SUBJECT: FOUNDATIONS FOR BUSINESS ANALYTICS

Module 3: Sampling, Correlation, and Regression Analysis

Exhaustive Applied Edition • Formula Verification Manual (Units 9 – 17)

9 & 10 Sampling and Estimation & Population Parameters vs. Sample Statistics

The Logic of Sampling and Estimation

In business analytics, it is often physically impossible, economically restrictive, or too time-consuming to audit every single item inside a massive dataset. For instance, evaluating the credit card behavior of all citizens in a country is unfeasible. Instead, we select a smaller, representative subset known as a **Sample**. The process of analyzing this sample to draw logical conclusions about the entire data universe is defined as **Statistical Estimation**. Estimation acts as an analytical bridge, translating fractional observations into macro corporate intelligence safely.

Population Parameters vs. Sample Statistics

To run predictive modeling, analytics managers must strictly separate the measurements of the entire data universe from the measurements of our specific sample subset:

- **Population Parameter:** A fixed, true numerical measurement describing a characteristic of the absolute entire population data universe. It is generally unknown and must be estimated. Denoted using classical Greek symbols (e.g., Population Mean = μ, Population Standard Deviation = σ).
- **Sample Statistic:** A variable numerical measurement computed directly from a specific sample data block. Its value is known exactly but varies depending on which random sample is pulled from the population. Denoted using standard Roman letters (e.g., Sample Mean = $\bar{x}$, Sample Standard Deviation = s).

Point Estimation vs. Interval Estimation

Statistical inference relies on two primary estimation methodologies:

1. **Point Estimation:** Using a single calculated sample statistic to serve as the best single guess for an unknown population parameter. For example, using a sample mean ($\bar{x} = ₹4,200$) to represent the true average customer expenditure (μ). While straightforward, it carries a high likelihood of error due to sampling variance.
2. **Interval Estimation (Confidence Intervals):** Constructing a range of values bounded by an upper and lower limit, within which the true population parameter is expected to fall with a specified level of statistical confidence (e.g., a 95% Confidence Interval). It incorporates a calculated safety margin known as the *Margin of Error* to account for sampling fluctuations.

11 Probability Sampling vs. Non-Probability Sampling

The validity of any business analytics insight depends entirely on how the underlying sample data is gathered. Sourcing techniques are divided into two primary structural frameworks:

I. Probability Sampling (Scientific & Representative)

Every single item within the population data universe has a known, non-zero mathematical chance of being selected into the sample. It removes personal human bias, allowing for the calculation of sampling error boundaries and the execution of parametric hypothesis tests. We evaluate four core types:

- **Simple Random Sampling (SRS):** Every individual item in the population has an identical, equal probability of selection, equivalent to an automated lottery sweep.
- **Systematic Sampling:** Selecting an initial item randomly from a list, and then picking every k-th element systematically thereafter (e.g., auditing every 50th transaction log crossing a payment gateway server).
- **Stratified Random Sampling:** Slicing a heterogeneous population into homogeneous, mutually exclusive sub-groups known as *Strata* based on shared traits (e.g., categorizing users by income brackets). A simple random sample is then pulled independently from each individual stratum proportional to its size, ensuring precise representation of minority segments.

- **Cluster Sampling:** Slicing a vast geographical population into natural, representative sub-sections known as *Clusters* (e.g., retail store locations). A few clusters are selected at random, and either all or a random sample of items within those selected clusters are audited completely.

II. Non-Probability Sampling (Subjective & Exploratory)

The selection of items rests on the personal judgment or convenience of the researcher. The mathematical probability of selection is unknown, meaning sampling error cannot be calculated scientifically. It is used primarily for fast, low-cost exploratory research, but results cannot be generalized to the entire population universe. Core types include:

- **Convenience Sampling:** Sourcing data from the most readily available subjects (e.g., surveying the nearest shoppers walking past an office door).
- **Judgmental / Purposive Sampling:** Hand-picking specific experts or data profiles that the researcher believes match the exact niche objective of the audit.
- **Quota Sampling:** Setting clear numerical target limits for specific demographic categories (e.g., ensuring exactly 50 male and 50 female users are interviewed), but relying on convenience methods to fill those targets.
- **Snowball Sampling:** Utilizing early respondents to identify and recruit subsequent participants, commonly used in niche market studies where target profiles are rare or difficult to locate.

