dv

A comprehensive cheat sheet covering measures of variability, correlation analysis, and causality in statistics. This guide includes formulas, interpretations, and common pitfalls to help you understand and apply these concepts effectively.

Measures of Variability

Basic Measures of Variability

Range	Simplest measure; difference between the maximum and minimum values in a dataset.
Variance	Average of the squared deviations from the mean. Indicates the spread of data points around the mean.
Standard Deviation	Square root of the variance. Measures the typical distance of data points from the mean.
Population Variance (σ^2)	$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$
Sample Variance (s^2)	$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$ (for n < 30)
Notation	σ^2: Population variance s^2: Sample variance μ: Population mean $\bar{x}$: Sample mean N: Population size n: Sample size

Tchebysheff's Theorem

Guarantees a minimum percentage of data within k standard deviations of the mean, regardless of the data's distribution.

Formula:
At least $1 - \frac{1}{k^2}$ of the data falls within k standard deviations of the mean.

Example:
At least 75% of data falls within 2 standard deviations of the mean (k = 2).

Empirical Rule (68-95-99.7 Rule)

Applies to bell-shaped (normal) distributions:

- Approximately 68% of data falls within 1 standard deviation of the mean.
- Approximately 95% of data falls within 2 standard deviations of the mean.
- Approximately 99.7% of data falls within 3 standard deviations of the mean.

Z-Scores and Bivariate Data

Z-Scores

Definition	Measures the distance between a data point and the mean in terms of standard deviations.
Formula	$z = \frac{x - \mu}{\sigma}$ (population) $z = \frac{x - \bar{x}}{s}$ (sample)
Interpretation	A z-score indicates how unusual or typical a data point is within its distribution. • $\|z\| > 2$: Potentially unusual. • $\|z\| > 3$: Potential outlier.

Covariance

Measures how two variables change together. Positive covariance indicates that the variables increase or decrease together, while negative covariance indicates an inverse relationship.

Limitations: Not standardized, difficult to interpret the strength of the relationship.

Correlation Coefficient (Pearson's r)

Definition	Standardized measure of the strength and direction of a linear relationship between two variables.
Range	−1 to +1 • −1: Perfect negative linear relationship. • 0: No linear relationship. • +1: Perfect positive linear relationship.
Formula	$r = \frac{1}{n} \sum_{i=1}^{n} \frac{x_i - \bar{x}}{s_x} \frac{y_i - \bar{y}}{s_y}$

Correlation Analysis

Assumptions of Pearson's Correlation

- **Linearity:** The relationship between X and Y should be linear.
- **Normality:** X and Y should be normally distributed.
- **Homoscedasticity:** The variance of Y at different values of X should be constant.

Checking Assumptions

Linearity	Scatterplots: Check for a linear pattern (oval shape).
Normality	Histograms/Q-Q plots: Assess normality.
Homoscedasticity	Residual plots (if performing regression): Check for constant variance.

Interpreting Correlation

Correlation measures the degree of linear association, but does not imply causation.

- **Pitfalls:**
  - Correlation does not equal causation.
  - Overlooking outliers or nonlinearity.
  - Using correlation alone to generalize about complex relationships.

Advanced Correlation Concepts & Causality

Regression Line

For every standard deviation σX increase above the average μX, Y grows ρ standard deviations σY above the average μY.

Formula:
$E(Y \mid X = x) = \mu Y + \rho * ((X - \mu X) / \sigma X) * \sigma Y$

Variance Explained

Conditioning on a random variable X can help reduce the variance of Y.

$SD(Y \mid X = x) = \sigma Y * \sqrt{(1 - \rho^2)}$

The variance decreases by ρ^2 percent.

Other Types of Correlations

Spearman's ρ	Rank-based correlation, measures monotonic relationships. Less sensitive to outliers.
Kendall's τ	Rank-based, interprets concordance among pairs. Used for smaller samples or data with many ties.
Partial Correlation	Measures the linear relationship between two variables while controlling for one or more additional variables.

Causality

Correlation is not causation.

Confounding Effect: A confounder is a third variable that affects both the independent (X) and dependent (Y) variables, leading to a spurious association.

Mediating Effect: A mediator is an intermediate variable that explains the relationship between X and Y.

Colliding Effect (Collider Bias): A collider is a variable influenced by both X and Y. Conditioning on it can introduce a spurious association.