Descriptive / Summary Statistics

Topics:

1. Descriptive Statistics:
   a. Measures of central tendency (mean, median, mode)
   b. Measures of dispersion (variance, standard deviation, range)
   c. Percentiles and quartiles
   d. Skewness and kurtosis

Descriptive Statistics/ Summary Statistics

Data can be presented in many different formats but, what are the main characteristics that describe the data set?

Descriptive statistics: summarize, organize, and present data concisely.

- They communicate something about the dataset without needing to understand the whole thing
- Summarize known data in a way that can be used for further predictions and analysis.
- Very useful for quickly understanding what’s going on
- Examples: Mean, median, mode

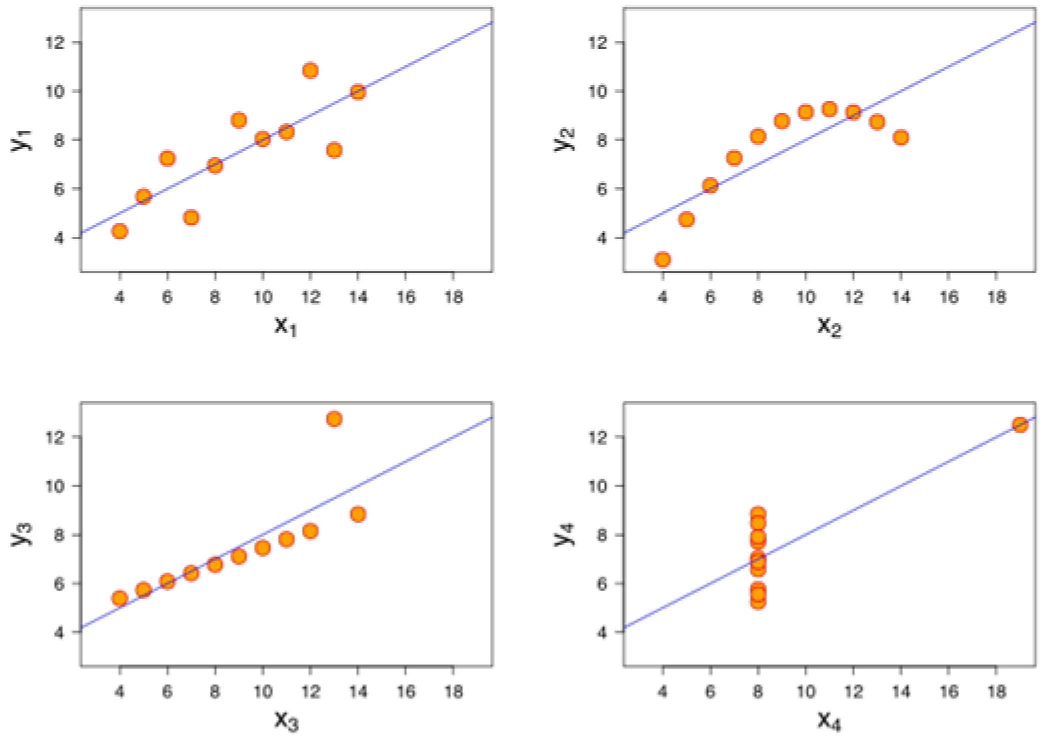
How not to use summary statistics

Do NOT take your dataset, compute the mean or median, and then call it a day. If you ever do this I will pop out of your computer and shake you

If you do not understand your data ahead of computing the summary statistics, they can end up being useless, or actively misleading.

Why not to rely on summary statistics

All four of these have the same mean and variance, but are clearly generated by four different processes.



