DegreeLive

BBA	SEMESTER VI	UNIVERSITY OF KERALA	COURSE CODE: UK6DSCMGT303

STATISTICS FOR BUSINESS DECISIONS

EXAM REVISION NOTES

JOIN WHATSAPP CHANNEL →

MODULE I

STATISTICS AND MEASURES OF CENTRAL TENDENCY

Definitional Scope and Functional Architecture of Statistics: In contemporary enterprise management, quantitative data constitutes the empirical foundation for rational decision-making, eliminating speculative guesswork. In its **Plural Sense**, statistics denotes aggregates of numerical facts affected to a marked extent by a multiplicity of causes, systematically collected for a predetermined purpose, and placed in relation to each other. In its **Singular Sense**, statistics represents the scientific discipline encompassing the collection, organization, presentation, analysis, and interpretation of numerical data. Indispensable managerial functions include data condensation, comparative performance benchmarking, statistical hypothesis testing, prospective corporate forecasting, quality control, and risk quantification.

The Distrust of Statistics and Ethical Safeguards: Widespread public and executive skepticism toward statistical arguments stems from pervasive misuse, fallacious interpretation, and deliberate commercial distortion rather than flaws in statistical theory. Common abuses include: **Selection Bias** (unrepresentative sampling); **Omission of Sample Size** (misleading percentage claims); **Spurious Correlation** (confusing association with causation); **Truncated Graphical Scaling** (exaggerating minuscule changes); and **P-Hacking** (data dredging for random significance). Professional safeguards require transparent disclosure of sample frames, publishing full methodology blueprints, reporting dispersion metrics alongside averages, and adhering to strict analytical objectivity.

Yule's Criteria for an Ideal Statistical Average: G. Udny Yule formulated six rigorous criteria governing an ideal measure of central tendency: **Rigid Definition** (unambiguous mathematical formulation producing identical results across independent analysts); **Inclusion of All Observations** (reflecting the full population distribution); **Simplicity and Intuitiveness** (clear conceptual meaning for non-technical stakeholders); **Amenability to Further Algebraic Treatment** (enabling combined group calculations without raw micro-data); **High Sampling Stability** (minimal variance across random samples); and **Resistance to Extreme Values** (immunity to outlier distortion).

Arithmetic Mean Foundations and Algebraic Properties: The Arithmetic Mean is the quotient obtained by dividing the sum of all observations by total sample size ((N)). It possesses profound mathematical properties: (1) the algebraic sum of deviations from the actual mean is identically zero ((sum (X - X)) = 0}), proving the mean acts as the distribution's center of gravity; (2) the sum of squared deviations from the mean is at an absolute minimum ((sum (X - X))^2 < sum (X - A)^2}), providing the justification for least-squares regression; (3) the product of the mean and sample size reproduces the grand aggregate sum ((N X) = sum X}); and (4) the **Combined Mean** of merged sub-groups is computed directly as (sum (N_i X)_i) / sum N_i}.

Weighted Arithmetic Mean and Step-Deviation Mechanics: When items carry unequal economic importance (such as weighting staple food spending in consumer price indices), the **Weighted Arithmetic Mean** ((sum (W · X) / sum W)) assigns numerical weights proportional to relative operational significance. In continuous frequency distributions with equal class intervals ((c)), the **Step-Deviation Method** ((A + [(sum fd' / N) c]), where (d' = (m - A)/c)) drastically compresses computational effort, eliminating large numerical products while preserving exact mathematical precision.

Positional Averages: Median and Partition Values: The **Median** ((M)) is the positional middle value partitioning an ordered array into two equal halves (50% below, 50% above). In continuous grouped data, the median class is identified by (N/2), and the exact value interpolated via (L + [((N/2) - cf)/f] c). The median is robust against extreme outliers and easily computed for open-ended classes. **Partition Values** subdivide ordered data: **Quartiles** divide data into four equal parts ((Q_1, Q_2 = M, Q_3)); **Deciles** divide into ten parts ((D_1) to (D_9)); and **Percentiles** divide into one hundred parts ((P_1) to (P_{99})}), widely deployed in psychometric testing and socioeconomic inequality benchmarking.