12 Sample Size Estimation for Mean of the Population

The Sample Size Equation

To estimate the minimum required sample size (n) to find a population mean within a specified error margin, we use the following structural algebraic model:

SAMPLE SIZE FORMULA FOR THE MEAN

$$n = [(Z \cdot \sigma) / E]^2$$

Variable	*Description*
n	*The calculated required sample size.*

Variable	Description
Z	*The confidence coefficient corresponding to the target confidence level (e.g., 1.96 for 95%).*
σ	*The population standard deviation measuring historical data volatility.*
E	*The allowable maximum margin of error.*

EXAMPLE PROBLEM & SOLUTION

Problem: A retail chain wants to estimate the true average customer transaction value. Past data indicates a population standard deviation (σ) of ₹150. The analytics team demands a 95% confidence level (approximating $Z = 2$ for simplicity) and wants the final estimation to be within an allowable margin of error (E) of ± ₹15. Calculate the required sample size (n).

Solution Strategy: Identify variables: $Z = 2$, $\sigma = 150$, $E = 15$. Substitute these directly into the sample size equation:

$$n = [\ (2 \times 150) / 15\ ]^2$$
$$n = [\ 300 / 15\ ]^2$$
$$n = [\ 20\ ]^2 = \mathbf{400}$$

Interpretation: The data team must extract a sample of at least **400 customer transactions** to satisfy their target margin and precision goals.

13 Central Limit Theorem

The Core Theorem

The **Central Limit Theorem (CLT)** states that for any underlying population distribution, the sampling distribution of the sample means will converge into a symmetrical Normal Distribution shape as the

sample size scales upwards ($n \geq 30$). The variability of this sampling distribution is governed by the **Standard Error (SE)** formula.

[Image of the Central Limit Theorem convergence process showing diverse population shapes smoothing into a normal curve]

STANDARD ERROR FORMULA (SE)

$$SE = \sigma_{\bar{x}} = \sigma / \sqrt{n}$$

Where σ is the population standard deviation, and n is the sample size.

EXAMPLE PROBLEM & SOLUTION

Problem: A high-frequency logistics company knows that the packing weight of boxes has a population standard deviation (σ) of 80 grams. If an automated quality assurance gate draws a random sample of $n = 64$ boxes, what is the exact Standard Error (SE) of the sample mean?

Solution Strategy: Identify inputs: $\sigma = 80$, $n = 64$. Apply the Standard Error equation:

$$SE = 80 / \sqrt{64}$$
$$SE = 80 / 8 = \mathbf{10\ grams}$$

Interpretation: The standard deviation of our sample means is 10 grams, showing how the variance compresses as sample size scales up.

14 & 15 Correlation Analysis: Karl Pearson's & Spearman's Rank Methods

Correlation analysis measures the direction and strength of the relationship between two variables, bounded strictly within $-1 \leq r \leq 1$.

[Image of scatter plots showing positive, negative, and zero correlation gradients]

I. Karl Pearson's Correlation Coefficient Formula

KARL PEARSON'S PRODUCT-MOMENT FORMULA (R)

$$r = [\, n\Sigma XY - (\Sigma X)(\Sigma Y) \,] \,/\, \sqrt{[\, [\, n\Sigma X^2 - (\Sigma X)^2 \,] \cdot [\, n\Sigma Y^2 - (\Sigma Y)^2 \,] \,]}$$

EXAMPLE PROBLEM & SOLUTION

Problem: A tech company tracks corporate ad spend (X) and web conversions (Y) over $n = 5$ testing blocks. The compiled data summaries are calculated as follows: $\Sigma X = 15, \Sigma Y = 20, \Sigma XY = 68, \Sigma X^2 = 55, \Sigma Y^2 = 90$. Calculate Karl Pearson's correlation coefficient (r).

Solution Strategy: Substitute the sum components directly into the algebraic matrix equation:

$$\text{Numerator} = 5(68) - (15)(20) = 340 - 300 = 40$$

$$\text{Denominator Term 1} = [\, 5(55) - (15)^2 \,] = 275 - 225 = 50$$
$$\text{Denominator Term 2} = [\, 5(90) - (20)^2 \,] = 450 - 400 = 50$$
$$\text{Full Denominator} = \sqrt{[\, 50 \times 50 \,]} = 50$$

$$r = 40 / 50 = \mathbf{0.80}$$

Interpretation: A coefficient of **+0.80** signals a strong positive linear correlation, confirming that increased ad spending is highly correlated with increased conversions.