Summary stats are for *after* you understand your data holistically (though you can use them to help that process)

Types of Descriptive/ Summary Statistics

Part of descriptive statistics, used to summarize data:

- Convey lots of information with extreme simplicity

- **Measures of central tendency/The location of the data**
  - e.g. mean, median, mode, etc.
- **The shape of the data:**
  - Measures of Skewness (asymmetry)
  - Modality
- **The spread of the data/ Measure of dispersion:**
  - Measures of variability (spread)
  - e.g. range, variance, etc…
- **Associations between variables: (LATER TOPIC)**
  - Understanding correlation, covariance etc.
  - Measuring correlation
  - Scatter plots and regression

Measures of Central Tendency

Measures of Central Tendency

Measures of Central Tendency tell you about the center of your distribution. It represents the whole set of data by a single value. These include:

- **The Pythagorean means**
  - **Arithmetic Mean**
  - **Geometric Mean**
  - **Harmonis Mean**
- **Median**
- **Mode**

Measures of Central Tendency

Let y denote a quantitative variable, with observations y1 , y2 , y3 , … , yn

a. Describing the center

Median: Middle measurement of ordered sample.

- The value in the dataset that has an equal number of items greater than and less than it.

Mean: Average of all observations in the dataset.

Mode: The most common item in your dataset

MEASURES OF LOCATION

Purpose: Identify the "center" of a distribution of values.

```
These are 30 hours of average defect data on sets of
circuit boards. Roughly what is the typical value?

1.45  1.65  1.50  2.25  1.65  1.60  2.30  2.20  2.70  1.70
2.35  1.70  1.90  1.45  1.40  2.60  2.05  1.70  1.05  2.35
1.90  1.55  1.95  1.60  2.05  2.05  1.70  2.30  1.30  2.35
```

Location and central tendency

- There exists a distribution of values
- We are interested in the “center” of the distribution

Two measures are the sample mean and the sample median.

- **They look similar, and measure the same thing**
- **They differ systematically (and predictably) when the data are not symmetric.**

THE MEAN OF AGGREGATE DATA

State	Listing	IncomePC	State	Listing	IncomePC	State	Listing	IncomePC
Hawaii	896800	24057	Rhode Island	432534	22251	Texas	266388	19857
California	713864	22493	Delaware	420845	22828	Mississippi	255774	15838
New York	668578	25999	Oregon	417551	20419	Tennessee	255064	19482
Connecticut	654859	29402	Idaho	415885	18231	Wisconsin	243006	21019
Dist.Columbia	577921	31136	Illinois	377683	23784	Michigan	241107	22333
Nevada	549187	24023	New Hampshire	361691	23434	Missouri	221773	20717
New Jersey	529201	23038	New Mexico	358369	17106	South Dakota	220708	19577
Massachusetts	521769	25616	Vermont	346469	20224	West Virginia	219275	17208
Wyoming	499674	20436	South Carolina	340066	17695	Arkansas	217659	16898
Maryland	480578	24933	North Carolina	330432	19669	Ohio	209189	20928
Utah	475060	17043	Georgia	326699	20251	Kentucky	208391	17807
Colorado	467979	22333	Alaska	324774	23788	Oklahoma	203926	17744
Arizona	448791	19001	Minnesota	306009	22453	Kansas	201389	20896
Florida	447698	21677	Maine	299796	19663	Indiana	200683	20378
Montana	446584	17865	Pennsylvania	295133	22324	Iowa	184999	20265
Virginia	443618	22594	Louisiana	280631	17651	North Dakota	173977	18546
Washington	440542	22610	Alabama	269135	18010	Nebraska	164326	20488

Average list price:
1/51 (\$898,800 + \$713,864 + … + \$164,326) = \$369,687

AVERAGING AVERAGES?

Calculating an overall average by simply taking the average of individual averages from different groups or categories, without considering the sizes of the groups (e.g., population size, sample size, etc.

Hawaii's average listing	= $896,800
Hawaii's population	= 1,275,194
Illinois' average listing	= $377,683
Illinois' population	= 12,763,371

When you average averages without considering population size, you can get misleading results.

Illinois and Hawaii each get an equal weight of 1/51 = .019607 when the mean is computed. Looks like Hawaii is getting too much influence …

WEIGHTED AVERAGE

Solution: Use a weighted average when observations have different levels of importance (e.g., populations in this case).