Graphical Determination of Positional Averages: The median and partition values can be derived graphically without mathematical formulas by constructing **Cumulative Frequency Ogives**. The X-coordinate of the intersection point between the *Less-than Ogive* and the *More-than Ogive* corresponds precisely to the **Median** on the horizontal axis. Similarly, the **Mode** can be located graphically from a **Histogram** by drawing diagonal intersecting lines across the peak modal bar and adjacent neighboring bars, projecting the intersection point downward to the horizontal X-axis.

Mode and Empirical Frequency Relationships: The **Mode** ((Z)) is that value occurring with the greatest frequency, representing peak consumer popularity or dominant operational demand. In continuous series, it is interpolated as (L + [(f_1 - f_0)/(2f_1 - f_0 - f_2)] c). In moderately asymmetrical skewed distributions, Karl Pearson's **Empirical Relationship** establishes: **Mode equals 3 Median minus 2 Mean** ((Z = 3M - 2X)}), enabling mode estimation when distributions exhibit ill-defined modal boundaries or open-ended tails.

Specialized Mathematical Averages: Geometric and Harmonic Means: The **Geometric Mean (GM)** is the (N)-th root of the product of values, computed logarithmically as (Antilog(sum log X / N)). GM is uniquely suited for calculating compound average growth rates (CAGR) and constructing index numbers. The **Harmonic Mean (HM)** is the reciprocal of the arithmetic mean of reciprocals, formulated as (N / sum (1/X)), deployed for averaging rates and speeds when distance or expenditure is held constant. In financial portfolio analytics, HM is mathematically required when averaging Price-to-Earnings (P/E) ratios across multiple equities to prevent upward distortion caused by extreme outlier multiples. For any positive unequal dataset, the immutable mathematical inequality establishes: **Arithmetic Mean >= Geometric Mean >= Harmonic Mean**.

MODULE II

MEASURES OF DISPERSION AND SKEWNESS

The Operational Significance of Dispersion: Central tendency alone provides an incomplete description; two distributions can share an identical mean while exhibiting radically divergent internal variations. **Dispersion** measures the extent to which individual observations scatter, spread, or deviate around the central average. Measuring dispersion enables managers to assess the reliability of an average, control manufacturing process variation in Six Sigma, and evaluate financial investment risk. **Absolute Measures** are expressed in raw physical units (Range, Quartile Deviation, Mean Deviation, Standard Deviation), whereas **Relative Measures** are dimensionless coefficients enabling cross-series comparisons.

Distance and Positional Measures of Dispersion: The **Range** ((R = L - S)) is the difference between largest and smallest values, highly sensitive to outliers but useful in stock trading and statistical quality control. The **Interquartile Range (IQR)** ((Q_3 - Q_1)) measures the spread of the middle 50% of the distribution. The **Quartile Deviation (QD)** (Semi-Interquartile Range) equals ((Q_3 - Q_1) / 2), with **Coefficient of QD** formulated as ((Q_3 - Q_1) / (Q_3 + Q_1)), serving as the optimal dispersion measure for open-ended and severely skewed datasets.

Mean Deviation (Average Deviation): Mean deviation is the arithmetic mean of the absolute deviations of values from an average, ignoring algebraic signs ((sum f|X - A| / N)). A fundamental mathematical theorem proves that the sum of absolute deviations is at an absolute minimum when taken from the **Median**; consequently, (MD_M <= MD_{X})}. The **Coefficient of Mean Deviation** divides the absolute mean deviation by the corresponding central average.

