either the total number of observations in the case of a population, or by 1 fewer than the total in the case of a sample
    - Take the square root of the result from step 4
- Standard deviation is the square root of the variance of a data set.
- Similar to median, the IQR is considered a more accurate measure of spread for data that is skewed or contains outliers. Alternatively, the mean and standard deviation are considered more accurate measures when the data is symmetric because they utilize all data points as opposed to just one or two measures.
- Standardizing the value includes finding the difference between the given value and the mean, and dividing that distance by the standard deviation. The resulting value is a number of standard deviations, and has no units associated with it.
- Standardized scores, called z-scores for the standard normal distribution, can result in positive and negative values. A negative indicates value less than the mean and a positive indicates a value that is greater than the mean.
- If a distribution is bell shaped, unimodal, and symmetric, the Empirical Rule states that:
    - about 68% of observations in a data set will be within one standard deviation of the mean
    - about 95% of observations in a data set will be within two standard deviations of the mean

- about 99.7% of the observations in a data set will be within three standard deviations of the mean

---

Key Equations

converting values into standardized scores

$z = \frac{x-\mu}{\sigma}$, where x represents the value of the observation, μ represents the population mean, σ represents the population standard deviation, and z represents the standardization value, or z-score

deviation from the mean

$(x - \bar{x})$, where x is the observation in the dataset, and $\bar{x}$ is the sample mean

interquartile range (IQR)

$Q3 - Q1$

lower outlier "fence" or boundary

$Q1 - 1.5(IQR)$, remember to multiply 1.5 by IQR first, and then subtract from $Q1$

mean

$\frac{sum\ of\ data\ values}{total\ number\ of\ data\ values}$ or $\bar{x} = \frac{\sum x}{n}$, where $\bar{x}$ is the mean, $\sum x$ is the symbol for "sum of", x represents the data values, and n is the total number of data values

standard deviation of a population

$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{n}}$, where $\sum$ is the summation of $(x - \mu)^2$ for each observation, (x) is the observation in the dataset, (μ) is the mean, and (n) is the number of observations

standard deviation of a sample

$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$, where $\sum$ is the summation of $(x - \bar{x})^2$ for each observation, (x) is the observation in the dataset, $(\bar{x})$ is the mean, and (n) is the number of observations

upper outlier “fence” or boundary

$Q3 + 1.5(IQR)$, remember to multiply 1.5 by IQR first, then add to $Q3$

variance of a population

$\sigma^2 = \frac{\sum(x-\mu)^2}{n}$, where $\sum$ is the summation of $(x - \mu)^2$ for each observation, (x) is the observation in the dataset, (μ) is the mean, and (n) is the number of observations

variance of a sample

$s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$, where $\sum$ is the summation of $(x - \bar{x})^2$ for each observation, (x) is the observation in the dataset, ($\bar{x}$) is the mean, and (n) is the number of observations

---

Glossary

s

the standard deviation of a sample of observations

σ

the standard deviation of a population of observations

s^2

the variation of a sample of observations

σ^2

the variance of a population of observations

deviation from the mean

the distance between an observation (x) in a data set and the mean ($\bar{x}$) of the data set

Empirical Rule

a guideline that predicts the percentage of observations within a certain number of standard deviations. Also known as the 68-95-99.7 Rule which states that in a bell-shaped, unimodal distribution, almost all of the observed data values, x, lie within three standard deviations, σ, to either side of the mean, μ. More specifically, about 68% of observations in a data set will be within one standard deviation of the mean $(\mu \pm \sigma)$, about 95% of the observations in a data set will be within two standard deviations of the mean $(\mu \pm 2\sigma)$, and about 99.7% of the observations in a data set will be within three standard deviations of the mean $(\mu \pm 3\sigma)$

first quartile

the value below which one quarter of the data lies, also equal to the 25th percentile; sometimes denoted $Q1$

five-number summary

the collection of the minimum, first quartile, median, third quartile, and maximum of the variable

interquartile range

the quantity $Q3 - Q1$, sometimes denoted IQR

left-skewed (negative skew)

when most of the data is bunched up to the right of the graph with a “tail” of infrequent values on the left (lower) end of the distribution

lower outlier

an observation that is less than $Q1 - 1.5(IQR)$

mean

an average of a set of values calculated by adding the values and then dividing the total by the number of values in the dataset

median

the “middlemost” value of a set of values listed in numerical order

outlier

an unusual or extreme value, given the other values in the dataset

range

the maximum (or largest) value – the minimum (or smallest) value

resistant

not affected by the skewness of a graph

right-skewed (positive skew)

when most of the data is bunched up to the left of the graph with a “tail” of infrequent values on the right (upper) end of the distribution

standard deviation

a measure of how spread out observations are from the mean

standardized value

the number of standard deviations an observation is away from the mean; also referred to as a z-score

symmetric

the left and right sides of the distribution (closely) mirror each other -- if you drew a vertical line down the center of the distribution and folded the distribution in half, the left and right sides would closely match one another

third quartile

the value below which three quarters of the data lay, also equal to the 75th percentile; sometimes denoted as $Q3$

upper outlier

an observation that is greater than $Q3 + 1.5(IQR)$

variability

a measure of how dispersed (spread out) the data are; often referred to as the spread, or dispersion, of a data set

variance

the standard deviation squared