Weighted Average Calculation Example

- Hawaii's Average Listing Price = $896,800
- Illinois' Average Listing Price = $377,683
- Hawaii's Population = 1,275,194
- Illinois' Population = 12,763,371

Step 1: Calculate the weights

- Total Population = 1,275,194 + 12,763,371 = 14,038,565
- Weight for Hawaii = $\frac{1{,}275{,}194}{14{,}038{,}565} = 0.0908$
- Weight for Illinois = $\frac{12{,}763{,}371}{14{,}038{,}565} = 0.9092$

Step 2: Weighted Average Formula

$$\text{Weighted Average} = \frac{(896,800 \times 0.0908) + (377,683 \times 0.9092)}{1}$$

$$\text{Weighted Average} = 409,234$$

New average is $409,234 compared to $369,687 without weights, an error of 11%

THE SAMPLE MEDIAN

Median:

- Sort the data
- Take the middle point*

Odd number:

- Central observation: Med[1,2,4,6,8,9,17]

Even number:

- Midpoint between the two central observations
  Med[1,2,4,6,8,9,14,17] = (6+8)/2=7

WHAT IS THE CENTER?

The mean and median measure the central tendency of data

- Generally, the center of a dataset is a point that is close to most of the data points
- Close? Need a distance metric between two points x and x_2
  - Absolute deviation: $| x_1 – x_2 |$
  - Squared deviation: $(x_1 – x_2)^2$

We'll define the center based on these metrics

- Median minimizes the sum of absolute deviations.
- Mean minimizes the sum of squared deviations.

Next

Pythagorean Means

Pythagorean Means

Pythagorean means refer to **three separate concepts:** the arithmetic mean, the geometric mean, and the harmonic mean.

- These means are named after the Pythagorean theorem because they are all forms of averaging.
- Provide different ways of summarizing a set of numbers.

Pythagorean Means

- **Arithmetic Mean**: Your typical average
  - **Sensitive to Outliers:** Affected by extreme values.
    - **Equal Weight:** Treats all values equally in calculation, regardless of its distance from the center of the dataset.

$$\mathrm{AM}(x_1, \ldots, x_n) = \frac{x_1 + \cdots + x_n}{n}$$

Pythagorean Means

Geometric Mean: A measure of central tendency that is particularly useful for data that are multiplicative or exponential in nature, such as growth rates, ratios, or percentages.

- Multiply all the values in a dataset and then take the nth root of the product (n = number of values).

$$\mathrm{GM}(x_1, \ldots, x_n) = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}} = \sqrt[n]{x_1 \cdot x_2 \cdots x_n}$$

Example:

- **Data**: Annual growth rates of an investment over 3 years: 1.10, 1.15, 1.20.
- **Calculation**:

$$\text{Geometric Mean} = (1.10 \times 1.15 \times 1.20)^{\frac{1}{3}} \approx 1.149$$

- **Interpretation**: The average annual growth rate is approximately **14.9%**.

Geometric Means is Less sensitive to outliers

- Why? It considers the relative magnitude of values, extreme values have less influence on the geometric mean)

Example: Consider two datasets of investment returns (in percentage):

Dataset A: 2%, 4%, 6%, 8% Dataset B (with an outlier): 2%, 4%, 6%, 100%

C

1. **Arithmetic Mean (sensitive to outliers):**
   - Dataset A: (2% + 4% + 6% + 8%) / 4 = 5%
   - Dataset B: (2% + 4% + 6% + 100%) / 4 = 28%
2. **Geometric Mean (less sensitive to outliers):**
   - Dataset A: $\sqrt[4]{2 \times 4 \times 6 \times 8} \approx 4.76$
   - Dataset B: $\sqrt[4]{2 \times 4 \times 6 \times 100} \approx 6.94$

Observe: In Dataset B, the geometric mean is less affected by the extreme 100% value compared to the arithmetic mean.

This is because the geometric mean looks at the product of the values (considering their relative sizes) rather than just their sum.