Standard Deviation and Variance Mathematical Superiority: Formulated by Karl Pearson, the **Standard Deviation** ((sigma)) is the positive square root of the mean of squared deviations from the arithmetic mean, eliminating negative signs through algebraic squaring rather than arbitrary sign-dropping. The square of standard deviation is **Variance** ((sigma^2)). Fundamental theorems include: (1) minimal root mean square deviation; (2) complete independence from change of origin; (3) linear responsiveness to change of scale ((sigma_{bX) = |b| sigma_X} and (Var(aX + b) = a^2 Var(X))); (4) **Charlier's Mathematical Check** ((sum f(d'+1)^2 = sum f(d')^2 + 2 sum fd' + N)) for verifying step-deviation calculations; and (5) **Combined Standard Deviation** formula for merged groups ((sqrt{[N_1(sigma_1^2 + d_1^2) + N_2(sigma_2^2 + d_2^2)] / (N_1 + N_2))}). In normal distributions, (mean +- 1 sigma) contains 68.27% of observations, (mean +- 2 sigma) contains 95.45%, and (mean +- 3 sigma) contains 99.73%. In moderately symmetrical series, the empirical approximation establishes: **6 QD approx 5 MD approx 4 Standard Deviations**.

The Coefficient of Variation (CV) in Risk Diagnostics: Formulated by Pearson, the **Coefficient of Variation** ((CV = (sigma / X)) 100}) expresses standard deviation as a percentage of the mean. It is the definitive relative metric for comparing consistency, stability, and risk. **Decision Rule:** The series possessing the **LOWER CV** is judged more consistent, uniform, and stable; the series with the **HIGHER CV** exhibits greater operational volatility and financial risk per unit of return.

The Lorenz Curve and Gini Concentration Ratio: The **Lorenz Curve** is a cumulative graphical device developed by Max O. Lorenz to measure and visually represent socioeconomic inequality in income, wealth distribution, or industrial sales concentration. The percentage of population is plotted on the horizontal axis against cumulative percentage of income/wealth on the vertical axis. The diagonal line at 45 degrees represents the **Line of Equal Distribution**. The greater the curvature and distance of the Lorenz curve away from the 45-degree line, the higher the inequality in the distribution. The area between the line of equal distribution and the Lorenz curve, expressed as a ratio of the total triangular area, is the **Gini Coefficient** (ranging from 0 for absolute equality to 1 for absolute monopoly).

Mathematical Foundations and Measures of Skewness: Skewness signifies the lack of symmetry in a statistical distribution. In a **Symmetrical Distribution**, Mean equals Median equals Mode ((skewness = 0)). In a **Positively Skewed Distribution**, an elongated tail stretches toward the right ((Mean > Median > Mode)). In a **Negatively Skewed Distribution**, an elongated tail stretches toward the left ((Mode > Median > Mean)). **Karl Pearson's Coefficients:** (Sk_p = (X) - Z) / sigma} or (3(X) - M) / sigma}, bounded between -3 and +3. **Bowley's Quartile Coefficient:** (Sk_b = (Q_3 + Q_1 - 2M) / (Q_3 - Q_1)), strictly bounded between -1 and +1. **Kelly's Measure of Skewness:** Formulated on deciles or percentiles as (Sk_k = (P_{90} + P_{10} - 2P_{50}) / (P_{90} - P_{10})}, providing robust percentile asymmetry metrics.

MODULE III

CORRELATION ANALYSIS

Bivariate Interdependence and Correlation Theory: Correlation analysis investigates and quantifies the degree and directional sympathy of association between two or more interrelated variables. It provides corporate leadership with empirical evidence regarding whether commercial metrics co-vary in unison or opposition, serving as the essential statistical foundation for regression modeling and investment portfolio diversification.

Philosophical Demarcation: Correlation versus Causation: Correlation establishes mathematical association, not physical cause-and-effect. Two variables may exhibit a high correlation coefficient under non-causal conditions: **Spurious Association** (both variables driven independently by a common confounding factor, such as ice cream sales and drowning incidents driven by ambient temperature); **Chance / Nonsense Association** (random statistical co-variation without theoretical plausibility); and **Reverse Causality**. Empirical causation requires establishing concomitant variation, temporal precedence, and the elimination of confounding third variables.

Taxonomy of Correlation Associations: Categorized across three dimensions: **By Direction:** Positive Correlation (variables move in identical direction, e.g., advertising and revenue) versus Negative Correlation (variables move in opposite directions, e.g., price and demand). **By Constancy of Ratio:** Linear Correlation (constant ratio of change forming a straight line) versus Non-Linear Curvilinear Correlation (varying ratio of change forming a curved path). **By Number of Variables:** Simple Correlation (two variables), Partial Correlation (two variables controlling for others), and Multiple Correlation (simultaneous joint association across three or more variables).

