****

**VIRGINIA COMMONWEALTH UNIVERSITY**

**Statistical analysis and modelling (SCMA 632)**

**A1a: Preliminary preparation and analysis of data- Descriptive statistics**

**Manoranjan Mohankumar**

**V01107254**

**Date of Submission: 16-06-2024**

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Title** | **Page No.** |
| **1.** | Introduction | **1** |
| **2.** | Results | **2** |
| **3.** | Interpretations | **3** |
| **4.** | Recommendations | **5** |
| **5.** | Codes | **7** |
| **6.** | References | **10** |

**Introduction**

Understanding consumption patterns is crucial for effective marketing, resource allocation, and strategic planning. Consumption data can reveal significant insights into the preferences and needs of different populations, which can be used to tailor products and services accordingly. In this context, rural and urban areas often exhibit distinct consumption behaviours due to differences in lifestyle, accessibility, socio-economic factors, and cultural practices.

This report compares the total consumption between rural and urban areas using a robust statistical approach. The analysis is conducted on a dataset comprising various consumption metrics, including quantities of food items, beverages, and meals consumed at home. By examining this data, the objective is to determine if there is a significant difference in total consumption between rural and urban regions.

The specific steps involved in this analysis include:

* Data Collection: Gathering consumption data for both rural and urban areas.
* Data Preprocessing: Cleaning and organizing the data to ensure it is suitable for analysis.
* Statistical Testing: Applying a Z-test to compare the means of total consumption between the two groups.
* Interpretation: Analysing the results to understand the extent and significance of the difference in consumption patterns.

The findings of this analysis will provide insights into how consumption behaviors vary between rural and urban areas, which can inform targeted marketing strategies, resource distribution, and further research initiatives.

**Results**

The analysis used a Z-test to compare the total consumption between rural and urban areas. The Z-test is a statistical method used to determine whether there is a significant difference between the means of two groups. Here are the detailed results of the Z-test:

* Z-Score: 15.349380604793554
* P-Value: 3.5758420317152064e-53

**Summary of Findings:**

The Z-test results reveal a highly significant difference in total consumption between rural and urban regions.

* The exceptionally high Z-score and the near-zero p-value provide statistically solid evidence that the mean total consumption in these two areas is unequal.
  + Z-Score: A Z-score of 15.35 indicates that the difference between the mean total consumption of rural and urban areas is 15.35 standard deviations away from the mean. This extremely high value suggests a substantial difference between the two groups.
  + P-Value: The p-value of approximately 3.58e-53 (which is effectively zero) indicates overwhelming evidence against the null hypothesis. In practical terms, this means a statistically significant difference in total consumption between rural and urban areas. A p-value this low suggests that the probability of observing such a difference by chance is virtually zero.
* These findings highlight that rural and urban populations have distinct consumption patterns, which could be attributed to various socio-economic and lifestyle factors.
  + Socio-Economic Factors: Rural areas might have different levels of income, access to resources, and consumption habits compared to urban areas.
  + Lifestyle Factors: Differences in lifestyle, such as dietary preferences and market access, can significantly influence consumption patterns.

Overall, the results of the Z-test confirm a significant disparity in total consumption between rural and urban areas, necessitating targeted approaches to address these populations' unique needs and preferences.

**Interpretations**

The results of the Z-test provide significant insights into the differences in total consumption between rural and urban areas. Here is a detailed interpretation:

Z-Score: 15.349380604793554

The Z-score represents the number of standard deviations the observed difference between the means of the two groups is from the overall mean. In this case, a Z-score of 15.35 is exceptionally high.

**Significance of the Z-Score:**

* A high Z-score (greater than 2 or 3) typically indicates a significant difference between the compared groups.
* A Z-score of 15.35 suggests that the difference in total consumption between rural and urban areas is 15.35 standard deviations away from the mean, which is a substantial deviation.
* This significant Z-score implies that the observed difference is not due to random variation or chance but is statistically significant.

P-Value: 3.5758420317152064e-53

The p-value indicates the probability of observing the test results under the null hypothesis, which assumes no difference between the group means.