Pythagorean Means: (Harmonic Mean)

- **Harmonic Mean**: A measure of central tendency used for data involving rates, ratios, or reciprocals.
  - Calculated by taking the reciprocal of the arithmetic mean of the reciprocals of all the values in a set.

$$\text{Harmonic Mean} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}}$$

Harmonic Means

- Used to calculate the mean of rates or ratios.
- Ideal for averaging rates or inverse values, such as speeds, work rates, or densities.
- Ex: Investment Returns: when calculating the average of multiple ratios or fractions, such as price-to-earnings ratios in finance.

- **Less sensitive to outliers:** Gives more weight to smaller values, providing robustness against extremely large values.
    - **Undefined for datasets containing zero values.**

Harmonic Means: Examples

Calculating Average Speed: Consider the following example where a car travels over equal distances at different speeds:

- **First half of the trip**: 60 km/h
- **Second half of the trip**: 40 km/h

Try Arithmetic Mean = (60+40)/2=50 km/h does not accurately reflect the total time taken for the trip because it **assumes equal weighting by speed rather than distance**.

Try Geometric Mean= √(60*40)/2 ~ 48.99 km/h: handle proportional growth rates better than the arithmetic mean, it still **does not correctly address the relationship between speed and time over equal distances**.

Solution: Harmonic Mean =2/ (1/60 +1/40) = 2/ (2/120 + 3/120)=2/(5/120)= (2×120)/5 =48 km/h

The harmonic mean accurately reflects the average speed over the entire trip because **it takes into account the time spent traveling at each speed**. It gives <u>more weight to lower speeds</u>, which is important when averaging rates where the time spent at each rate matters.

2. Measures of Skewness: Other Descriptors

The shape of the data

SKEWNESS

Extreme observations distort means but not medians.

Outlying observations distort the mean:

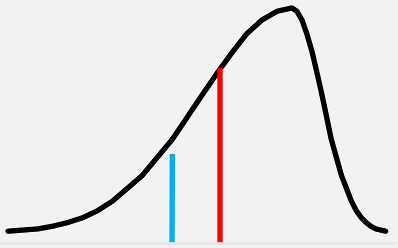
- Med [1,2,4,6,8,9,17] = 6
- Mean[1,2,4,6,8,9,17] = 6.714
- Med [1,2,4,6,8,9,17000] = 6 (still)
- Mean[1,2,4,6,8,9,17000] = 2432.8 (!)

Typically occurs when there are some outlying observations, such as in cross sections of income or wealth and/or when the sample is not very large.