Partial and Multiple Correlation Mathematical Formulations: In multivariate business phenomena where three variables ((X_1, X_2, X_3)) interact, the **Partial Correlation Coefficient** ((r_{12.3)}) measures the net linear association between variables 1 and 2 after eliminating the linear influence of variable 3: (r_{12.3) = (r_{12} - r_{13} r_{23}) / [sqrt{1 - r_{13}^2} sqrt{1 - r_{23}^2}]}. The **Multiple Correlation Coefficient** ((R_{1.23)}) measures the joint linear association between variable 1 and the combined predictors 2 and 3: (R_{1.23) = sqrt{(r_{12}^2 + r_{13}^2 - 2 r_{12} r_{13} r_{23}) / (1 - r_{23}^2)}}, bounded strictly between 0 and +1.

Exploratory Scatter Diagram Interpretation: A Cartesian plot of paired coordinates ((X_i, Y_i)) reveals relationship form visually: points clustering on an upward straight diagonal corridor indicate high positive linear correlation; points along a downward diagonal indicate high negative correlation; an amorphous circular cloud indicates zero correlation. While intuitive and exposing outliers, scatter diagrams provide no precise numerical coefficient.

Karl Pearson's Product-Moment Coefficient (r): The preeminent parametric metric of linear association between continuous variables, defined as Covariance divided by the product of individual standard deviations ((r = sum xy / sqrt{sum x^2 sum y^2)}). Core mathematical properties: (1) bounded strictly between **-1.0 and +1.0**; (2) pure dimensionless number; (3) perfectly symmetrical ((r_{xy) = r_{yx}}); (4) completely invariant to change of origin and scale; and (5) geometric mean of regression coefficients ((r = +- sqrt{b_{yx) b_{xy}}}).

Probable Error and Statistical Significance Testing: The **Standard Error of r** is ((1 - r^2) / sqrt{N)}, and **Probable Error** is (PE_r = 0.6745 [(1 - r^2) / sqrt{N)]}. **Decision Criteria:** If (r < PE_r), correlation is statistically insignificant; if (r > 6 PE_r), correlation is highly significant and dependable. Population correlation limits are bounded by (r +- PE_r).

Spearman's Rank Correlation Coefficient: Formulated for qualitative, ordinal, or non-normally distributed data where individuals are ranked. For **Untied Ranks**, (ρ = 1 - [6 sum D^2 / (N(N^2 - 1))]), where (D = R_1 - R_2). For **Tied / Repeated Ranks**, tied items receive the average of ranks they would occupy, and an algebraic correction factor of ([m(m^2 - 1)] / 12) is added to (sum D^2) for every tied group of size (m).

Mathematical Properties and Invariants of Rank Differences: In Spearman's rank correlation computation, a vital mathematical invariant states that the algebraic sum of rank differences across all pairs is identically zero ((sum D = sum (R_1 - R_2) = 0)). This invariant provides an essential arithmetic check for human calculators before squaring rank differences. Furthermore, when two judges assign identical ranks to every single candidate ((R_1 = R_2)), all rank differences are zero ((sum D^2 = 0)), producing perfect positive rank correlation ((ρ = 1 - 0 = +1.0)). Conversely, when two judges assign perfectly opposite, inverted ranks, (sum D^2 = [N(N^2 - 1)] / 3), producing perfect negative rank correlation ((ρ = 1 - [6 N(N^2 - 1) / (3 N(N^2 - 1))] = 1 - 2 = -1.0)).

The Coefficient of Determination in Predictive Modeling: The square of Pearson's coefficient ((r^2)) represents the **Coefficient of Determination**, measuring the proportion of total variation in the dependent variable explained by the independent variable. The residual unexplained variation is the **Coefficient of Non-Determination** ((k^2 = 1 - r^2)), while the **Coefficient of Alienation** is (k = sqrt{1 - r^2)}.