**Significance of the P-Value:**

* A p-value is used to determine the statistical significance of the test results. Typically, a p-value less than 0.05 is considered statistically significant.
* In this analysis, the p-value is approximately 3.58e-53, which is extraordinarily low (close to zero).
* A p-value of this magnitude provides strong evidence against the null hypothesis. It suggests that the probability of observing such a significant difference by chance is virtually zero.
* Therefore, we can confidently reject the null hypothesis and conclude that there is a statistically significant difference in total consumption between rural and urban areas.

Implications of the Results

**Statistical Significance:**

* The combination of a very high Z-score and an extremely low p-value clearly indicates that the difference in mean total consumption between rural and urban regions is statistically significant.
* This statistical significance means that the observed differences are highly unlikely to be due to random variation, and there are underlying factors contributing to the distinct consumption patterns in these regions.

**Practical Significance:**

* The significant difference in consumption patterns can be attributed to various socio-economic, cultural, and lifestyle differences between rural and urban populations.
* Rural areas may have different access to goods and services, dietary preferences, and economic conditions compared to urban areas, which could explain the variation in total consumption.

**Decision-Making and Strategy:**

* These findings have practical implications for businesses, policymakers, and researchers.
* Businesses can tailor their marketing and distribution strategies to meet the specific needs of rural and urban consumers.
* Policymakers can design targeted interventions to address the unique consumption patterns and improve resource allocation.
* Further research can be conducted to explore the specific factors driving these differences and develop strategies to bridge the consumption gap.

Overall, the results highlight the importance of understanding and addressing the distinct consumption behaviors of rural and urban populations to optimize strategies and improve outcomes in various sectors.

**Recommendations**

Based on the significant difference in total consumption between rural and urban areas, several strategic recommendations can be made to optimize marketing efforts, resource allocation, and further research:

**Targeted Marketing Strategies**

Customized Marketing Campaigns:

* Develop marketing campaigns specifically tailored to the unique consumption behaviours and preferences of rural and urban populations.
* For rural areas, emphasize accessibility, affordability, and practicality of products. Highlight features that cater to rural lifestyles and needs.
* For urban areas, focus on convenience, premium quality, and modern lifestyle benefits. Use digital marketing channels more extensively to reach tech-savvy urban consumers.

Localized Advertising:

* Utilize local media channels, such as regional newspapers, radio, and community events, to reach rural consumers effectively.
* In urban areas, leverage social media platforms, online advertising, and urban billboards to capture the attention of a more diverse and digitally connected audience.

**Resource Allocation**

Efficient Distribution:

* Allocate distribution resources based on the consumption patterns observed in rural and urban areas. Ensure that products are readily available where they are most needed.
* For rural areas, enhance the supply chain infrastructure to improve product availability and reduce delivery times. Consider partnerships with local distributors to streamline logistics.
* In urban areas, optimize inventory management to ensure high-demand products are always in stock, reducing the risk of stockouts and enhancing customer satisfaction.

**Product Development:**

* Develop product variations that cater specifically to the tastes and preferences of rural and urban consumers.
* For rural consumers, consider creating value-for-money product packages and sizes that are economically accessible.
* For urban consumers, focus on innovative, premium, and convenience-oriented product offerings.

**Policy Recommendations**

Support for Rural Development:

* Advocate for policies that support rural development, such as improving infrastructure, enhancing access to education and healthcare, and promoting economic opportunities. These efforts can indirectly influence consumption patterns by improving the overall quality of life in rural areas.

**Urban Planning:**

* Support urban planning initiatives that address the challenges of urbanization, such as traffic congestion, pollution, and housing shortages. Improved urban living conditions can positively impact consumption patterns and overall well-being.

By implementing these recommendations, businesses and policymakers can better address the unique needs and preferences of rural and urban consumers, leading to more effective strategies, optimized resource allocation, and improved market outcomes.

