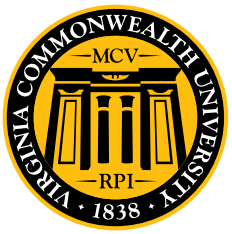

VIRGINIA COMMONWEALTH UNIVERSITY

Statistical Analysis and Modelling (SCMA 632)

A1: Data Cleaning and Statistical Analysis

UJJWAL AKSHITH MONDRETI

V01153239

Date of Submission: 12-06-2025

CONTENTS

Introduction

This report details the analysis of consumption patterns in Meghalaya using data from the 68th round of the National Sample Survey Organisation (NSSO68). The analysis involves data cleaning, imputation of missing values, outlier removal, calculation of total consumption, and comparative statistical tests between different demographic groups and districts within the state.

Results

- Data Overview and Preprocessing
    - The initial dataset for Meghalaya (state code 17) contained 763 observations and 209 variables. After selecting relevant columns, missing values in *Meals_At_Home* and *No_of_Meals_per_day* were imputed using their respective means. Outliers in *ricetotal_v*, *wheattotal_v*, *Milktotal_v*, and *pulsestot_v* were removed using the Inter-Quartile Range method. A new variable, *total_consumption_v*, was created by summing up consumption values for rice, wheat, milk, pulses, non-vegetarian items, and fruits. The final processed dataset for Meghalaya contained 1259 observations and 384 variables after filtering and outlier removal.

- Consumption Summaries
    - District Consumption: The analysis identified the top and bottom consuming districts base on *total_consumption_v*.
        - Top Consuming District: East Khasi Hills (total consumption: 70280.35)
        - Bottom Consuming District: South Garo Hills (total consumption: 19920.00)
    - Region Consumption:
        - Region 1: 269421.14
    - Sector Consumption:
        - Rural: 184565.41
        - Urban: 84855.73
- Z-Test Results
    - Rural vs. Urban Consumption
        - Mean consumption in Rural areas: 228.99
        - Mean consumption in Urban areas: 271.97
        - Z-Test p-value: 0.00000
        - Conclusion: The p-value is less than 0.05, leading to the rejection of the null hypothesis. There is a significant difference between the mean consumption of urban and rural areas.
    - Top Consuming District (East Khasi Hills) Consumption vs. Bottom Consuming District (South Garo Hills) Consumption
        - Mean consumption in East Khasi Hills: 252.80
        - Mean consumption in South Garo Hills: 245.93
        - Z-Test p-value: 0.37814
        - Conclusion: The p-value is greater than 0.05, leading to a failure to reject the null hypothesis. There is no significant difference between the mean consumptions of East Khasi Hills and South Garo Hills.

Interpretations

The analysis reveals interesting consumption patterns across Meghalaya. The Z-test comparing rural and urban consumption indicates a statistically significant difference, with urban areas showing a higher mean per-person consumption. This suggests that urban dwellers in Meghalaya, on average, consume more than their rural counterparts, possibly due to factors such as higher disposable income, better access to diverse goods, or different lifestyle patterns.

Conversely, the Z-test comparing the top and bottom consuming districts (East Khasi Hills and South Garo Hills) **does not show a statistically significant difference** in their mean consumption values. While East Khasi Hills has a substantially higher *total* consumption, likely attributable to a larger population or higher number of surveyed households, the mean consumption per person is quite like that of South Garo Hills. This suggests that while overall consumption volume differs, the individual consumption habits or levels between an average person in the most consuming district and the least consuming district are not significantly different when considered on a per-person mean basis.

Recommendations

1. **Investigate Nuances of Urban-Rural Divide:** Further research is needed to pinpoint the specific factors contributing to the significant difference in mean consumption between urban and rural areas. This could include examining income disparities, market accessibility, and dietary preferences.
2. **Population Context for District Consumption:** When evaluating district-level consumption, it's crucial to consider population size alongside total consumption figures. The current analysis suggests that while total consumption varies significantly by district, mean per-person consumption might not, indicating population size as a primary driver of aggregate consumption differences.
3. **Detailed Food Category Analysis:** A more granular analysis of specific food categories (e.g., rice, milk, non-veg) could provide deeper insights into nutritional patterns and potential food security issues in different regions.
4. **Policy Implications:** Understanding these consumption dynamics is vital for policymakers to design targeted interventions, especially in areas with lower overall consumption, to ensure equitable access to food and improve living standards.

References

- National Sample Survey Organisation (NSSO). (2014). *Report No. 558: Household Consumption of Various Goods and Services in India, 2011-12*. Ministry of Statistics and Programme Implementation, Government of India. https://mospi.gov.in/sites/default/files/publication_reports/Report_no558_rou68_30june14.pdf