II. Spearman's Rank Correlation Coefficient Formula

SPEARMAN'S NON-PARAMETRIC RANK FORMULA (R_S)

$$R_s = 1 - [\, [\, 6 \cdot \Sigma d^2 \,] / [\, n \cdot (n^2 - 1) \,] \,]$$

Where d is the numerical difference between paired ranks, and n is the observation count.

EXAMPLE PROBLEM & SOLUTION

Problem: Two managers independently rank $n = 5$ employee candidates. The sum of the squared differences between their assigned ranks is calculated as $\Sigma d^2 = 4$. Find Spearman's Rank Correlation (R_s).

Solution Strategy: Identify variables: $n = 5$, $\Sigma d^2 = 4$. Substitute into the rank formula:

$$R_s = 1 - [(6 \times 4) / (5 \times (5^2 - 1))]$$
$$R_s = 1 - [24 / (5 \times (25 - 1))]$$
$$R_s = 1 - [24 / (5 \times 24)]$$
$$R_s = 1 - [24 / 120]$$
$$R_s = 1 - 0.20 = \mathbf{0.80}$$

Interpretation: A rank correlation score of **0.80** indicates strong agreement between the two managers' evaluation profiles.

16 & 17 Regression Analysis & Regression vs. Correlation

Regression builds a predictive mathematical functional equation to model a dependent variable (Y) based on an independent variable (X).

[Image of a linear regression line tracking through a scatter plot showing intercept, slope, and error residuals]

I. Linear Regression Equations Framework

THE LINES OF REGRESSION AND SLOPES

Line of Y on X: $Y - \bar{Y} = b_{yx} \cdot (X - \bar{X})$ | *Slope:* $b_{yx} = [n\Sigma XY - (\Sigma X)(\Sigma Y)] / [n\Sigma X^2 - (\Sigma X)^2]$

Line of X on Y: $X - \bar{X} = b_{xy} \cdot (Y - \bar{Y})$ | *Slope:* $b_{xy} = [n\Sigma XY - (\Sigma X)(\Sigma Y)] / [n\Sigma Y^2 - (\Sigma Y)^2]$

EXAMPLE PROBLEM & SOLUTION

Problem: Utilizing the exact same data from our previous tech company dataset ($n = 5, \Sigma X = 15, \Sigma Y = 20, \Sigma XY = 68, \Sigma X^2 = 55, \Sigma Y^2 = 90$), calculate the means ($\bar{X}, \bar{Y}$), the regression slope (b_{yx}), and construct the final linear regression line equation of Y on X.

Solution Strategy: First find the means, then calculate the slope coefficient, and arrange into the linear line formula:

1. Calculate Means: $\bar{X} = 15 / 5 = 3 \mid \bar{Y} = 20 / 5 = 4$

2. Calculate Slope b_{yx}*:*
Numerator $= 5(68) - (15)(20) = 40$ *(Derived in Pearson example)*
Denominator $= [\, 5(55) - (15)^2 \,] = 275 - 225 = 50$
$b_{yx} = 40 / 50 =$ ***0.80***

3. Assemble Line Equation:
$Y - 4 = 0.80 \cdot (X - 3)$
$Y - 4 = 0.80X - 2.4$
$\boldsymbol{Y = 0.80X + 1.6}$

Interpretation: The final predictive model is **Y = 0.80X + 1.6**. For every additional unit increase in ad spending (X), conversions (Y) are predicted to grow by 0.80 units, starting from a base baseline of 1.6 units.

II. Standard Error of Estimate Formula

STANDARD ERROR OF ESTIMATE EQUATION (S_{YX})

$$S_{yx} = \sqrt{[\, \Sigma(Y - \hat{Y})^2 / (n - 2) \,]}$$

Where Y represents actual data points, $\hat{Y}$ is the predicted line values, and n is the sample count.

EXAMPLE PROBLEM & SOLUTION

Problem: An analyst calculates the prediction errors (residuals) for a regression model across $n = 5$ data points. The sum of the squared errors is found to be $\Sigma(Y - Y)^2 = 12$. Calculate the Standard Error of Estimate (S_{yx}).

Solution Strategy: Substitute variables directly into the variance denominator layout:

$$S_{yx} = \sqrt{[\ 12 / (5 - 2)\ ]}$$
$$S_{yx} = \sqrt{[\ 12 / 3\ ]}$$
$$S_{yx} = \sqrt{4} = \mathbf{2}$$

Interpretation: The Standard Error of Estimate is **2**, providing a precise standard metric of the typical variance or margin of error around the model's predictions.

End of Module 3 • Subject: Foundations for Business Analytics