SKEWED DATA

Left-Skewed

Mean < Median

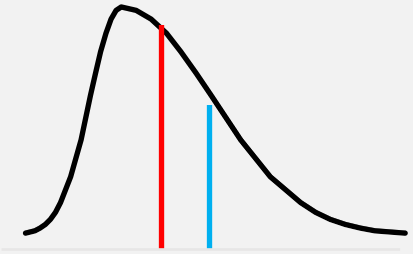


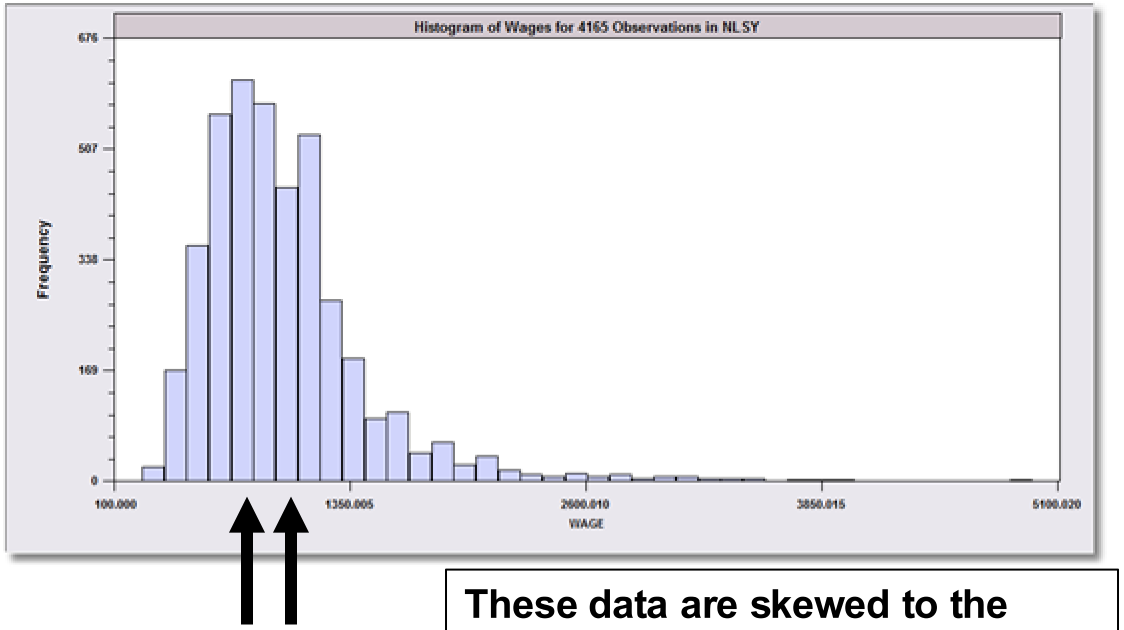
Symmetric

Mean = Median

Right-Skewed

Median < Mean





Median Mean

These data are skewed to the right.

Monthly Earnings
N = 595,
Median = 800
Mean = 883

The mean will exceed the median when the distribution is skewed to the right.

Skewness is in the direction of the long tail

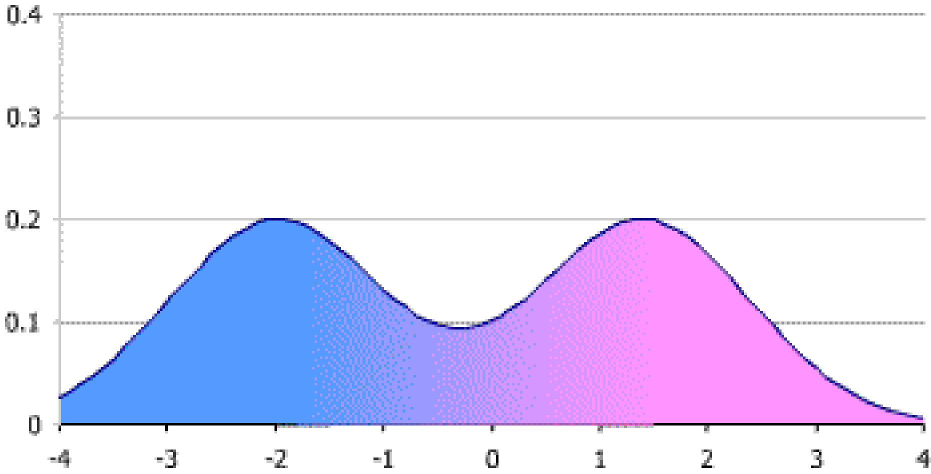
The shape of the data

3. Modality: Other Descriptors

Understand the number and nature of peaks in a distribution, which provides insight into the data's structure and characteristics

3. Modality: number of peaks or modes in a distribution

Modality is useful because it reveals the underlying structure and characteristics of the data by identifying the number of distinct groups or clusters within a dataset.

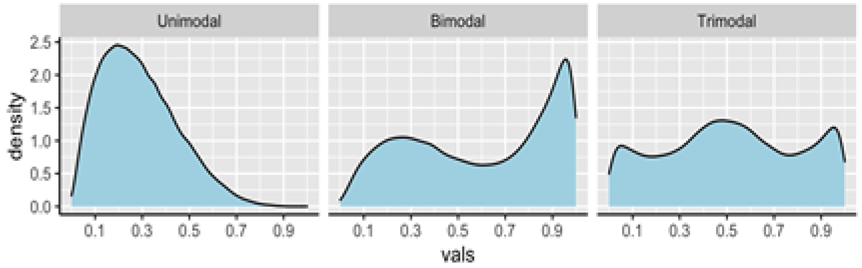


- **Unimodal Distribution:** One peak, showing a single trend.

