Using Statistics to Summarize Data Sets

It focuses on **numerical summaries** of data sets, providing a quick overview of key characteristics like center, spread, and the relationship between variables.

Sample Mean

- **Sample Mean ($\bar{x}$)**: Average of the data values.

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  Where:

  - n= number of observations
  - xi = each data point

- **Deviations**: Differences from the mean, $xi - \overline{x}$. The sum of deviations is always zero.

🔷 **Key Point:** Mean **measures central tendency**, sensitive to outliers.

Sample Median

- **Median**: Middle value when data is ordered.
  - If n is odd, median is the middle value.
  - If n is even, median is the average of the two middle values.
- **Comparison**: Median is **less affected by outliers** compared to the mean.

🔷 **Key Point:** Use the median for **skewed distributions**.

Sample Mode

- **Mode**: The data set's most frequently occurring value(s).
  - **Unimodal**: One mode.
  - **Bimodal**: Two modes.

- **Multimodal**: More than two modes.

🔷 **Key Point:** Mode is useful for **categorical data**.

Sample Variance and Sample Standard Deviation

- **Sample Variance** (s^2): Measures data spread around the mean

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

- **Why** $n-1$**?** It's the **degrees of freedom —** adjusts for sample size.
- **Sample Standard Deviation (**s**)**: Square root of the variance, gives spread in original units.

$$s = \sqrt{s^2}$$

A standard deviation (or σ) **measures how dispersed the data is in relation to the mean**. A low or small standard deviation indicates data are clustered tightly around the mean, and a high or large standard deviation indicates data are more spread out.

🔷 **Key Point:** Higher variance/standard deviation = **greater spread**.

Normal Data Sets and the Empirical Rule

- For **normal distributions**, about:
  - **68%** of data falls within $\overline{x} \pm s$
  - **95%** within $\overline{x} \pm 2s$
  - **99.7%** within $\overline{x} \pm 3s$
- **The empirical rule** helps estimate the spread of data quickly if the distribution is roughly normal.

🔷 **Key Point:** Use the empirical rule for **quick checks** on data spread.

Sample Correlation Coefficient

- **Correlation Coefficient (**r**)**: Measures the strength and direction of a linear relationship between two variables.

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

Where r ranges from -1 to 1

- $r = 1$: Perfect positive correlation.
- $r = -1$: Perfect negative correlation.
- $r = 0$: No linear correlation.

Chebyshev's Inequality

Imagine you have a dataset with many numbers and calculate the average (mean) and the spread (variance or standard deviation). You might wonder:

- How much of the data is **close** to the mean?
- How much is **far away** from it?

Chebyshev's inequality tells us that for any dataset (no matter the shape of its distribution), **at least a certain proportion** of the values will lie within a given number of standard deviations from the mean.

Mathematically, Chebyshev's inequality states:

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

or equivalently,

$$P(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}$$

Where

- X is a random variable,
- μ is the mean (expected value) of X,
- σ is the standard deviation of X,
- k is any number greater than 1.

What Does This Mean?

- The inequality tells us that **at least** $1 - \frac{1}{k^2}$ of the values will be within k standard deviations from the mean.
- This is useful because it works **for any probability distribution**, whether it's normal, skewed, or even unknown.

Key Results from Chebyshev's Inequality

For different values of k:

k	At Least This Much Data Lies Within k Standard Deviations
k=2	At least **75%** $(1 - \frac{1}{4})$
k=3	At least **88.89%** $(1 - \frac{1}{9})$
k=4	At least **93.75%** $(1 - \frac{1}{16})$
k=5	At least **96%** $(1 - \frac{1}{25})$

This means that:

- At least **75%** of the data is within **2 standard deviations** of the mean.
- At least **88.89%** of the data is within **3 standard deviations** of the mean.
- This applies **no matter what** the shape of the distribution is!

Real-Life Example

Example: Exam Scores

Suppose the average score on an exam is **70,** with a standard deviation of **10**. You want to estimate how many students scored between **50 and 90**.

1. The distance from the mean:
    a. 90−70=20,   70−50=20

    This is **2 standard deviations** (k=2).
2. Applying Chebyshev's inequality:

    $P(|X - 70| < 2(10)) \geq 1 - \frac{1}{4} = 0.75$

    This means **at least 75%** of students scored between 50 and 90.

Even though we don't know the exact distribution of scores, we can confidently say at least 75% of the scores are in this range.

Conclusion

Chebyshev's inequality is a powerful tool for estimating how much data is clustered around the mean, even when the exact shape of the distribution is unknown. It provides a lower bound on probabilities, making it useful in situations where we don't assume normality.

One-Sided Chebyshev's Inequality

The standard **Chebyshev's inequality** bounds the probability of a random variable being far from the mean **in both directions**. However, in some cases, we only care about values **above** or **below** the mean. This is where **one-sided Chebyshev's inequality** comes in.

The formula for One-Sided Chebyshev's Inequality

For a random variable X with mean μ and standard deviation σ, the one-sided Chebyshev's inequality states:

$$P(X - \mu \geq k\sigma) \leq \frac{1}{1 + k^2}$$

or

$$P(X - \mu \leq -k\sigma) \leq \frac{1}{1 + k^2}$$

where:

- X is a random variable,
- μ is the mean of X,
- σ is the standard deviation,
- k is a positive number (i.e., how many standard deviations away from the mean we are considering).

Interpretation

- The standard **two-sided** Chebyshev inequality tells us **at least** how much probability is within a certain range.
- The **one-sided** Chebyshev inequality tells us the **maximum** probability that values exceed a given threshold **in only one direction** (above or below the mean).

This is useful when we are only interested in **extreme values in one direction**, such as:

- **Risk management**: Probability of extreme losses (low tail).
- **Quality control**: Probability of defects exceeding a certain limit.
- **Stock market**: Probability of a crash or a price surge.

Key Differences from Standard Chebyshev's Inequality

Aspect	Two-Sided Chebyshev	One-Sided Chebyshev
Direction	Both sides (above & below the mean)	Only one side (above or below)
Formula	(P(	X - \mu
Interpretation	How much probability is within a range	Maximum probability of extreme values in one direction

Understanding Quartiles

A **quartile** divides ordered data into **four equal parts**:

1. **Q1 (First Quartile - 25th Percentile)**: The median of the lower half of the data (excluding the overall median if n is odd).
2. **Q2 (Second Quartile - 50th Percentile)**: The median of the dataset.
3. **Q3 (Third Quartile - 75th Percentile)**: The median of the upper half of the data.(including the overall median if n is odd).

Interquartile Range (IQR) Formula:

$$IQR = Q3 - Q1$$

This represents the range within which the central 50% of the data lies.

Interpretation

- **If IQR is large**, the middle 50% of the data is widely spread → **high variability**.
- **If IQR is small**, the data is tightly clustered → **low variability**.