Statistical formulae, possible variations, and alternatives, ensuring a comprehensive understanding of the key concepts.

|  |  |  |  |
| --- | --- | --- | --- |
| Formula | Description of Terms | Why It's Used? | Variations/Alternatives |
| X̄ = ΣXᵢ / n | Mean (Average): Xᵢ are data values, n is the total number of observations. | To find the central tendency of a dataset. | Median, Mode (for skewed distributions). |
| σ² = Σ(Xᵢ - X̄)² / n | Variance: Measures data spread; Xᵢ are data points, X̄ is the mean, n is the number of observations. | To measure dispersion and variability in data. | Standard Deviation (square root of variance), Mean Absolute Deviation (MAD). |
| σ = √(Σ(Xᵢ - X̄)² / n) | Standard Deviation: Square root of variance, represents dispersion. | Provides a measure of dispersion in the same units as the data. | Coefficient of Variation (CV) for relative dispersion. |
| Z = (X - μ) / σ | Z-score: Standardization formula where X is data value, μ is the mean, σ is the standard deviation. | To compare different data points across distributions. | T-score (for small samples), Percentile ranking. |
| P(X = k) = (nCk) p^k (1-p)^(n-k) | Binomial Distribution: n = trials, k = successes, p = probability of success. | Used for binary outcomes (success/failure). | Poisson distribution (if trials are infinite). |
| P(X = k) = (e^(-λ) λ^k) / k! | Poisson Distribution: λ = average rate, k = occurrences. | To model rare events over a fixed interval. | Binomial distribution (for small n), Exponential distribution. |
| f(x) = (1 / (σ√2π)) e^(-(x - μ)² / 2σ²) | Normal Distribution: μ = mean, σ = standard deviation. | To describe naturally occurring datasets. | Student's t-distribution (for small samples). |
| r = Σ(Xᵢ - X̄)(Yᵢ - Ȳ) / √[Σ(Xᵢ - X̄)² Σ(Yᵢ - Ȳ)²] | Pearson Correlation: Measures strength of relationship between variables. | To measure linear relationships between two variables. | Spearman's Rank Correlation (for ordinal data), Kendall's Tau. |
| Y = a + bX | Regression Equation: a = intercept, b = slope, X = independent variable. | To model relationships between dependent and independent variables. | Polynomial regression, Logistic regression (for classification). |
| t = (X̄ - μ) / (σ / √n) | t-test formula: Tests hypothesis on sample mean (X̄) vs population mean (μ). | Used for hypothesis testing with small samples. | Z-test (for large samples), Paired t-test (for related samples). |
| F = variance₁ / variance₂ | F-test: Compares two sample variances. | To compare the variance of two populations. | Bartlett’s test (for homogeneity), Levene’s test (for normality violation). |
| χ² = Σ(Oᵢ - Eᵢ)² / Eᵢ | Chi-square test: Oᵢ = observed values, Eᵢ = expected values. | To test categorical data for independence or goodness of fit. | Fisher's exact test (for small samples), G-test. |
| Z = (X̄ - μ) / (σ / √n) | Z-test: Used when population variance is known. | To test population means for large samples. | t-test (when variance is unknown). |

**Coverage Check**

You have appropriately covered:

1. **Measures of central tendency and dispersion.**
2. **Probability distributions (Binomial, Poisson, Normal).**
3. **Correlation and regression.**
4. **Hypothesis testing (t-test, F-test, Chi-square, Z-test).**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Concept** | **Description** | **Different Approaches** | **Variations** | **Applicability** |
| **Measures of Central Tendency** | Describes the center point of data (mean, median, mode). | Arithmetic mean, Median (middle value), Mode (most frequent). | Mean ≠ Median ≠ Mode in skewed distributions; Mean = Median = Mode in normal distribution. | Use mean for symmetric data, median for skewed data, mode for categorical data. |
| **Measures of Dispersion** | Quantifies the spread of data. | Range, Variance, Standard Deviation, Interquartile Range (IQR). | High variance indicates greater spread; IQR used for skewed data. | Use SD for normal distribution; IQR for outlier-prone data. |
| **Skewness & Kurtosis** | Skewness measures asymmetry, Kurtosis measures peakness of data distribution. | Pearson’s skewness, Moment coefficient of skewness, Fisher’s kurtosis. | Skewness: Positive, Negative; Kurtosis: Leptokurtic, Mesokurtic, Platykurtic. | Normal distributions have zero skewness and mesokurtic kurtosis. |
| **Correlation vs. Causation** | Correlation indicates association; causation implies one variable affects another. | Pearson correlation (linear), Spearman rank (monotonic), Causality tests (Granger causality). | Strong correlation doesn’t imply causation; spurious correlations exist. | Useful in predictive analytics but requires careful interpretation. |
| **Regression Analysis** | Models relationships between dependent and independent variables. | Simple Linear Regression, Multiple Regression, Logistic Regression. | Linear vs. Non-linear; Parametric vs. Non-parametric regression. | Used in forecasting and impact analysis across industries. |
| **Sampling Techniques** | Methods to select representative data from a population. | Simple random, Stratified, Systematic, Cluster sampling. | Probability vs. Non-probability sampling. | Important for unbiased analysis, generalization of results. |
| **Hypothesis Testing** | Determines if a hypothesis about population parameters is valid. | One-sample, Two-sample, Paired tests, ANOVA. | Parametric (t-test, Z-test) vs. Non-parametric (Mann-Whitney U). | Used in scientific studies and business decision-making. |
| **Confidence Intervals** | A range of values likely to contain the population parameter. | Mean-based, Proportion-based, Bootstrapping. | Affected by sample size, confidence level (90%, 95%, 99%). | Provides uncertainty estimation in predictions. |
| **Probability Distributions** | Describes likelihood of outcomes in an experiment. | Discrete (Binomial, Poisson), Continuous (Normal, Exponential). | Symmetric (Normal) vs. Skewed (Exponential). | Used in risk analysis, reliability studies, quality control. |
| **Time Series Analysis** | Examines data points collected over time. | Moving Averages, Exponential Smoothing, ARIMA models. | Seasonal vs. Non-seasonal; Additive vs. Multiplicative models. | Used in stock market prediction, climate modeling, demand forecasting. |
| **Non-Parametric Tests** | Statistical tests without assuming specific data distribution. | Chi-square, Mann-Whitney U, Kruskal-Wallis test. | Used when normality assumption is violated. | Suitable for ordinal or categorical data. |
| **Bayesian vs. Frequentist Statistics** | Two competing approaches to statistical inference. | Bayesian updates beliefs with prior data; Frequentist relies on long-run frequency. | Bayesian considers prior probability; Frequentist doesn’t. | Bayesian useful in dynamic environments; Frequentist in fixed trials. |
| **Outliers Detection** | Identifies data points that deviate significantly. | Z-score, IQR method, Box plot analysis. | Mild vs. Extreme outliers. | Critical in fraud detection, anomaly detection. |
| **Data Transformation** | Converts data into a suitable format for analysis. | Normalization, Standardization, Log transformation. | Linear vs. Non-linear transformations. | Used to improve model performance and comparability. |
| **Correlation Coefficient** | Measures the strength and direction of a relationship between two or more variables. | Pearson (linear), Spearman (rank-based), Kendall’s Tau (ordinal data). | - **Simple Correlation:** Between two variables.- **Multiple Correlation:** Between one dependent and multiple independent variables.- **Partial Correlation:** Measures relationship while controlling other variables.- **Canonical Correlation:** For multiple dependent and independent variables. | Used in trend analysis, prediction modeling, and feature selection in machine learning. |
| **Inferences from Correlation** | Helps understand how variables move together but does not imply causation. | Correlation matrix, Heatmaps for visualization. | - Positive correlation (+r) means variables move in the same direction.- Negative correlation (−r) means they move in opposite directions.- Zero correlation means no relationship. | Useful in identifying predictive features, but must be cautious of confounding factors. |

**Key Inferences and How to Use Correlation:**

1. **Correlation ≠ Causation:** A high correlation does not imply that one variable causes another.
2. **Multi-Collinearity Issues:** High correlation among independent variables can affect regression models.
3. **Interpretation:**
   * r>0.7r > 0.7r>0.7: Strong positive correlation.
   * 0.3<r<0.70.3 < r < 0.70.3<r<0.7: Moderate correlation.
   * r<0.3r < 0.3r<0.3: Weak correlation.
   * r=0r = 0r=0: No correlation.
4. **Choosing the Right Type:**
   * Use **Pearson correlation** for continuous numerical data with a linear relationship.
   * Use **Spearman correlation** when data is ordinal or non-linear.
   * Use **Partial correlation** to control for additional variables' effects.