Example: Adult human heights having one peak around the average height.

- **Bimodal Distribution:** Two peaks, indicating two groups.

Example: Exam scores with high and low achievers.



- **Trimodal (three peaks) Example:** Monthly website traffic, with peaks around product launches, seasonal sales, and special events.
- **Multimodal (more than three peaks):** suggesting various factors. **Example:** Retail sales with peaks during holidays and promotions.

4. Measures of Variance

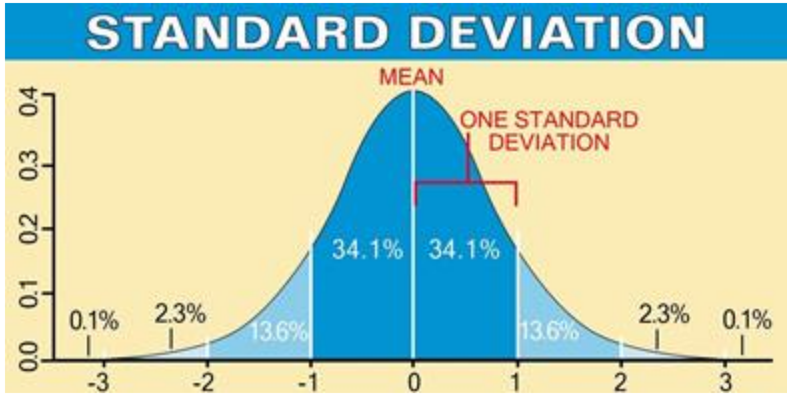
The spread of the data/ Measure of dispersion:

An important characteristic of any set of data is the variation in the data; it reflects how tightly or widely data points are distributed around the mean.

The **standard deviation and variance** are the most common measures of this spread.

The standard deviation σ and The Variance

1. **Variance**:
   - Measures the **average squared deviation** of each data point from the mean.
   - Formula: $\text{Variance} = \frac{\sum(x_i - \mu)^2}{N}$
   - **Units**: Squared units of the data (e.g., meters^2 if measuring height).
   - **Interpretation**: Larger variance = greater spread in the data.



2. **Standard Deviation**:
   - Measures the **average deviation** of each data point from the mean, in the original units.
   - Formula: $\text{Standard Deviation} = \sqrt{\text{Variance}}$
   - **Units**: Same as the data (e.g., meters if measuring height).
   - **Interpretation**: Easier to interpret than variance, as it's in the same units as the data.

Good one: https://www.shiksha.com/online-courses/articles/variance-and-standard-deviation/

Measures of Sample Variance and Sample Std. Deviation

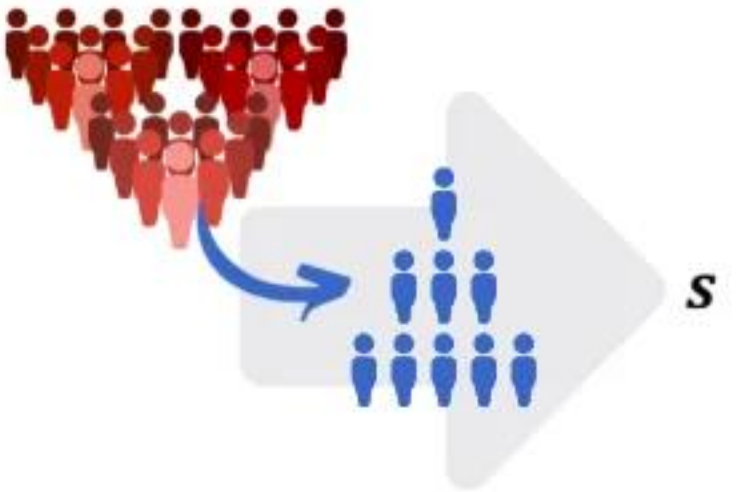
The procedure to calculate the standard deviation depends on whether the **numbers are the entire population or are data from a sample**. The calculations are similar, but not identical.