MODULE IV

REGRESSION ANALYSIS

The Predictive Philosophy of Regression: Introduced historically by Sir Francis Galton, regression analysis measures the average functional relationship between variables in terms of original measurement units, estimating the unknown value of a dependent criterion variable ((Y)) from a known independent predictor variable ((X)). Unlike correlation (which is symmetric and unit-free), regression is asymmetrical and expresses concrete rates of change.

The Principle of Least Squares and the Two Regression Lines: Bivariate datasets generate two distinct regression lines: **Line of Y on X** estimates (Y) given (X) by minimizing vertical squared residuals ((sum (Y - \hat{Y))^2}); **Line of X on Y** estimates (X) given (Y) by minimizing horizontal squared residuals ((sum (X - \hat{X))^2}). Both lines intersect at their arithmetic means (((X), \bar{Y})). When correlation is perfect ((r = +- 1.0)), the lines coincide; when correlation is zero ((r = 0)), the lines are perpendicular ((90^{\circ})). The acute angle between lines satisfies (tan θ = [(1 - r^2)/|r|] [(sigma_X sigma_Y)/(sigma_X^2 + sigma_Y^2)]).

Mathematical Properties of Regression Coefficients: The slopes are (b_{yx} = r(sigma_Y / sigma_X) = sum xy / sum x^2} and (b_{xy} = r(sigma_X / sigma_Y) = sum xy / sum y^2}. Core algebraic theorems: (1) (r = +- sqrt{b_{yx} b_{xy}}}; (2) both regression coefficients and the correlation coefficient share the identical algebraic sign; (3) if one coefficient exceeds unity ((> 1.0)), the other must be less than unity ((< 1.0)), since (b_{yx} b_{xy} = r^2 <= 1.0}; (4) the arithmetic mean of regression slopes is greater than or equal to correlation (((b_{yx} + b_{xy})/2 >= r}); and (5) regression slopes are independent of origin changes but sensitive to scale changes.

Formulation of Linear Regression Equations: Using deviation form: **Line of Y on X:** (Y - \bar{Y}) = b_{yx}(X - \bar{X})}, simplifying to (Y = a + b_{yx}X}, where (a = \bar{Y}) - b_{yx}\bar{X}} is the Y-intercept. **Line of X on Y:** (X - \bar{X}) = b_{xy}(Y - \bar{Y})}, simplifying to (X = c + b_{xy}Y}, where (c = \bar{X}) - b_{xy}\bar{Y}} is the X-intercept. In raw data normal equations: (sum Y = Na + b sum X) and (sum XY = a sum X + b sum X^2).

Standard Error of Estimate and Forecasting Bands: The **Standard Error of Estimate** ((S_{yx} = sigma_Y sqrt{1 - r^2}}) measures the standard deviation of actual observations around the fitted regression line. Under normal residual assumptions, management constructs **Forecasting Confidence Bands** around point predictions: (\hat{Y} +- 1.96(S_{yx}}) provides a 95% statistical confidence interval for prospective corporate planning.

Econometric Diagnostics in Multiple Regression: Multiple regression models simultaneous predictors of an outcome. Essential diagnostics include auditing **Multicollinearity** via the Variance Inflation Factor (VIF < 5 to 10), checking residual **Autocorrelation** via the Durbin-Watson statistic (target ~2.0), and confirming **Homoscedasticity** to prevent biased standard errors.

MODULE V

TIME SERIES ANALYSIS

Chronological Structure of Time Series Data: A time series is a sequence of quantitative observations recorded at successive, chronological, equal intervals of time ((Y = f(t))). Time series analysis isolates underlying forces driving historical movements to understand past operational dynamics, forecast future demand, and evaluate strategic policy interventions.

The Four Structural Components of a Time Series: Observed variations reflect four combined forces: **Secular Trend (T):** Smooth, long-term continuous movement over decades reflecting demographic growth, technological innovation, and capital accumulation. **Seasonal Variations (S):** Rhythmic, periodic fluctuations repeating regularly within a single year due to natural climatic forces or cultural/institutional conventions. **Cyclical Fluctuations (C):** Wave-like oscillations over multi-year macroeconomic business cycles (Prosperity, Recession, Depression, Recovery). **Irregular Variations (I):** Erratic, unpredictable shocks from catastrophic natural disasters, wars, strikes, or pandemics.

