Summarizing Numerical Data
Descriptive Statistics

DASC 512

Overview

- Central Tendency
- Skewness
- Variability
- Quantiles
- Parametric and Nonparametric interpretations

Measures of Central Tendency

<u>Central Tendency</u>: the "center" or "typical" value of a variable

- Mean
- Median
- Mode

Mean

Let each observation in a quantitative data set be represented by the variable x. Then a set of n observations is represented

$$x_1, x_2, x_3, \dots, x_n$$

The mean is the value at the center of mass for those observations.

$$\bar{x} = \frac{1}{n} \times \sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Median

If we sort those observations, the set of ordered observations is represented

$$x_{(1)}, x_{(2)}, x_{(3)}, \dots, x_{(n)}$$

The <u>median</u>, or 50th percentile, is the middle observation when sorted

- If there are an even number of observations, it is the mean of the middle two

$$\text{median} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd} \\ \dfrac{\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right)}{2} & \text{if } n \text{ is even} \end{cases}$$

Mode

The mode is the most frequently occurring observation in the data

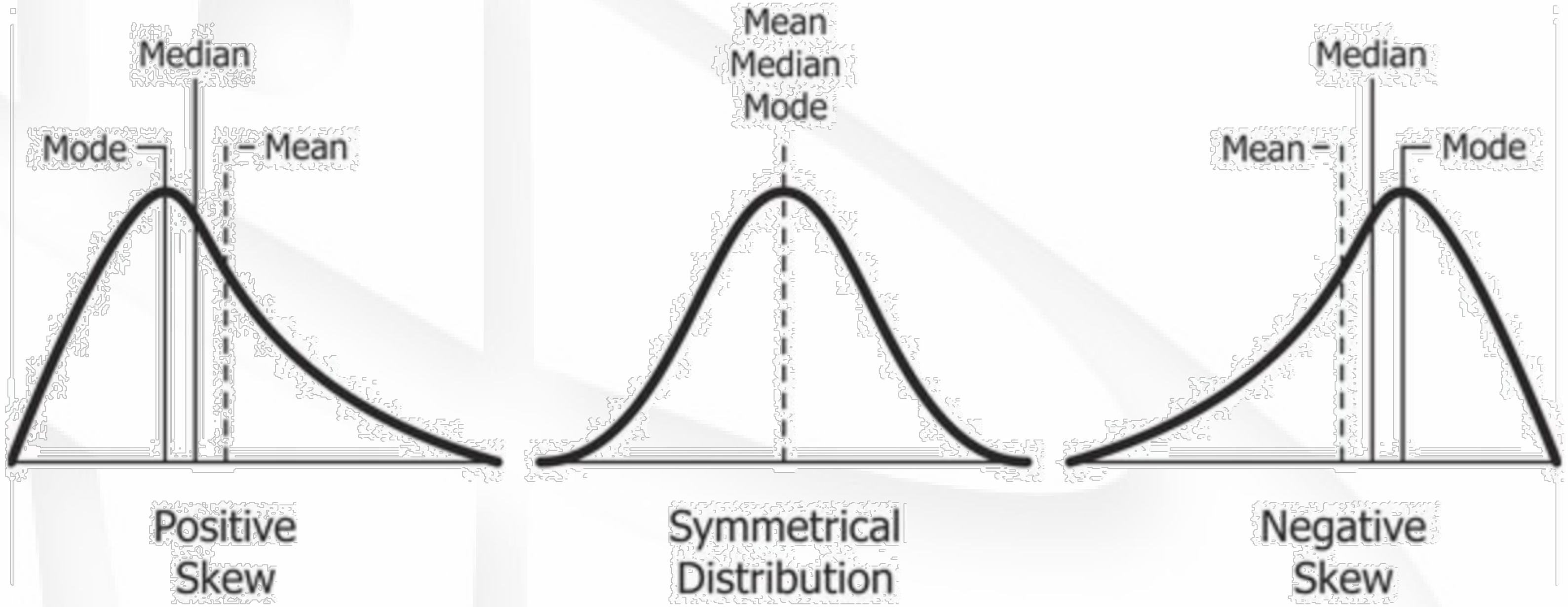
There can be multiple modes in a dataset.

Skewness

Skewness: describes the shape of the data.

- Right/Positive skew indicates a right tail – mean > median > mode
- Symmetric indicates equal tails – mean = median = mode
- Left/Negative skew indicates a left tail – mean < median < mode

"The tail tells the tale"



Measures of Variability

Variability: the spread of the data

- Range
- Inter-Quartile Range
- Variance
- Standard Deviation

Range

<u>Range</u> is the difference between the largest and smallest observations

$$\text{Range} = \max\{x_1, x_2, \dots, x_n\} - \min\{x_1, x_2, \dots, x_n\}$$

This can be a useful value, but it increases with sample size.

Inter-Quartile Range (IQR)

A quantile is one part of equal partitions of the data

- Quartiles divide the data into four parts, each containing 25% of the data
- Percentiles divide the data into 100 parts, each containing 1% of the data

Commonly used quantiles include:

- 5^{th} and 95^{th} Percentile
- Q_1 and Q_3: First and third quartile
- Median or Q_2: Second quartile

Inter-Quartile Range (IQR)

The IQR is the difference between the third and first quartiles

$$IQR = Q_3 - Q_1$$

Variance

Population Variance is the mean squared deviation from the mean.
If you perform a census, you can calculate this directly.