**Codes**

# In[1]:

import os, pandas as pd, numpy as np

# In[3]:

os.chdir("E:\VCU Bridge course\Python\_WD")

# In[4]:

df=pd.read\_csv("NSSO68.csv",encoding="Latin-1", low\_memory=False)

# In[6]:

df.head()

# In[7]:

UP = df[df['state\_1']=="UP"]

# In[8]:

UP.isnull().sum().sort\_values(ascending = False)

# In[9]:

df.columns

# In[11]:

UP\_new = UP[['state\_1', 'District', 'Sector','Region','State\_Region','ricetotal\_q','wheattotal\_q','moong\_q','Milktotal\_q','chicken\_q','bread\_q','foodtotal\_q','Beveragestotal\_v','Meals\_At\_Home']]

# In[12]:

UP\_new.isnull().sum().sort\_values(ascending = False)

# In[13]:

UP\_clean = UP\_new.copy()

# In[15]:

UP\_clean.loc[:, 'Meals\_At\_Home'] = UP\_clean['Meals\_At\_Home'].fillna(UP\_new['Meals\_At\_Home'].mean())

# In[16]:

UP\_clean.isnull().any()

# In[17]:

# Outlier Checking

# In[18]:

import matplotlib.pyplot as plt

# Assuming UP\_clean is your DataFrame

plt.figure(figsize=(8, 6))

plt.boxplot(UP\_clean['ricetotal\_q'])

plt.xlabel('ricetotal\_q')

plt.ylabel('Values')

plt.title('Boxplot of ricetotal\_q')

plt.show()

# In[20]:

rice1 = UP\_clean['ricetotal\_q'].quantile(0.25)

rice2 = UP\_clean['ricetotal\_q'].quantile(0.75)

iqr\_rice = rice2-rice1

up\_limit = rice2 + 1.5\*iqr\_rice

low\_limit = rice1 - 1.5\*iqr\_rice

# In[22]:

UP\_clean=UP\_new[(UP\_new['ricetotal\_q']<=up\_limit)&(UP\_new['ricetotal\_q']>=low\_limit)]

plt.boxplot(UP\_clean['ricetotal\_q'])

# In[24]:

UP\_clean['District'].unique()

# In[25]:

# Replace values in the 'Sector' column

UP\_clean.loc[:,'Sector'] = UP\_clean['Sector'].replace([1, 2], ['URBAN', 'RURAL'])

# In[26]:

UP\_clean.columns

# In[30]:

UP\_clean.loc[:, 'total\_consumption'] = UP\_clean[['ricetotal\_q', 'wheattotal\_q', 'moong\_q', 'Milktotal\_q', 'chicken\_q', 'bread\_q', 'foodtotal\_q', 'Beveragestotal\_v']].sum(axis=1)

# In[32]:

UP\_clean.head()

# In[33]:

UP\_clean.groupby('Region').agg({'total\_consumption':['std','mean','max','min']})

# In[34]:

UP\_clean.groupby('District').agg({'total\_consumption':['std','mean','max','min']})

# In[35]:

total\_consumption\_by\_districtcode=UP\_clean.groupby('District')['total\_consumption'].sum()

# In[36]:

total\_consumption\_by\_districtcode.sort\_values(ascending=False).head(3)

# In[40]:

UP\_clean.loc[:,"District"] = UP\_clean.loc[:,"District"].replace({34: "Kanpur Nagar", 67: "Varanasi", 2: "Muzaffarnagar"})

# In[41]:

total\_consumption\_by\_districtname=UP\_clean.groupby('District')['total\_consumption'].sum()

# In[42]:

total\_consumption\_by\_districtname.sort\_values(ascending=False).head(3)

# In[43]:

from statsmodels.stats import weightstats as stests

# In[45]:

rural=UP\_clean[UP\_clean['Sector']=="RURAL"]

urban=UP\_clean[UP\_clean['Sector']=="URBAN"]

# In[46]:

rural.head()

# In[47]:

urban.head()

# In[48]:

cons\_rural=rural['total\_consumption']

cons\_urban=urban['total\_consumption']

# In[49]:

z\_statistic, p\_value = stests.ztest(cons\_rural, cons\_urban)

# Print the z-score and p-value

print("Z-Score:", z\_statistic)

print("P-Value:", p\_value)

**References**

* Dataset Source: The dataset used for this analysis was provided and loaded into 'rural' and 'urban' DataFrames.
* ChatGPT