Mathematical Decomposition Models: Additive Model: Assumes components are independent and in identical measurement units ((Y_t = T_t + S_t + C_t + I_t)). **Multiplicative Model:** Standard commercial practice assuming components are interdependent ((Y_t = T_t × S_t × C_t × I_t)), where trend is in original physical units while seasonal, cyclical, and irregular movements are dimensionless percentage multipliers centered at 100%.

Trend Measurement Methodologies: Freehand Graphic Method: Visual curve fitting through plotted points (subjective). **Semi-Averages Method:** Partitions data into equal halves, averages each segment, and connects midpoints with a straight line. **Moving Averages Method:** Non-parametric smoothing calculating overlapping arithmetic means over a span equal to cycle length, eliminating cyclical noise; even periods require centered moving averages. Primary limitation is boundary data loss preventing future projections.

The Method of Least Squares for Linear Trend: The definitive objective mathematical technique fitting (Y_t = a + bX). By placing the origin at the exact middle time point ((sum X = 0)), parameters simplify into: (a = sum Y / N) (mean of series) and (b = sum XY / sum X^2) (annual growth rate). For **Odd Years**, origin is the middle year ((X) step = 1 year). For **Even Years**, origin is midway between two central years ((X) step = half-year, (b) represents semi-annual growth, annual growth = (2b)).

Seasonal Indices and Deseasonalization: Seasonal patterns are quantified using the **Ratio-to-Moving-Average Method**: dividing raw data by centered moving averages isolates seasonal-irregular components ((Y/TC = SI)); monthly/quarterly averaging eliminates irregular shocks, producing normalized Seasonal Indices summing to 400 (quarterly) or 1,200 (monthly). **Deseasonalization** (([Actual / Seasonal Index] 100)) purges seasonal distortions to reveal authentic business performance.

Link Relatives and Ratio-to-Trend Methods: Under the **Link Relatives Method** (Karl Pearson), each period's value is expressed as a percentage of the immediately preceding period (([Y_t / Y_{t-1}] 100}); link relatives are averaged, chained into chain relatives, adjusted for trend drift, and scaled to a 100% base. In the **Ratio-to-Trend Method**, observed values are expressed as percentages of least-squares trend values (([Y / T] 100)), effectively isolating seasonal movements under the multiplicative paradigm.

Adaptive Short-Term Forecasting via Exponential Smoothing: In dynamic corporate inventory control and retail replenishment, **Simple Exponential Smoothing (SES)** provides computationally agile short-term forecasts without storing extensive historical vectors: (F_{t+1) = α Y_t + (1 - α) F_t}, where (α) (smoothing constant bounded between 0 and 1) weights recent actual sales versus past forecast memory. Higher values of (α) (e.g., 0.8) make the model hyper-responsive to recent demand shifts, whereas lower values (e.g., 0.2) generate conservative, highly smoothed forecasts. When time series exhibit linear trends, **Holt's Two-Parameter Model** incorporates a dedicated trend smoothing constant ((β)) to adjust projected slopes dynamically.

Forecasting Error Metrics in Supply Chain Planning: Forecast accuracy is evaluated via: **Mean Absolute Deviation (MAD)** (sum |Y - Ŷ| / n), directly sizing safety stock (Safety Stock approx 1.25 · z · MAD); **Mean Squared Error (MSE)** (sum (Y - Ŷ)^2 / n), penalizing large outliers; and **Mean Absolute Percentage Error (MAPE)**, expressing error as a percentage of actual sales for cross-category benchmarking.

Forecasting Protocol Selection Matrix: Strategic demand planners choose between moving averages, exponential smoothing, and least squares based on operational horizons: **Short-term operational replenishment (1 to 4 weeks):** Simple Exponential Smoothing or 3-period moving averages minimize data overhead and react to volatile swings; **Medium-term budgeting (1 to 4 quarters):** Multiplicative decomposition with seasonal indices adjusts for recurring festival peaks and working capital cycles; **Long-term capital planning (3 to 10 years):** Secular least-squares linear trend equations project factory capacity additions and infrastructure investments.

DegreeLive