Sample Variance

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

Sample Standard Deviation

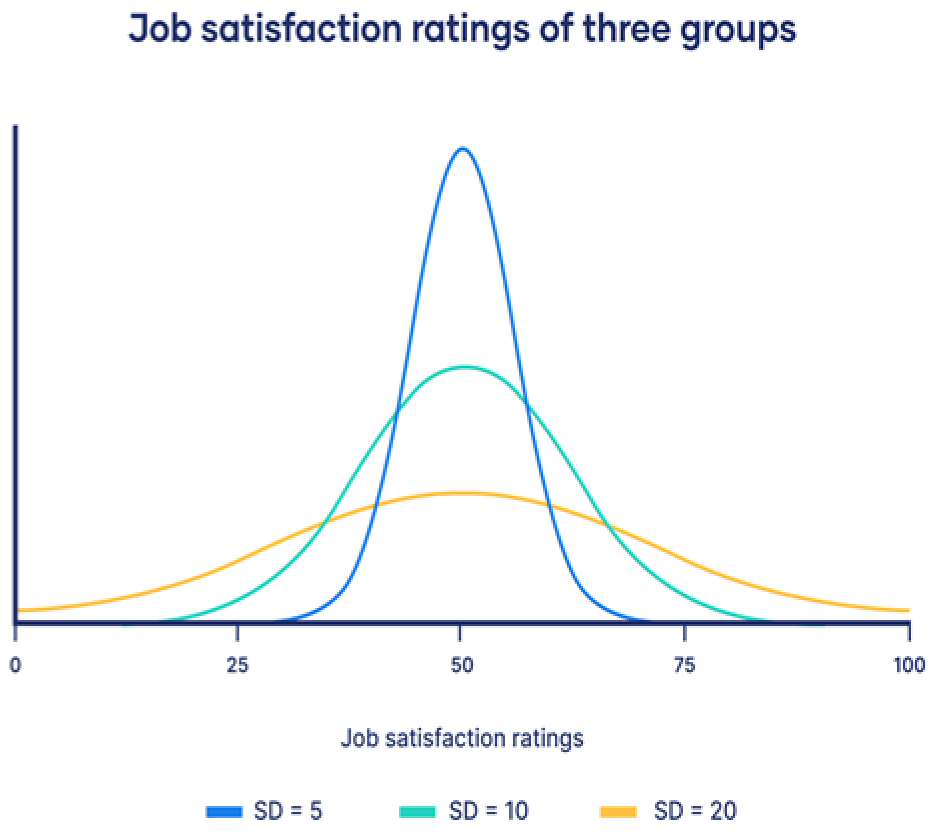
$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n - 1}}$$



NOTE: In general, the sample standard deviation is preferred over the sample variance

- ***Interpretability***: standard deviation measured in same units as the original data
- **Scale Sensitivity:** Variance involves squaring differences, making comparisons across datasets with different units or scales challenging.

Interpreting Variability with Standard Deviation



Greater Variability (a wide, flat distribution curve): Larger standard deviation.
Less Variability (tall, spike-like distribution curve): Smaller standard deviation (data points are closer to the mean).

This visual representation helps us understand the level of variability in the data.

5. Correlation and Relationships (Later Topics)

Correlation: Measures the strength and direction of the relationship between two variables.

Understanding Correlation:

- Positive: Variables move in the same direction.
- Negative: Variables move in opposite directions.

Measuring Correlation: Pearson's Correlation Coefficient (r): Ranges from -1 to 1.

- Example: Variables: Hours studied vs. Exam scores Correlation: Positive (r≈0.8).