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

Sample Variance is an "unbiased estimator" of the population variance.
This must account for a lost degree of freedom.

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

Standard Deviation

Standard Deviation is the square root of the Variance. For samples,

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

For a census,

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

Interpreting Descriptive Statistics

Depends on the assumptions that you are willing to make

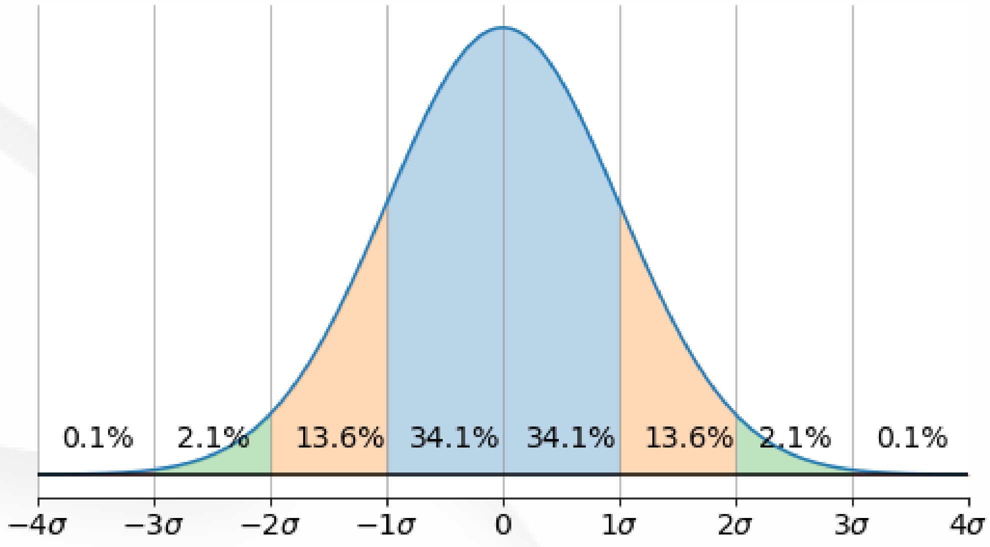
- Parametric statistics assume some distribution (typically normality)
  - Empirical rule
- Nonparametric statistics make no distributional assumptions
  - Chebyshev's Rule

Later we'll get much more inference using both parametric and nonparametric hypothesis tests and confidence intervals.

Empirical Rule

- If we know that a dataset is unimodal and symmetric, we can approximate tighter bounds on the data by assuming normality
  - About 68% of observations will fall within 1 standard deviation
  - About 95% of observations will fall within 2 standard deviations
  - About 99.7% of observations will fall within 3 standards deviations

- This is the origin of “Six Sigma”

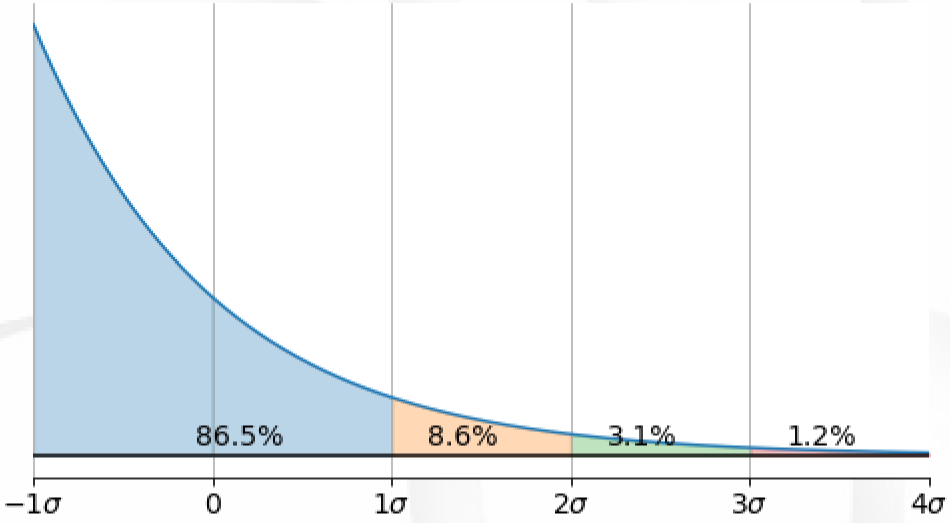


Chebyshev’s Rule

Even without making any distributional assumptions,

- At least 75% of observations will fall within 2 standard deviations of the mean
- At least 8/9 of observations will fall within 3 standard deviations of the mean
- For k > 1, $\left(1 - \frac{1}{k^2}\right)$ observations will fall within k standard deviations of the mean

These are very wide bounds, but useful for strange distributions.



→ $0.95 > \frac{3}{4}$ within 2 standard deviations

→ $0.98 > \frac{8}{9}$ within 3 standard deviations

→ $0.99 > \frac{15}{16}$ within 4 standard deviations

Recap

- Central Tendency
- Skewness
- Variability
- Quantiles
- Parametric and Nonparametric interpretations